

# **Towards an Universal Person Description Framework for Looking at People Applications**

Von der Fakultät IV - Elektrotechnik und Informatik  
der Technischen Universität Berlin  
zur Verleihung des akademischen Grades  
Doktor der Ingenieurwissenschaften  
- Dr.-Ing. -  
genehmigte Dissertation

vorgelegt von  
Dipl.-Ing. Lutz Goldmann

## **Promotionsausschuss**

Vorsitzender: Prof. Dr.-Ing. Reinhold Orglmeister

1. Gutachter: Prof. Dr.-Ing. Thomas Sikora

2. Gutachter: Prof. Dr. Francesc Tarrés

Tag der wissenschaftlichen Aussprache: 07.09.2009

Berlin 2010

D 83



*To the light of my life*





# Acknowledgement

This thesis emerged during my work as a research assistant at the Communication Systems Group of the Technical University of Berlin. At this point I would like to thank all the people that have supported me on this long way.

First of all, I would like to express my sincere gratitude to my supervisor Prof. Dr.-Ing. Thomas Sikora for the chance to work on this challenging and interesting research topic. I appreciate the freedom he gave me in my research and the fruitful advice I received from him every time I was stuck in a problem. I also want to thank Prof. Dr. Francesc Tarrés for the detailed review of my thesis, the encouraging feedback and interesting discussions.

Special thanks go to my colleagues of the group for sharing time with me inside and outside the office. In particular, I would like to thank my roommates Mustafa Karaman and Amjad Samour for the helpful discussions, encouraging words and long evenings to meet a particular deadline. Furthermore, I would like to appreciate all the organizational support by Birgit Boldin who is the kind soul of the institute.

Within the scope of several European projects I had the chance to meet and exchange ideas with numerous researchers working in the same research field. In particular, I would like to express my gratitude to Toni Rama for the interesting discussions and the fruitful collaboration that lead too several joint articles.

This work would not have been the same without the help and commitment of diploma and master students under my supervision. Special thanks go to Ullrich Moenich and Lars Thiele for their excellent work that is referenced within this thesis.

There are some moments in life when it is important to look back and remember where we come from. This work would not have been possible without the constant support of my dear parents and grandparents at home. At the same time I want to thank all my friends that have continuously reminded me that there is more in life than work.

All the things we see and hence this whole work are based on the existence of light. There is a light that has guided me along the way. This light is called Sylvia.



# Eidesstattliche Erklärung

Ich versichere an Eides statt, dass ich die von mir vorgelegte Dissertation selbstständig angefertigt und alle benutzten Quellen und Hilfsmittel vollständig angegeben habe.

Eine Anmeldung der Promotionsabsicht habe ich an keiner anderen Fakultät oder Hochschule beantragt.



# Abstract

During the last decade computers and the internet have become an important aspect in our everyday life. We use this technology to communicate, study, work, shop, and entertain ourselves. The vision of the future is to embed this computing technology into our home, transportation and working environments. The ultimate goal is to develop intelligent machines that are aware of humans and can assist them if required. Therefore, visual data needs to be analyzed with respect to humans which is often referred to as "looking at people". So far the developments within this area have been largely influenced by the interests and needs of specific applications (surveillance, biometrics, human computer interaction).

The objective of this dissertation is to move towards an universal framework for the visual analysis of humans, that describes humans at several levels including different body parts (body, face, hands) and features (color, texture, shape and motion). Then, in analogy to the human visual perception, an appropriate subset of the provided information can be chosen, depending on environmental or application specific criteria. Within the scope of such a framework, this work provides scientific contributions in several areas. For face detection a novel component based face detection approach has been developed that combines techniques from the statistical and structural pattern recognition domain for improved performance especially in the presence of partial occlusions. It is not only able to detect faces despite occlusions, but can also provide additional occlusion information to subsequent face analysis steps. Based on that, existing appearance based face recognition approaches have been extended through occlusion awareness by selecting the most reliable representation. For appearance based body recognition both holistic and component based representations and a large set of color and texture features have been considered to determine the optimal description of a person's clothes.

The developed framework has been used within several applications to prove its versatility. The first original application, that has been developed, is an efficient system for the audiovisual search of persons based on facial appearance and voice characteristics. A high retrieval performance is achieved through the combination of multimodal fusion and relevance feedback. For the second application, an original system for visual person search, a different query paradigm was used. It provides an intuitive query interface through an automatically derived human visual thesaurus that groups people based on their visual similarity. Finally, the appearance based analysis was combined with motion based analysis

for a personalized human computer interface that detects, tracks and identifies humans and interprets their gestures for the use in an intelligent cash machine scenario.

Although this dissertation focuses only on the appearance based description of face and body, ideas and findings may also be applied to other channels (hands, limbs), features (shape, motion) and tasks (tracking). Therefore it contributes to the gradual change from an application specific view towards a universal framework for the visual analysis of humans, which will enable machines to sense and react to humans in a more natural way.

# Zusammenfassung

Im vergangenen Jahrzehnt sind der Computer und das Internet zu einem wichtigen Bestandteil unseres täglichen Lebens geworden. Wir verwenden diese Technologien um zu kommunizieren, zu arbeiten, einzukaufen, und für unsere Unterhaltung. Die Zukunft sieht eine stärkere Einbettung dieser Technologien in unsere tägliche Umgebung (Heim, Büro und öffentliche Räume) vor. Dabei besteht das Ziel in der Entwicklung intelligenter Maschinen, die in der Lage sind, Menschen in einer Umgebung wahrzunehmen und helfend zur Seite zu stehen. Dafür bedarf es einer auf den Menschen fokussierten Analyse der hauptsächlich visuellen Daten, was auch als "Looking at People" bezeichnet wird. Die bisherigen Entwicklungen in diesem Forschungsgebiet sind stark von den Anforderungen bestimmter Anwendungen (z.B. Überwachung, Biometrie, Mensch-Maschine-Interaktion) geprägt.

Das Ziel dieser Dissertation ist die Entwicklung eines universellen Systems für die visuelle Analyse des Menschen, welches diesen auf mehreren Ebenen anhand verschiedener Teile (Körper, Gesicht, Hände) und Merkmale (in Farbe, Textur, Form und Bewegung) beschreibt. In Analogie zur menschlichen Wahrnehmung kann dann abhängig von Anwendung oder Umgebungsbedingungen ein geeigneter Teil der Beschreibung berücksichtigt werden. Im Rahmen dieses universellen Systems liefert diese Dissertation wissenschaftliche Beiträge in verschiedenen Bereichen. Für die Gesichtsdetektion wurde ein neuer komponentenbasierter Ansatz entwickelt, der Techniken der statistischen und der strukturellen Mustererkennung miteinander vereint, um auch teilweise verdeckte Gesichter zu detektieren. Darüber hinaus ist der entwickelte Ansatz in der Lage, zusätzliche Informationen über das Vorhandensein und die Lage der Verdeckungen den folgenden Analyseschritten zur Verfügung zu stellen. Darauf basierend wurden existierende Ansätze der Gesichtserkennung durch eine intelligente Fusion erweitert, um die Robustheit bei teilweisen Verdeckungen zu erhöhen. Für die optimale Beschreibung der äußeren Erscheinung (Kleidung) eines Menschen wurden sowohl ganzheitliche als auch komponentenbasierte Modelle und eine große Auswahl an Farb- und Texturmerkmalen berücksichtigt.

Um das System zur Detektion und Beschreibung von Personen auf seine Vielseitigkeit zu testen, wurden verschiedene Anwendungen entwickelt. Bei der ersten Anwendung handelt es sich um ein effizientes System zur audiovisuellen Suche von Personen in Videos anhand ihres Gesichts und ihrer Stimme. Die hohe Genauigkeit der Suche wurde dabei durch die Fusion der verschiedenen Modalitäten und die Integration des Nutzers in den Suchprozeß

erreicht. Für die zweite Anwendung, einem System zur visuellen Suche von Personen in Bildern, wurde ein grundlegend anderer Ansatz verfolgt. Er basiert auf einer intuitiven Suchanfrage mittels eines visuellen Thesaurus, welcher die Personen in einer Datenbank anhand ihrer Ähnlichkeit gruppiert. In der letzten Anwendung wurde die Analyse der äußeren Erscheinung mit der von Bewegungen in einem System für personalisierte Mensch-Maschine-Interaktion kombiniert. Dieses System detektiert und verfolgt Personen, deren Gesichter und Hände, identifiziert sie und interpretiert ihre Gesten zur Steuerung eines intelligenten Bankautomaten.

Obwohl sich diese Dissertation nur mit der Detektion und Beschreibung von Gesicht und Körper, basierend auf der äußeren Erscheinung befasst, können die grundlegenden Ideen auch auf andere Körperteile (Gliedermaßen, Hände), Merkmale (Form, Bewegung) und Aufgaben (Verfolgung) angewendet werden. Damit leistet diese Arbeit einen Beitrag zum schrittweisen Wandel von einer anwendungsbezogenen zu einer universellen visuellen Analyse des Menschen, welche es Maschinen ermöglicht, Menschen wahrzunehmen und auf natürliche Art und Weise zu reagieren.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Looking at people . . . . .	2
1.3	Objectives . . . . .	4
1.4	Contributions . . . . .	6
1.5	Organization . . . . .	7
<b>2</b>	<b>Looking at people and related fields</b>	<b>11</b>
2.1	Introduction . . . . .	11
2.2	Surveillance . . . . .	11
2.2.1	History and trends . . . . .	12
2.2.2	Modules . . . . .	13
2.2.3	Discussion . . . . .	16
2.3	Biometrics . . . . .	16
2.3.1	Biometric traits . . . . .	17
2.3.2	Multi biometrics . . . . .	18
2.3.3	Discussion . . . . .	20
2.4	Multimedia search and retrieval . . . . .	21
2.4.1	Content description . . . . .	22
2.4.2	Content retrieval . . . . .	23
2.4.3	Discussion . . . . .	25
2.5	Conclusion . . . . .	26
<b>3</b>	<b>Fundamental techniques</b>	<b>27</b>
3.1	Introduction . . . . .	27
3.2	Image processing . . . . .	27
3.2.1	Point operations . . . . .	28
3.2.2	Geometric transformations . . . . .	29
3.2.3	Image analysis . . . . .	30
3.3	Machine learning . . . . .	30
3.3.1	Matching . . . . .	31

3.3.2	Feature reduction . . . . .	33
3.3.3	Density estimation . . . . .	38
3.3.4	Clustering . . . . .	41
3.3.5	Classification . . . . .	44
3.4	Information fusion . . . . .	48
3.4.1	Premapping fusion . . . . .	48
3.4.2	Postmapping fusion . . . . .	49
3.4.3	Score normalization . . . . .	51
3.5	Graph theory . . . . .	52
3.5.1	Graph concepts . . . . .	53
3.5.2	Graph matching . . . . .	53
<b>4</b>	<b>Visual person description framework</b>	<b>57</b>
4.1	Introduction . . . . .	57
4.2	Human description . . . . .	59
4.3	System overview . . . . .	60
4.4	Human anthropometry and modeling . . . . .	62
4.4.1	Body anthropometry . . . . .	62
4.4.2	Face anthropometry . . . . .	63
4.5	Conclusion . . . . .	65
4.5.1	Summary . . . . .	65
4.5.2	Future work . . . . .	66
<b>5</b>	<b>Body recognition</b>	<b>67</b>
5.1	Introduction . . . . .	67
5.1.1	Related work . . . . .	68
5.1.2	Challenges . . . . .	69
5.1.3	Objective . . . . .	70
5.2	Approach . . . . .	70
5.2.1	Representation . . . . .	71
5.2.2	Description . . . . .	76
5.2.3	Recognition . . . . .	90
5.2.4	Fusion . . . . .	90
5.3	Experiments . . . . .	90
5.3.1	Dataset . . . . .	91
5.3.2	Evaluation . . . . .	91
5.3.3	Results . . . . .	92
5.4	Conclusion . . . . .	97
5.4.1	Summary . . . . .	97
5.4.2	Future work . . . . .	97

<b>6</b>	<b>Face detection</b>	<b>99</b>
6.1	Introduction . . . . .	99
6.1.1	Related work . . . . .	99
6.1.2	Challenges . . . . .	102
6.1.3	Objective . . . . .	103
6.2	Holistic approach . . . . .	104
6.2.1	Description . . . . .	104
6.2.2	Classification . . . . .	107
6.3	Component based approach . . . . .	109
6.3.1	Component detection . . . . .	110
6.3.2	Topology verification . . . . .	110
6.3.3	Face localization . . . . .	115
6.3.4	Occlusion localization . . . . .	116
6.4	Experiments . . . . .	117
6.4.1	Dataset . . . . .	118
6.4.2	Evaluation . . . . .	119
6.4.3	Results . . . . .	120
6.5	Conclusion . . . . .	126
6.5.1	Summary . . . . .	126
6.5.2	Future work . . . . .	127
<b>7</b>	<b>Face recognition</b>	<b>129</b>
7.1	Introduction . . . . .	129
7.1.1	Related work . . . . .	130
7.1.2	Challenges . . . . .	134
7.1.3	Objective . . . . .	136
7.2	Approach . . . . .	136
7.2.1	Description . . . . .	136
7.2.2	Representation . . . . .	139
7.2.3	Reduction . . . . .	140
7.2.4	Recognition . . . . .	141
7.2.5	Fusion . . . . .	141
7.3	Experiments . . . . .	142
7.3.1	Dataset . . . . .	142
7.3.2	Evaluation . . . . .	144
7.3.3	Results . . . . .	144
7.4	Conclusion . . . . .	146
7.4.1	Summary . . . . .	146
7.4.2	Future work . . . . .	147

<b>8</b>	<b>Multimodal person search</b>	<b>149</b>
8.1	Introduction . . . . .	149
8.1.1	Motivation . . . . .	149
8.1.2	Related work . . . . .	149
8.1.3	Objective . . . . .	150
8.2	System overview . . . . .	150
8.2.1	Audio analysis . . . . .	151
8.2.2	Video analysis . . . . .	153
8.2.3	Query by example (QBE) . . . . .	153
8.2.4	Relevance feedback (RF) . . . . .	154
8.2.5	Multimodal fusion . . . . .	155
8.3	Experiments . . . . .	155
8.3.1	Dataset . . . . .	155
8.3.2	Evaluation . . . . .	156
8.3.3	Results . . . . .	157
8.4	Conclusion . . . . .	159
8.4.1	Summary . . . . .	159
8.4.2	Future work . . . . .	159
<b>9</b>	<b>Visual person search</b>	<b>161</b>
9.1	Introduction . . . . .	161
9.1.1	Motivation . . . . .	161
9.1.2	Related work . . . . .	162
9.1.3	Objectives . . . . .	163
9.2	System overview . . . . .	163
9.2.1	Body analysis . . . . .	164
9.2.2	Face analysis . . . . .	164
9.2.3	Visual thesaurus creation . . . . .	165
9.2.4	Query by visual thesaurus . . . . .	168
9.3	Experiments . . . . .	172
9.3.1	Dataset . . . . .	174
9.3.2	Evaluation . . . . .	174
9.3.3	Results . . . . .	174
9.4	Conclusion . . . . .	177
9.4.1	Summary . . . . .	177
9.4.2	Future work . . . . .	178
<b>10</b>	<b>Personalized human computer interaction</b>	<b>179</b>
10.1	Introduction . . . . .	179
10.1.1	Motivation . . . . .	179

10.1.2	Related work . . . . .	179
10.1.3	Objective . . . . .	180
10.2	System overview . . . . .	180
10.2.1	Body detection and tracking . . . . .	181
10.2.2	Face detection and tracking . . . . .	182
10.2.3	Face recognition . . . . .	182
10.2.4	Hand detection and tracking . . . . .	183
10.2.5	Gesture recognition . . . . .	184
10.3	Experiments . . . . .	186
10.3.1	Database . . . . .	186
10.3.2	Skin detection . . . . .	186
10.3.3	Face recognition . . . . .	187
10.3.4	Gesture Recognition . . . . .	188
10.4	Conclusion . . . . .	188
10.4.1	Summary . . . . .	188
10.4.2	Future work . . . . .	189
<b>11</b>	<b>Conclusion</b>	<b>191</b>
11.1	Summary . . . . .	191
11.2	Major contributions . . . . .	193
11.3	Outlook . . . . .	194
<b>A</b>	<b>Database overview</b>	<b>197</b>
A.1	Introduction . . . . .	197
A.2	Image databases . . . . .	198
A.2.1	Neckermann Database . . . . .	198
A.2.2	Free Character Database . . . . .	199
A.2.3	AR Face Database . . . . .	200
A.2.4	UMIST Face Database . . . . .	201
A.2.5	VISNET II Face Database . . . . .	202
A.3	Video databases . . . . .	202
A.3.1	VISNET II Cash Machine Database . . . . .	203
A.4	Multimodal databases . . . . .	204
A.4.1	VALID Database . . . . .	204
<b>B</b>	<b>Evaluation methodologies</b>	<b>207</b>
B.1	Confusion matrix . . . . .	207
B.2	Detection evaluation . . . . .	208
B.2.1	Receiver operating characteristic (ROC) curve . . . . .	209
B.2.2	Detection error tradeoff (DET) curve . . . . .	210
B.3	Retrieval evaluation . . . . .	210

---

B.3.1	Precision recall (PR) curves . . . . .	210
B.3.2	Ranks . . . . .	211
B.4	Recognition evaluation . . . . .	211
B.4.1	Recognition and error rate . . . . .	212
B.4.2	Cumulative match characteristic (CMC) curve . . . . .	212
B.5	Segmentation evaluation . . . . .	212
B.6	Clustering evaluation . . . . .	213
B.6.1	Internal measures . . . . .	214
B.6.2	External measures . . . . .	216
<b>Bibliography</b>		<b>217</b>
<b>Publications</b>		<b>234</b>

# List of Figures

1.1	The vision of intelligent machines that can interact with humans from well-known science fiction stories. . . . .	2
1.2	Typical samples of selected looking at people application scenarios. . . . .	4
1.3	Overview of the various dimensions of the looking at people domain with considered (solid) and ignored (dashed) parts. . . . .	5
2.1	Overview of a typical surveillance system with its individual modules. . . . .	13
2.2	Different object representations used for object (human) tracking. . . . .	14
2.3	Illustration of the different biometric traits. . . . .	17
2.4	Illustration of the different multi biometric sources. . . . .	19
2.5	Biometric fusion at different levels. . . . .	20
2.6	Different application scenarios for multimedia retrieval. . . . .	21
2.7	Characterization of the user system interaction for a multimedia search engine. . . . .	25
3.1	Technologies used within the looking at people domain. . . . .	27
3.2	Image enhancement with different point operations. . . . .	29
3.3	Hierarchy of geometric 2D transformations. . . . .	29
3.4	Connected component labeling applied to a binary example. . . . .	31
3.5	Overview of a typical pattern recognition system. . . . .	32
3.6	Illustration of the “curse of dimensionality” for a fixed number of samples and different number of dimensions. . . . .	34
3.7	Illustration of the different feature projection approaches for a two dimensional sample. . . . .	37
3.8	Illustration of a 2D multivariate Gaussian distribution with different covariance matrix types. . . . .	41
3.9	Position of the different fusion approaches in the overall processing chain. . . . .	49
4.1	Illustration of the scale space character of human motion analysis. . . . .	58
4.2	Views as an external criterion that determines the visible interval of all possible scales. . . . .	58

4.3	Questionnaires for describing suspicious persons in a surveillance scenario from two different countries. . . . .	59
4.4	Sample of a generic hierarchical visual description with the low level region tree (left) and the corresponding high level object tree (right). . . . .	61
4.5	Overview of the proposed hierarchical human analysis framework with the channels along the vertical axis and the tasks along the horizontal axis. . . . .	61
4.6	Idealized (left side) and simplified (right side) body anthropometry imposed on the “Vitruvian Man”. . . . .	62
4.7	Body model with holistic representation (red) and component based representation (blue). . . . .	63
4.8	Face anthropometry with most important features and average distances. . . . .	64
4.9	Face model with facial features (green), holistic representation (red) and component based representation (blue). . . . .	65
5.1	Taxonomy of recent body recognition approaches based on the used body representation. . . . .	68
5.2	Overview of the body recognition module. . . . .	71
5.3	Illustration of the top-down body modeling approach. . . . .	72
5.4	Illustration of the bottom-up body modeling approach. . . . .	73
5.5	Illustration of the hybrid body modeling approach. . . . .	75
5.6	Two images with similar color distributions but different spatial structure. . . . .	79
5.7	Extraction of the MPEG-7 scalable color descriptor (SCD). . . . .	81
5.8	Extraction of the MPEG-7 color structure descriptor (CSD). . . . .	82
5.9	Extraction of the MPEG-7 color layout descriptor (CLD). . . . .	83
5.10	Extraction of the grey level cooccurrence matrix (GCM). . . . .	85
5.11	Illustration of the different Tamura features for some samples. . . . .	85
5.12	Illustration of the different Gabor filters used for the extraction of the MPEG-7 homogeneous texture descriptor (HTD). . . . .	87
5.13	Extraction of the MPEG-7 edge histogram descriptor (EHD). . . . .	89
5.14	Samples of the Neckermann Database with a large variety of costumes with different colors, textures and shapes. . . . .	91
5.15	Samples of the Free Character Database with a large variety of costumes reaching from sports over casual to business. . . . .	92
5.16	Visual samples of the different body modeling approaches. . . . .	93
5.17	Objective evaluation of the different extraction approaches for the component based body representation. . . . .	94
6.1	Taxonomies of face detection approaches. . . . .	100
6.2	Overview of the occlusion aware face analysis system. . . . .	104
6.3	Standard set of Haar features. . . . .	105



6.4	Extended set of Haar features. . . . .	105
6.5	Illustration of the major idea behind integral images. . . . .	106
6.6	Classifier cascade as a degenerate tree of several classifiers. . . . .	108
6.7	Overview of the component based face detection method. . . . .	110
6.8	Illustration of the component based face detection method. . . . .	110
6.9	Facial reference graph with size and distance ratios. . . . .	111
6.10	Connected component labeling to split the overall face graph into individual face candidates. . . . .	113
6.11	Cost function with unary (gray), binary (dashed) and wildcard (black) cost terms for different cases. . . . .	114
6.12	Graph matching with detected (solid) and wildcard components (dashed). . .	114
6.13	Comparison of the individual component detectors using ROC curves. . . . .	120
6.14	Comparison of the different face detection approaches based on ROC curves. .	122
6.15	Comparison of the face detection approaches for different occlusions based on visual samples. . . . .	123
6.16	Comparison of the face detection approaches for different views based on visual samples. . . . .	123
6.17	Comparison of the face detection approaches for different sizes based on visual samples. . . . .	124
6.18	Subset of the AR Face Database with occlusions. . . . .	125
6.19	Subset of the VISNET II Face Database with real occlusions. . . . .	125
6.20	Performance of the face/component detection/classification tasks over different variations. . . . .	125
7.1	Face recognition beyond determining the identity of a person. . . . .	129
7.2	Small inter subject variations of faces. . . . .	130
7.3	Large intra subject variations of faces. . . . .	130
7.4	Taxonomies of face recognition approaches. . . . .	131
7.5	Overview of the face recognition system. . . . .	137
7.6	Illustration of the texture template extraction by applying a 2D similarity transformation based on four feature points. . . . .	138
7.7	Illumination compensation methods for different illumination directions. . .	138
7.8	Illustration of the different face representations. . . . .	139
7.9	Multiple expert fusion for the different face representations with and without a priori information. . . . .	142
7.10	Subset of the AR Face Database with occlusions. . . . .	143
7.11	Subset of the VISNET II Face Database with real occlusions. . . . .	143
7.12	Face recognition performance of the different versions over the variations of the AR Face Database. . . . .	145

7.13	Face recognition performance of the different versions over the variations of the VISNET II Face Database. . . . .	146
8.1	Application scenarios for multimodal person search and retrieval. . . . .	150
8.2	Overview of the system for multimodal person search and retrieval. . . . .	151
8.3	Web based user interface (WUI) of the multimodal person search application. . . . .	152
8.4	Sample of the VALID Database showing an individual in 5 different environments. . . . .	156
8.5	Performance of the different modalities and retrieval approaches over the number of iterations for a result set size of 45 items. . . . .	158
8.6	Performance of the different modalities and retrieval paradigms over the result set size for 3 iterations. . . . .	158
9.1	Overview of the system for visual person search based on a human visual thesaurus. . . . .	164
9.2	Color distribution and predicted clusters and ground truth classes within the Lab color space. . . . .	166
9.3	Determine the optimal number of clusters by finding the knee in cohesion vs. number of cluster curves. . . . .	168
9.4	Query interface with individual thesauri for the different body parts and some selected categories (red). . . . .	170
9.5	Illustration of the neighbor selection for the range queries. . . . .	171
9.6	Result sets for the individual thesauri and after the logical query combination. . . . .	173
9.7	Internal and external quality assessment of the visual face thesaurus for different clustering methods. . . . .	175
9.8	Internal and external quality assessment of the visual body thesaurus for different color spaces. . . . .	175
9.9	Internal and external quality assessment of the visual body thesaurus for different clustering methods. . . . .	176
10.1	Overview of the system for personalized human computer interaction. . . . .	181
10.2	Intermediate results of the individual tracking modules. . . . .	181
10.3	Distribution of skin color within the YCbCr space for all users (red) and three individual users (green). . . . .	183
10.4	Samples for each of the users considered within the database. . . . .	186
10.5	Comparison of different skin detection approaches (general, hybrid, specific) based on visual examples. . . . .	187
A.1	Visual samples of the Neckermann Database. . . . .	200
A.2	Visual samples of the Free Character Database. . . . .	200
A.3	Visual samples of the AR Face Database. . . . .	201

---

A.4	Visual samples of the UMIST Face Database. . . . .	201
A.5	Visual samples of the VISNET II Face Database. . . . .	202
A.6	Visual samples of the VISNET II Cash Machine Database. . . . .	204
A.7	Visual samples of the VALID Database. . . . .	205
B.1	Confusion matrix as a common tool for unsupervised and supervised learning.	208
B.2	Common evaluation curves for unary/binary classification evaluation such as detection and retrieval tasks. . . . .	210
B.3	Cumulative match characteristic curve for the evaluation of recognition tasks.	213
B.4	Illustration of the segmentation evaluation process. . . . .	213
B.5	Clustering evaluation via correlation. . . . .	214
B.6	Estimating the optimal number of clusters via the sum of squared errors (SSE).	215
B.7	Illustration of the silhouette coefficient for a single data point and two clusters.	216



# List of Tables

2.1	Comparison of different biometric traits based on different characteristics. . .	18
4.1	Definition of the rectangular patches for the holistic and the component based representation relative to the anthropometric face region. . . . .	65
5.1	Overview of recent body recognition approaches. . . . .	69
5.2	Overview of the considered visual low level features grouped into color and texture types as well as standard (MPEG-7) and non-standard ones. . . . .	77
5.3	Body matching performance of the individual features across the different body parts. . . . .	95
5.4	Body matching performance of different feature fusion methods across the different body parts. . . . .	96
5.5	Body matching performance of different part fusion methods across the different features. . . . .	96
6.1	Comparison of selected face detection approaches. . . . .	102
6.2	Optimized parameter set for the component based face detection approach. .	117
6.3	Overview of the component detectors (upper part) and holistic face detectors (lower part). . . . .	118
6.4	Overview of available databases for face detection grouped into considered (upper part) and discarded (lower part) ones. . . . .	119
6.5	Overall performance of the different component/face detection/classification tasks for the AR Face and the VISNET II Face Database. . . . .	126
7.1	Overview of the different representations with interesting characteristics including number of experts, dimensionality, correlation and fusion possibility.	140
7.2	Comparison of selected databases suitable for face recognition. . . . .	143
7.3	Face recognition performance of the versions over the different databases. .	144
8.1	Performance of the different modalities (audio, video, multimodal) and retrieval approaches (QBE, SG, SVM) based on two measures after convergence.	157
10.1	Possible labels of the different stage 1 classifiers. . . . .	185

10.2	Comparison of different gesture recognition approaches (general, user specific) based on the recognition rate. . . . .	188
A.1	Overview of available image databases for looking at people research with considered scenario and channel. . . . .	199
A.2	Overview of available video databases for looking at people research with considered scenario and channel. . . . .	203
A.3	Overview of available multimodal (audiovisual) databases for looking at people research with considered scenario and channel. . . . .	204

# Acronyms

AC average color

AGG agglomerative clustering

AV audiovisual

BC Bayesian classifier

BIC Bayesian information criterion

BIO biometrics

BM Bayesian model

CCV color coherence vector

CH color histogram

CLD color layout descriptor

CMC cumulative match characteristic

CM color moments

CS color spatiogram

DCT discrete cosine transform

DET detection error tradeoff

EHD edge histogram descriptor

ER error rate

FCM fuzzy c-means

FNR false negative rate

FPR false positive rate

F f-measure

GCM grayscale cooccurrence matrix

GMM Gaussian mixture model

GT ground truth

HCI human computer interaction

HTD homogeneous texture descriptor

H head

ICA independent component analysis

IM intensity moments

KM k-means

kNN k-nearest neighbor

LDA linear discriminant analysis

MAP maximum a posteriori

MDM minimum distance to means

MFCC Mel frequency cepstrum coefficients

ML maximum likelihood

PCA principal component analysis

PR precision/recall

P precision

ROC receiver operating characteristic

RR recognition rate

R recall

SCD scalable color descriptor

SG single Gaussian model

SRT smart room technologies

SUR surveillance

SVM support vector machine



TF Tamura features

TNR true negative rate

TPR true positive rate

U upper body

W whole body



# Chapter 1

## Introduction

*The future lies in designing and selling computers that people don't realize are computers at all.*

---

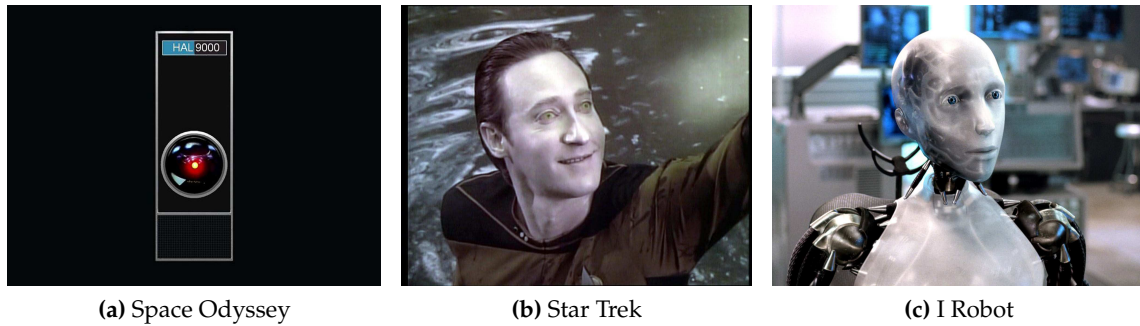
Adam Osborne

### 1.1 Motivation

The last two decades have experienced a rapid evolution with respect to hardware development, especially in the fields of data acquisition, processing, communication and storage. *Sensors* have become smaller and cheaper and can acquire data (image, video and audio) at higher speed and with better quality. *Processors* have become more powerful and much smaller which makes them usable for highly complex coding or analysis tasks in both stationary and portable devices. The development of faster and more flexible *communication* technologies and the invention of the world wide web (WWW) have lead to a time where information is easily accessible and more important than ever before. With *storage* devices becoming much smaller, providing more capacity and being more robust, huge amounts of multimedia data can be captured and stored.

All this has lead to a time where computers and the internet have become an important aspect in our everyday life. We use this technology to communicate, study, shop and entertain ourselves. The vision of the future is to embed this computing technology into our home, transportation, and working environments. In this vision of the future, often referred to as *ubiquitous computing* [Weiser, 1991] or *ambient intelligence* [Aarts, 2005], machines will be able to monitor their environments and analyze the behavior of the inhabitants. The ultimate goal is to develop smart machines that are aware of humans and can assist them if required.

While the hardware is the foundation for these intelligent machines, software that is able to analyze the acquired data with respect to humans is also required. This human centered analysis is commonly referred to as *sensing people* in the case of audiovisual data and *looking*



**Figure 1.1:** The vision of intelligent machines that can interact with humans from well-known science fiction stories. The humanoid look of them already implies to have human-like behavior.

*at people* in the case of visual only data. Although audio information may be interesting for some applications, visual information is considered to be the most important [Essa, 1999].

The following sections provide an overview of the looking at people research field (section 1.2), define the overall objectives (section 1.3), summarize the key contributions (section 1.4) and describe the overall structure (section 1.5) of this thesis.

## 1.2 Looking at people

The major goal of the looking at people research domain is to build machines that can interact with persons in an environment [Essa, 1999]. Traditionally science fiction writers have described the ultimate goals through their characters, which are shown in figure 1.1: HAL in 2001 Space Odyssey, Commander Data in Star Trek The Next Generation, and NS5 in I Robot. The exact ability of the machine depends largely on its task. In general, one can distinguish between machines with and without the ability to sense humans. The former group contains intelligent machines that are able to work with us, support our needs and be our helpers. For these personal assistants the ability to sense people is essential. In contrast, the latter group contains machines which are not aware of humans in their environment such as industrial vision systems aimed at extracting defects on an assembly line or computers used for email writing and text processing.

The key technical goal of the looking at people domain is to determine the context with respect to humans. This involves answering several questions such as:

- Are there persons?
- Where are the persons?
- How many persons are there?
- Where are the persons moving?
- What are the persons doing?

- Who are the persons?
- To which group do they belong?

These questions are directly related to different analysis tasks, which are necessary to answer the questions mentioned above. In general the following tasks can be distinguished:

**Detection:** The goal of this task is to determine the presence of humans in an environment, their number and location. Depending on the considered information several detection tasks can be distinguished including object, body detection, face and hand detection.

**Tracking:** While detection tries to find humans within images or individual video frames, the goal of tracking is to establish the correspondence of the human or individual parts between consecutive frames. Again several tracking tasks can be distinguished including object, body, face and hand tracking.

**Recognition:** This actually corresponds to a group of tasks which extract additional information regarding the detected and tracked person. *Person recognition* is related to the identity of a person and usually considers face and gait as primary sources of information. *Activity recognition* analyzes the global motion of persons within an environment to interpret their activities. On the other hand, *behavior recognition* considers the local motion of individual body parts to interpret the behavior or intention of a person.

The combination of the previously described tasks leads to a large variety of applications (some samples are shown in figure 1.2) [Essa, 1999]:

**Human computer interaction (HCI):** The goal is to develop machines that can interact with us in a similar way as we interact with each other, using gestures and speech. Such systems are able to know where someone is looking, estimate pointing directions, interpret gestures and their intention. These interfaces are an integral part toward more human centered interfaces, which are applicable in areas where traditional interfaces (keyboard, mouse) are not efficient enough.

**Smart room technologies (SRT):** Intelligent machines can be installed within rooms to detect people in it, identify them and interpret their behaviour. These rooms can be used to monitor children, senior citizen and handicapped people and provide assistance if needed. Furthermore, this technology can be integrated in seminar rooms to provide automatic indexing and summarization of meetings, discussions and presentations. Such smart room technologies could become an integral part of our daily activities.

**Surveillance and security:** This is one of the more traditional applications, which involves the verification of a persons identity for access control and the interpretation of human



**Figure 1.2:** Typical samples of selected looking at people application scenarios. Depending on the scenario the focus of the human analysis may be quite different.

actions within an environment for active surveillance. The major idea is to support security personnel in their laborious work through detecting events, selecting interesting footage and providing visual enhancements.

**Entertainment and education:** Recently these two areas have experienced a rapid growth. Non-invasive tracking and interpretation of human behaviour could revolutionize these two areas. An intelligent tutor could judge the actions and moods of the students and react accordingly. Similarly, systems that can analyze human motions could be used in sports training and dance teaching.

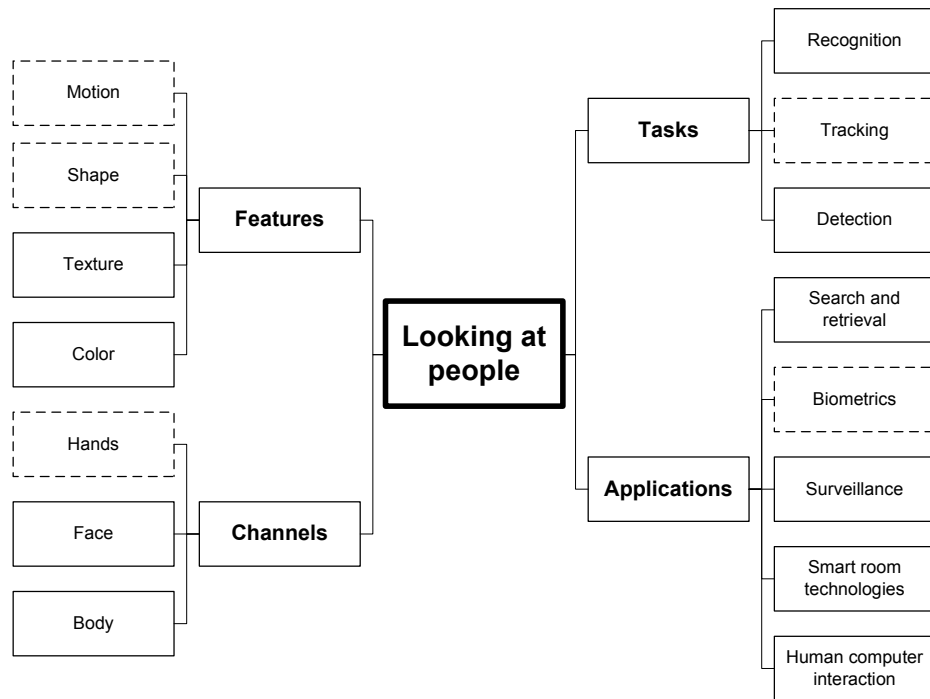
**Video conferencing:** The analysis of facial motion can be used for animation and model based coding. With these methods low bitrate video phones can be built. Furthermore, synthetic avatars can be used to hide the identity of the speakers.

**Digital libraries and annotation:** With the rapid increase of available multimedia data, systems for efficient indexing and retrieval gain more and more importance. Since a large amount of this data contains humans, detecting, tracking and recognizing them plays a significant role in the automatic annotation of this data.

**Human augmentation and wearable computing:** The idea is to develop systems that can interpret the activities of humans in an environment and provide assistance to impaired people by translating the missing communication modality into another modality that can be perceived. Furthermore, this technology can be used to provide more efficient ways of communication.

### 1.3 Objectives

The major objective of this thesis is to develop a hierarchical framework for the visual analysis of humans. The goal is to mimic the behavior of the human visual system (HVS), that extracts information at various levels and chooses the appropriate level depending on the



**Figure 1.3:** Overview of the various dimensions of the looking at people domain with considered (solid) and ignored (dashed) parts. Not all of the developed applications are described within this thesis.

current task or condition [Pers et al., 2003]. The work focuses on analyzing the appearance (color, texture) of the body and the face, but it can be assumed that most of the ideas are applicable to other channels (e.g. hands) and characteristics (e.g. shape, motion). Based on that, methods for the visual detection and recognition of humans are developed and extended. Special emphasis is laid on the robustness regarding typical challenges such as varying illuminations, different views and partial occlusions. The developed modules are used within several applications to assess their performance in real world scenarios.

The scope of this thesis is illustrated in figure 1.3 which shows the various dimensions of the looking at people domain with the considered (solid) and ignored (dashed) parts.

Regarding face detection the objective is to study existing approaches and analyze their robustness with respect to the different challenges. Furthermore, a component based approach will be developed that detects faces by considering the appearance and spatial relationship between facial components.

For body recognition a hierarchical approach will be developed that describes the appearance of the human body hierarchically. Therefore, different representations and visual features is considered and compared to each other. Furthermore, the fusion of body parts and features are explored.

Regarding face recognition the goal is to develop an appearance based approach that describes the appearance of a face with different representations. In order to address specifically the challenge of partial occlusions, different fusion methods with and without addi-

tional occlusion information are explored.

Finally, the developed framework will be used for a variety of applications to prove its versatility. The chosen applications will consider human computer interaction, and multimedia search and retrieval.

Within another application the developed framework is applied to the audiovisual search of humans within multimedia documents. The objective of this work is to combine two biometric traits of a person, namely face and speech, with each other to resolve ambiguities within the individual modalities. Furthermore, this work will explore the use of typical content based image retrieval techniques such as query by example and relevance feedback for multimodal person search.

The whole framework will be applied to the hierarchical visual search of humans within images. The objective of this work is to explore the query by visual thesaurus paradigm as an intuitive and efficient alternative to classical query by example. Therefore, the original visual thesaurus approach will be extended towards a universal human visual thesaurus.

Finally, the framework will be used for personalized human computer interaction where the objective is to develop a system that integrates the recognition of the identity and the behaviour of a person and explore how the exchange of information between these different analysis parts can improve the overall performance.

## 1.4 Contributions

This thesis studies different aspects of the looking at people domain and makes the following major contributions:

- Comprehensive review of the looking at people research domain with most important analysis tasks and application scenarios
- Condensed overview of the fundamental techniques reaching from image processing over machine learning to information fusion
- Proposal of an original hierarchical framework for human analysis that supports different application scenarios and environmental conditions
- Novel component based face detection approach that combines tools from statistical and structural pattern recognition domain and robustly detects faces in the presence of partial occlusions
- Extension of appearance based face recognition approach that utilizes a priori occlusion information for improved performance in the presence of occlusions
- Novel appearance based body recognition approach that considers different body representations and a large variety of visual features for describing the appearance of the human body



- Original system for multimodal person search that combines audiovisual analysis with relevance feedback for efficient search and retrieval
- Adaptation of the query by visual thesaurus paradigm towards a human visual thesaurus for intuitive and efficient visual person search

Parts of this thesis have been published in several articles in international conferences and journals. A detailed publication list is provided on page 234.

## 1.5 Organization

The rest of this thesis is organized into several chapters which are shortly summarized below:

*Chapter 2* provides a comprehensive review of the most important application scenarios, already mentioned in section 1.2, to illustrate the diversity of interests and approaches in the looking at people domain. The goal is not to discuss all available methods in detail, but provide a review of the individual goals, challenges and conditions of these applications scenarios. Furthermore, it will be shown that each them concentrates on specific features and characteristics of humans.

The looking at people domain combines techniques from several fields including image processing, computer vision, machine learning and information fusion. *Chapter 3* summarizes the fundamental techniques that are used within this work and provides links to the chapters where they are considered. The intention is to provide a condensed overview of the required tools along with the most important references.

Following the major objective, *chapter 4* describes the proposed hierarchical framework for human analysis and discusses analogies to the classical scale space theory and focus of attention. Furthermore, it provides an overview of the overall system and links to the chapters that describe the individual modules. Finally, it describes the hierarchical human model which is inspired by the human visual perception and reviews anthropometrical models that form the basis for the visual analysis.

*Chapter 5* describes the developed body recognition module which considers a holistic and a component based representation of the human body. Several color and texture features with different characteristics are considered for describing the appearance of the individual body parts. Furthermore, the fusion of complementary information in form of body parts and visual features is explored. The experiments provide interesting insights on the choice of suitable features and representations, the fusion of complementary information and the tradeoff between the complexity of representation and description.

*Chapter 6* explores holistic and component based approaches for face detection and their limits with respect to the various challenges. It describes a novel component based face detection approach that combines techniques from the statistical and structural pattern recognition domain. It allows to detect faces even under partial occlusions and provides addi-

tional information such as the presence and location about them. Extensive experiments show the large performance improvement in comparison to the holistic approach and explore the limits of both approaches.

*Chapter 7* describes the developed appearance based face recognition module that considers a holistic, a component based and a lophoscopic representation of the face. Given the additional occlusion information provided by the face detection module, an adaptive fusion method is proposed that considers only non-occluded parts of the face for the recognition. The experiments provide an in-depth analysis of the different representations under a large variety of occlusions and show that the adaptive fusion of several components improves the performance considerably.

Based on the developed framework with its individual modules, several applications have been developed. Depending on the application certain parts of the overall framework are integrated.

In *chapter 8* a novel system for content based search of persons within multimedia documents is described. It combines the analysis of biometric traits (face, voice) with retrieval techniques such as query by example and relevance feedback. This leads to a very efficient search tool that significantly reduces the user effort in contrast to the manual search. Within the experiments various aspects of the system such as multimodal fusion and relevance feedback are explored to find the optimal tradeoff between retrieval performance and required user interaction.

*Chapter 9* describes an original system for visual person search within images based on the query by visual thesaurus paradigm. The original approach is extended towards a human visual thesaurus that summarizes present humans at different levels and provides an intuitive and efficient query interface. For the creation of the visual thesaurus different color spaces and clustering methods have been considered. The conducted experiments assess the quality of the visual thesaurus and provide a summary of the subjective impressions on the use of this query paradigm.

*Chapter 10* describes a system for personalized human computer interaction that combines face and gesture recognition in a tightly integrated manner. It explores ways to improve the performance of the overall system by exchanging information between the usually independently working modules. The experiments demonstrate the improved performance for an intelligent cash machine scenario.

*Chapter 11* concludes this thesis by summarizing the achievements and drawing general conclusions regarding the visual analysis of humans in images and videos. Based on these findings, an outlook towards a universal human analysis framework is given and open challenges are discussed.

*Appendix A* gives an overview about available and created datasets for looking at people research. It considers typical characteristics such as media type, number of files, included variations and discusses the use for certain tasks or applications. In addition, it provides a detailed description of the datasets that have been used and referenced throughout the

thesis.

*Appendix B* summarizes the considered evaluation methodologies for the different tasks and applications, to complement the specifics described within the individual chapters. It includes methodologies for the evaluation of detection, segmentation, recognition and retrieval and clustering problems.



## Chapter 2

# Looking at people and related fields

### 2.1 Introduction

As already mentioned in section 1.2 the looking at people domain consists of a large number of research fields with diverse interests and characteristics. The goal of this chapter is to provide a comprehensive review of the most important fields including surveillance, biometrics, and search and retrieval and discuss their overlap with the looking at people domain.

### 2.2 Surveillance

The idea of smart surveillance systems is to assist human observers in monitoring environments [Regazzoni et al., 2001]. Since the human perception and reasoning capabilities are quite limited, the goal is to extend these capabilities by audiovisual analysis.

More specifically smart surveillance systems provide the following advantages over traditional surveillance systems [Hampapur et al., 2003]:

- Ability to prevent incidents through real time alerting for suspicious behavior
- Enhanced forensic capabilities through content based search and retrieval
- Situation awareness through joint analysis of location, identity and activity of objects within an environment

This functionality can be used in a large variety of scenarios [Regazzoni et al., 2001; Hu et al., 2004], including:

- Safety in transportation such as railway stations, underground stations, airports, motorways and maritime environments
- Quality control in industrial applications such as nuclear power plants and industrial processing cycles

- Improved security in public spaces such as banks, supermarkets and facilities
- Crowd flux statistics and congestion analysis for shopping malls, parking lots and highways
- Military surveillance for strategic infrastructure and battlefields

### 2.2.1 History and trends

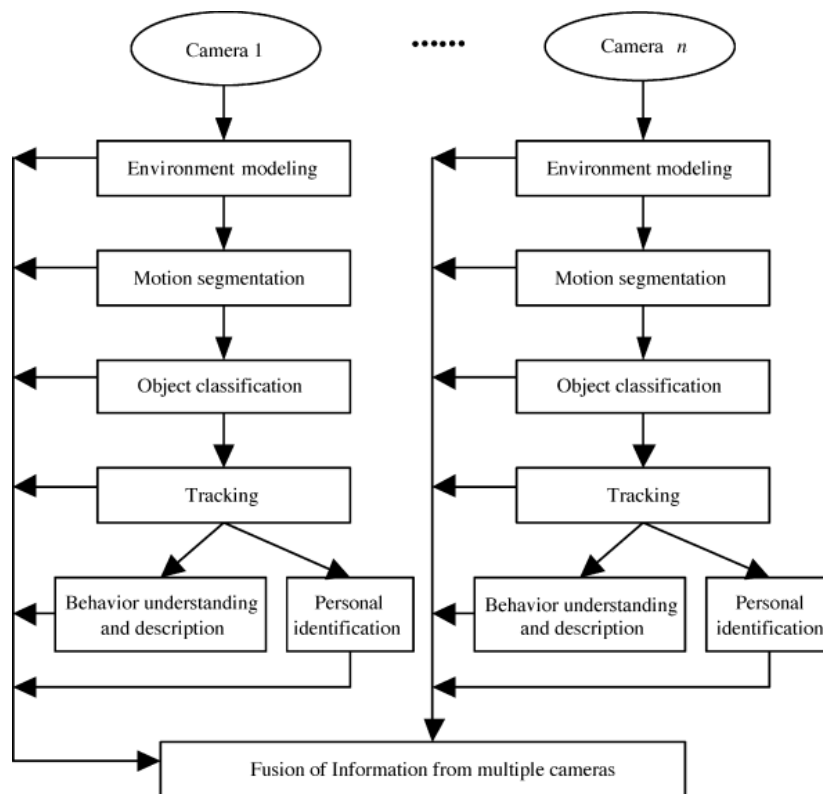
From a technological perspective video surveillance systems can be classified into three successive generations that follow roughly the evolution of communication, processing and storage technology [Regazzoni et al., 2001]:

**First generation surveillance systems (1GSS)** (1960–1980) basically serve as an extension of the human perception in the spatial sense. Analog cameras capture visual information from multiple remote locations and transfer it to a single control room where a human operator analyzes them on several screens. 1GSS are solely based on analog video capturing, transmission and processing technology. The advantage of these systems is given by the telepresence of a human operator with respect to multiple remote places. The major disadvantages of these systems are caused by the analog technology in terms of large bandwidth, low video quality and large storage requirements.

**Second generation surveillance systems (2GSS)** (1980–2000) benefited from the early advances in digital video communications (acquisition, transmission and coding) that led to decreasing bandwidth and storage requirements and improved video quality. Furthermore, automatic video analysis techniques were used to provide assistance to the human operator. This includes real time detection and tracking of objects as well as activity recognition and event detection for prefiltering of the incoming video data. In this way, 2GSS support the simultaneous monitoring of a larger number of remote places and longer storage of interesting video data for forensic analysis.

**Third generation surveillance systems (3GSS)** (from 2000) are targeted towards “full digital” solutions starting from the acquisition up to the presentation of visual information. Therefore, they take advantage of the progress in computing, networking and human computer interaction technology. This will lead to large surveillance systems consisting of smart sensors (cameras, microphones, range) that communicate over heterogeneous networks (WLAN, GPRS) and can be accessed with various types of terminals (PC, PDA). At the same time more sophisticated analysis techniques will provide better assistance to the human operators.

Currently the trend goes towards the development of multimedia surveillance systems, that further extend the capabilities of smart surveillance systems [Cucchiara, 2005]. One idea is to utilize *different camera types* (fixed, pan tilt zoom, omnidirectional, stereo, multispectral) to obtain different views of a scene. Furthermore, *additional modalities* (audio, range,



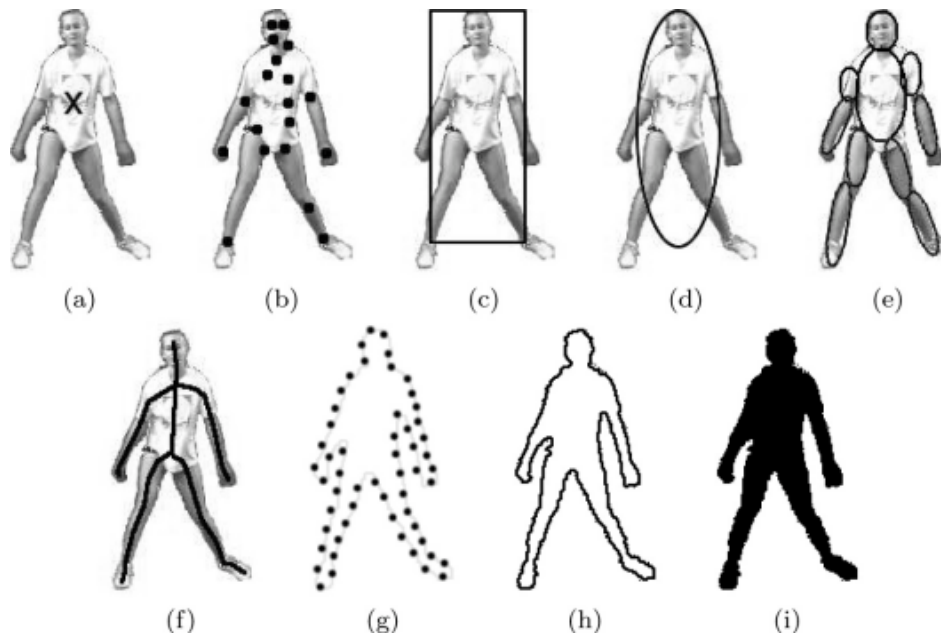
**Figure 2.1:** Overview of a typical surveillance system with its individual modules [Hu et al., 2004]. Most of the modules apart from the behavior understanding and the personal identification consider humans usually as generic objects.

radar) are used to improve the visual modality. Finally, the trend goes towards extensible *distributed sensor networks* that can monitor large areas without any complex calibration. Besides the analysis aspects, data management and storage of the extracted information are of increasing importance. This involves the use of content description and querying standards such as MPEG-7 [Manjunath et al., 2002].

### 2.2.2 Modules

Although the objectives of a surveillance system may differ depending on the application scenario, there are several modules that are present in any system (see figure 2.1) [Hu et al., 2004]:

**Object detection:** The goal of the object detection stage is to detect and localize objects of interest within an environment [Hu et al., 2004]. Within surveillance systems the basic idea is usually to segment foreground objects from the background. This step is further subdivided into environment modeling, motion segmentation and object classification. Within the *environment modeling* step a model of the background is built and updated. For fixed cameras the idea is simply to construct a pixel wise background model and update it for dynamic environments. For pure translation cameras,



**Figure 2.2:** Different object representations used for object (human) tracking [Yilmaz et al., 2006]: (a) centroids, (b) features, (c) boxes, (d) ellipses, (e) parts, (f) stick figures, (g) contours, (h) silhouettes, and (i) regions. The large variety of models shows already the diversity of human analysis approaches.

a panorama image can be built by patching individual background images together. For mobile cameras motion compensation can be used to construct temporary background images. Based on the extracted environment model, the *motion segmentation* step aims at detecting regions that correspond to foreground objects. Approaches are based on background subtraction [Haritaoglu et al., 2000; McKenna et al., 2000; Stauffer and Grimson, 1999], temporal differencing and optical flow [Meyer et al., 1998; Barron et al., 1994]. Since the extracted foreground regions may correspond to different objects, an *object classification* step is required to recognize different object types (persons, cars). Traditionally approaches are based either on shape [Lipton et al., 1998; Collins et al., 2000] or motion characteristics [Cutler and Davis, 2000; Lipton, 1999].

**Object tracking:** Object tracking establishes the temporal correspondence of detected objects between consecutive frames [Hu et al., 2004]. Tracking approaches can be divided into 4 categories, depending on the object representation which is used for the tracking (see figure 2.2 for some examples). *Region based* approaches are matching the detected regions between consecutive frames based on shape, motion, color and texture features [McKenna et al., 2000]. Often multiple abstraction levels (region, object) are used to handle merges, splits and occlusions of regions. *Contour based* approaches track objects by representing their outlines using a parametric description and updating it dynamically between consecutive frames [Paragios and Deriche, 2000; Peterfreund, 2000; Isard and Blake, 1996]. *Feature based* approaches perform tracking of objects by



detecting features (edges, corners) within an object and matching these features between consecutive images. Depending on how these features are combined global, local and dependence-graph based approaches can be further distinguished. Finally, *model based* approaches utilize a priori knowledge about the object type, to match projected object models to the individual images. Depending on the object type non rigid object tracking (e.g. humans) and rigid object tracking (e.g. cars) can be distinguished. The general idea for both categories is following an analysis by synthesis approach, that maps a predicted model onto the image plane and compares it to the current image.

**Behavior understanding:** After successful tracking objects between consecutive frames, the understanding of object behavior is of major interest [Hu et al., 2004]. Although the term *behavior recognition* is not very well defined, it usually refers to the interpretation of human motion patterns. On the other hand, *activity recognition* considers general object motion within an environment usually based on the objects trajectory. Behavior recognition involves the analysis and recognition of motion patterns and the extraction of high level descriptions of actions and interactions. Existing methods for behavior understanding are based on dynamic time warping (DTW), finite state machines (FSM), hidden Markov model (HMM), time delay neural networks (TDNN), and self organizing maps (SOM).

**Person recognition:** Beside behavior understanding, person recognition is of increasing importance for surveillance applications [Hu et al., 2004]. Since video surveillance can be seen as an uncooperative scenario, face and gait are considered as suitable biometric traits (see section 2.3). *Face recognition* is based on the facial appearance of a person and requires a frontal, close up view. Since this is in contrast to the wide angle view of static surveillance cameras, moving cameras are used for face recognition. On the other hand, *gait recognition* is based on the unique walking characteristics of a person and works for wide angle views. To improve the robustness and reliability of person recognition, the different biometric traits can be fused.

**Multiple camera fusion:** While the modules described above are used within a single camera surveillance system, multiple camera fusion provides larger field of view and different views of an environment. While multiple camera views provide more information that can be used to handle occlusions or other ambiguities within a single camera view, they also raise some additional questions. The *camera installation* has a large influence on the real time performance and the costs of a surveillance system. While a lack of cameras may cause blind spots or reduce the reliability, redundant cameras increase both the processing time and the installation cost. *Camera calibration* is required to relate the different camera views to each other and to the real world. The transformation between different camera views can be either done offline by using a set of pre-

defined points or online using spatio-temporal information. *Object matching* involves establishing the correspondence between objects within different views. While appearance based approaches can be used for overlapping and non-overlapping views, geometry based approaches are only suitable for overlapping views. *Camera switching* is used to switch between available cameras, when an object moves out of the field of view of a camera or a camera can not provide a good view of the object. *Data fusion* between the different camera views is important for continuous object tracking and occlusion handling.

### 2.2.3 Discussion

One of the keys to develop smart surveillance systems is the visual analysis of objects within an environment. Although other sensors (audio, range, infrared) are also used the visual information is by far the most important. Beside others, humans are the most interesting objects within this application scenario. The goal is usually to detect and track them throughout the environment and interpret their activities and behavior. This usually requires a visual analysis at object level with respect to the environment and other objects.

## 2.3 Biometrics

Establishing the identity of a person is commonly referred to as person recognition. In general there are three different ways to achieve this, based on [Jain et al., 2004]:

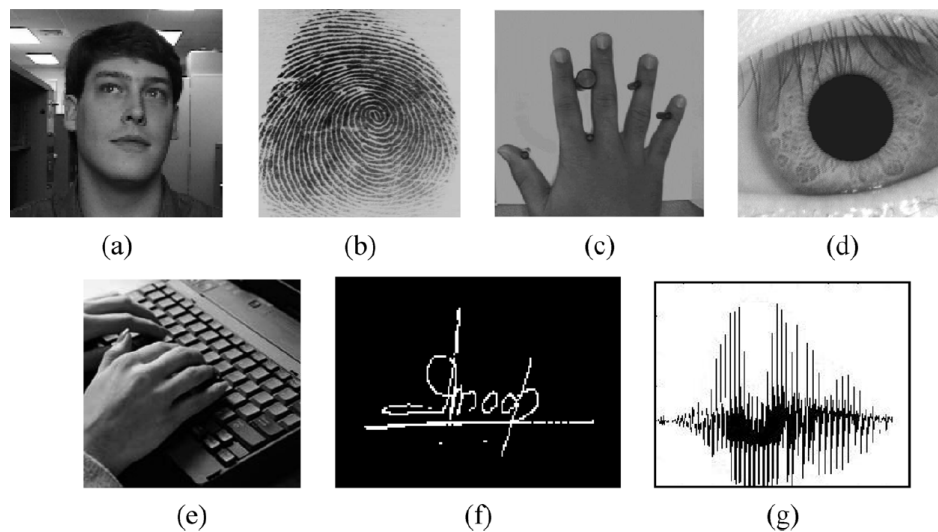
**Possession:** Something you carry (e.g. key, ID card)

**Knowledge:** Something you know (e.g. password, PIN)

**Traits:** Something you are (e.g. face, voice)

The major problem with possession based systems is that keys and ID cards can be easily misplaced, shared or stolen. Knowledge based systems such as passwords can be easily hacked or guessed. On the other hand physical and behavioral traits, called *biometrics*, offer a natural and reliable solution, since they provide a strong and permanent link between a person and his identity.

In order to support different applications biometric systems typically provide several functionalities [Nandakumar, 2005]. During the *enrollment* a model for the user is built and stored in the database. The recognition itself supports several scenarios. In the *verification* scenario the user claims an identity and the system either accepts or rejects the claim. Within the *identification* scenario the system identifies the person assuming that it has been enrolled before (closed set scenario). On the other hand, in the *watchlist* (screening) scenario the system determines whether a person belongs to the group of enrolled identities (open set scenario).



**Figure 2.3:** Illustration of the different biometric traits [Jain et al., 2006]: (a) face, (b) fingerprint, (c) hand geometry, (d) iris, (e) keystroke, and (g) voice. Except for face and voice all other biometric traits require user cooperation.

A biometric system typically involves four major components including sensor, feature extractor, matcher and decision [Nandakumar, 2005]. The *sensor* acquires the biometric data from an individual. The *feature extractor* creates a compact and robust representation of the biometric trait. During the recognition a *matcher* compares the actual feature to a previously extracted template and determines the degree of similarity (dissimilarity) between them. Finally, the *decision* module decides upon the identity of the person based on the score provided by the matcher.

### 2.3.1 Biometric traits

Several physical and behavioral traits can be used for biometric recognition, as it is shown in figure 2.3. Physical or anatomical traits include face, fingerprint, iris, palmprint, hand geometry and ear shape while behavioral traits are gait, signature and keystroke. Voice can be considered as a physical or behavioral trait depending on the characteristics of the voice that are analyzed.

Each biometric trait has its strengths and weaknesses which are summarized in table 2.1 and the choice depends largely on the application [Jain et al., 2006]. No single biometric trait is expected to meet all of the following requirements:

**Universality:** Is it available across all people?

**Distinctiveness:** How well people can be distinguished?

**Permanence:** How permanent is the trait?

**Collectable:** Can it be easily acquired and quantified?

Factors →							
Biometric identifier ↓	Universality	Distinctiveness	Permanence	Collectable	Performance	Acceptability	Circumvention
Face	H	H	M	H	L	H	H
Fingerprint	M	H	H	M	H	M	M
Hand geometry	M	M	M	H	M	M	M
Iris	H	H	H	M	H	L	L
Keystroke	L	L	L	M	L	M	M
Signature	L	L	L	H	L	H	H
Voice	M	L	L	M	L	H	H

**Table 2.1:** Comparison of different biometric traits based on different characteristics [Jain et al., 2006]. The ratings are based on a 3-level scale including high (H), middle (M), and low (L).

**Performance:** What is the accuracy and the speed?

**Acceptability:** Is it well accepted by the people?

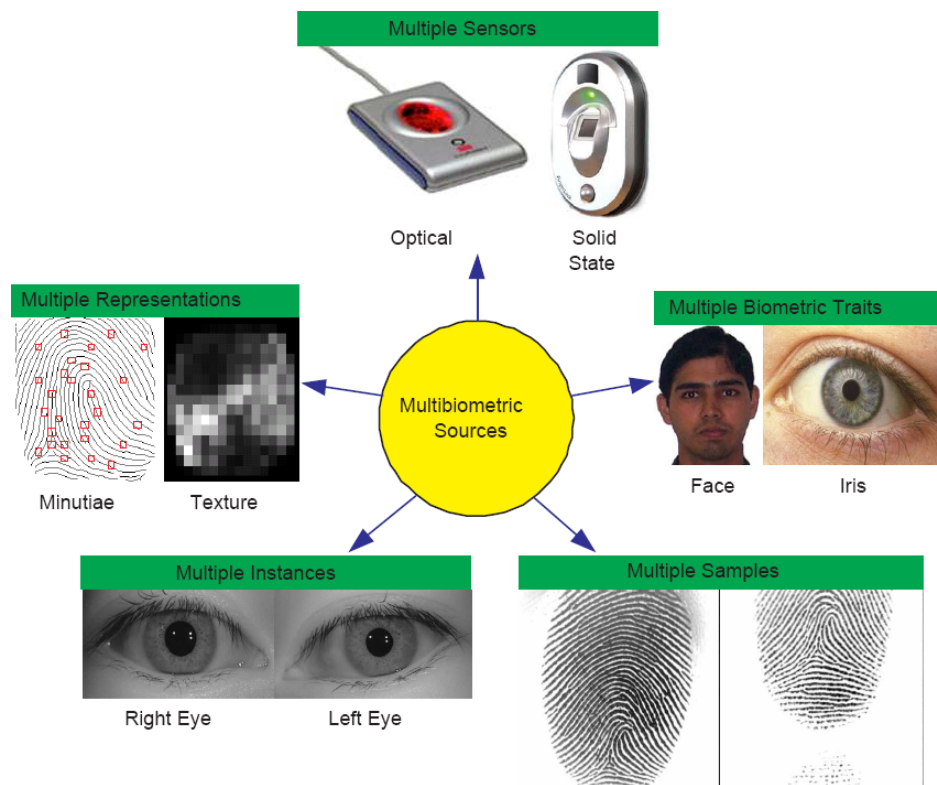
**Circumvention:** Is it foolproof?

More recently *soft biometrics* have been proposed by Jain et al. [2005b] as additional traits that can be combined with other biometrics to improve their robustness. These soft biometric traits are defined as characteristics that provide some information about an individual, but lack the distinctiveness and permanence to differentiate between any two individuals alone. Soft biometrics can be either continuous, e.g. height, weight, or discrete, e.g. gender, eye color and ethnicity.

### 2.3.2 Multi biometrics

Biometric systems in operational scenarios have to cope with a lot of challenges [Jain et al., 2006] including:

- Noise in sensed data
- Intraclass variations
- Distinctiveness
- Non-universality
- Spoof attacks

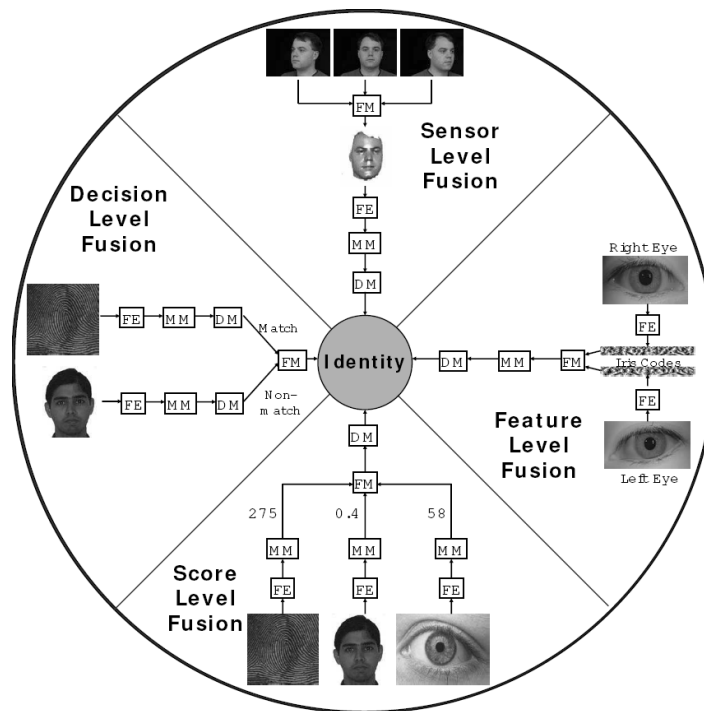


**Figure 2.4:** Illustration of the different multi biometric sources [Nandakumar, 2008]. Only systems that consider multiple biometric traits are called multimodal biometric systems.

Some of these challenges can be overcome by combining multiple biometrics into a multi-biometric system, since the different biometric sources usually compensate for the inherent limitations of the other sources [Hong et al., 1999]. More specifically multi biometrics systems offer the following advantages:

- Combining the evidence of multiple sources can improve the overall accuracy
- Different traits reduce the non-universality problem by providing alternatives
- Multiple traits provide flexibility within different application scenarios
- The availability of the different sources allows to reduce the noise effects
- Large databases can be searched in a more efficient way

Different sources of biometric information can be distinguished for multi biometric systems (see figure 2.4) including: *multiple biometric traits* (e.g. face and voice) and a single biometric trait captured by *multiple sensors* (e.g. optical and range sensor for face), described by *multiple representations* (e.g. texture and minutiae of fingerprint), with *multiple instances* (left and right iris), and with *multiple samples* (e.g. two face samples over time). Systems combining multiple biometric traits are usually called multimodal biometric systems.



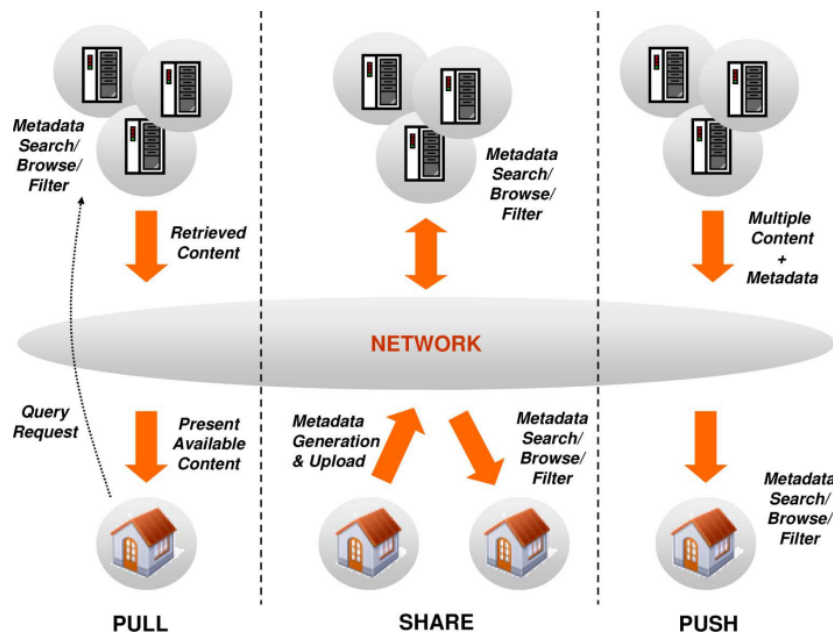
**Figure 2.5:** Biometric fusion at different levels [Nandakumar, 2008]. The different levels are defined according to position of the fusion module (FM) in relation to the overall processing chain of a biometric system which consists of feature extraction (FE), matching module (MM) and a decision module (DM).

The combination of multiple biometric sources can be either serial or parallel. Both architectures have their own advantages and disadvantages. Within the *serial* architecture the output of a biometric trait is usually used to narrow down the number of possible identities before the next biometric trait is used. In general, the advantages are a faster processing speed and a higher user convenience since only a subset of the traits might be used. On the other hand, the *parallel* architecture typically achieves a higher performance since they combine the evidence of multiple sources directly.

The fusion within a multi-biometric system can take place at different levels (sensor, feature, score, decision) depending on the type of information that is fused. Each of the levels (shown in figure 2.5) corresponds to the output of one module that constitutes a typical biometric system.

### 2.3.3 Discussion

Biometric traits usually correspond to the analysis of different modalities (audio, image, video, 3D). The most important visual traits are face and gait since they can be acquired by a standard video camera which is the most universal and accepted type of sensor. These traits are based on the visual analysis of humans at face and body level.



**Figure 2.6:** Different application scenarios for multimedia retrieval [Manjunath et al., 2002]. The major difference between the scenarios is the location where the meta-data, that describes the content, is generated and searched.

## 2.4 Multimedia search and retrieval

Multimedia data and related technologies are becoming a very important part of our everyday life [Pereira et al., 2008]. This is mainly caused by the rapid development of hardware and software, that increases the ease of consumers to acquire, process, store, transmit and share multimedia data. Due to the increasing volume of this multimedia data content retrieval and delivery become central issues. While retrieval refers to the identification of interesting content, delivery describes the transport and consumption of this content.

While the ultimate goal is the consumption of interesting multimedia data, experience has shown that content description plays a central role for multimedia retrieval [Pereira et al., 2008]. The description (meta-data) typically provides the key information about the content, that makes it searchable as text.

The extracted metadata can be used for different applications, which can be roughly grouped into 3 scenarios (see figure 2.6) [Manjunath et al., 2002]:

**Pull:** The content is stored on a server which provides also search, browse and filter functionality. In this way, only the target content is transferred to the client. Examples are large multimedia repositories, such as Flickr<sup>1</sup> and YouTube<sup>2</sup>.

**Push:** In this scenario the content along with the corresponding meta-data is transferred from the server to client, where the content can be searched, filtered and browsed.

<sup>1</sup><http://www.flickr.com>

<sup>2</sup><http://www.youtube.com>

Typical examples include distributed surveillance systems integrating several smart cameras and the electronic program guide (EPG) for TV broadcast.

**Share:** A relatively new scenario is created by the increase of peer to peer networks and file sharing platforms. Within this scenario each client may provide content and/or meta-data as well as search, filter and browse the available content based on the meta-data.

The goal of the following sections is to provide an overview of current developments in the field of multimedia information retrieval including techniques for content description and content retrieval.

### 2.4.1 Content description

As it has been already stated above, the description of content based on metadata provides the key to search, filter and retrieve interesting items from a large set of multimedia data. Meta-data can describe different aspects of multimedia items including content, rights and context. Depending on the semantic level content meta-data can be either:

**Low level:** This group considers numerical audiovisual low level features, that can be automatically extracted from the multimedia data. Visual features include color, texture, shape and motion. Typical audio features include temporal and spectral characteristics.

**High level:** This group contains textual high level features, that are either manually annotated by humans or semi-automatically extracted by a machine.

**Structural:** Structural meta-data describes the organization or arrangement of the multimedia data in terms of spatial or temporal segmentation, audio and video streams, objects and collections.

A standard that facilitates the description of all these aspects is MPEG-7 [Manjunath et al., 2002] also known as multimedia content description interface. It includes standardized tools (language, descriptors and description schemes) for a detailed description of audiovisual content at different granularities (collection, video, image, region) and in different areas (content description, management, organization, navigation, and user interaction).

Formally, the MPEG-7 standard (referred to as ISO 15938) is organized into several parts [Chang et al., 2001]:

**Systems:** MPEG-7 systems specifies system level functionalities such as the preparation of the descriptions for storage, transportation, synchronization of the content and the descriptions, and development of conformant decoders.

**Description Definition Language (DDL):** The MPEG-7 DDL is a standardized language for defining new description schemes (DSs) and descriptors (Ds) as well as extending or modifying existing DDs and Ds.



**Visual:** MPEG-7 visual specifies a set of standardized visual Ds and DSs. Visual Ds describe typical visual features such as color, texture, shape and motion. Furthermore, several DSs support the description of locations and

**Audio:** MPEG-7 audio specifies a set of standardized audio Ds and DSs. Audio Ds consider different classes of audio signals such as music, speech, sounds. They describe audio features such as silence, timbre and melody.

**Multimedia description schemes (DSs):** The MPEG-7 multimedia description schemes provide a framework that allows a high level description of all kinds of multimedia data. It consists of different levels that describe different aspects.

**Reference software:** MPEG-7 reference software aims at providing a reference implementation of the relevant parts of the MPEG-7 standard. The focus of the software is on creating standard compliant descriptions with the normative syntax, instead of extracting features and creating content descriptions.

**Conformance:** MPEG-7 conformance aims at providing guidelines and procedures for testing the performance of MPEG-7 implementations.

### 2.4.2 Content retrieval

Initially multimedia retrieval systems were solely *text based* [Liu et al., 2007]. In such approaches, the content is manually annotated by textual descriptions (keywords, captions, free text), which are used by text search engines to retrieve the multimedia content. In general there are two disadvantages with this approach. First of all, a considerable level of human labor is required to manually annotate large sets of multimedia content. Secondly, it is usually quite difficult to provide rich and reliable textual descriptions of general multimedia data due to the subjectivity of the annotation process. This problem is commonly referred to as *linguistic gap*.

To overcome the problems inherent to the text based retrieval, *content based retrieval* systems were introduced in the early 1980s. The major idea is to index multimedia content based on an audiovisual description that can be automatically extracted from the content.

Although continued efforts have been put into solving the fundamental issues in robust multimedia understanding, several challenges remain. They can be summarized by three major gaps. The *sensory gap* describes the difference between an object in the real world and its computational description derived during the recording. For visual data this commonly refers to the mapping of the 3D data into the 2D image which causes a loss of structural information. On the other hand, the *numerical gap* describes the retrieval problems caused by incomplete or confusing descriptions of the multimedia content. This gap can be reduced by selecting rich and faithful features. Finally the *semantic gap* describes the difference between the low level feature that can be automatically extracted and the high level semantics that

humana usually consider for searching content. In general there is no direct link between these high level concepts and the low level features.

Current techniques for bridging the semantic gap and deriving high level semantics can be categorized into 5 major categories [Liu et al., 2007]:

**Object ontologies:** These techniques are based on simple vocabularies (object ontologies) that can be derived from our daily language. Widely used vocabularies are color names that quantize a color space into a predefined set of colors. These mid level descriptors (vocabularies) are mapped to high level semantics (keywords) based on a priori knowledge. An example is the term “sky” which can be defined as a region of “light blue” color, “uniform” texture and “upper” location.

**Machine learning:** In order to derive high level semantic features machine learning techniques are widely used to predict categories using supervised learning or organize and cluster data using unsupervised learning. Typically these approaches are trained offline to detect certain objects (e.g. cars, persons, faces) or concepts (e.g. explosion, mountains, beach) based on the extracted low level features.

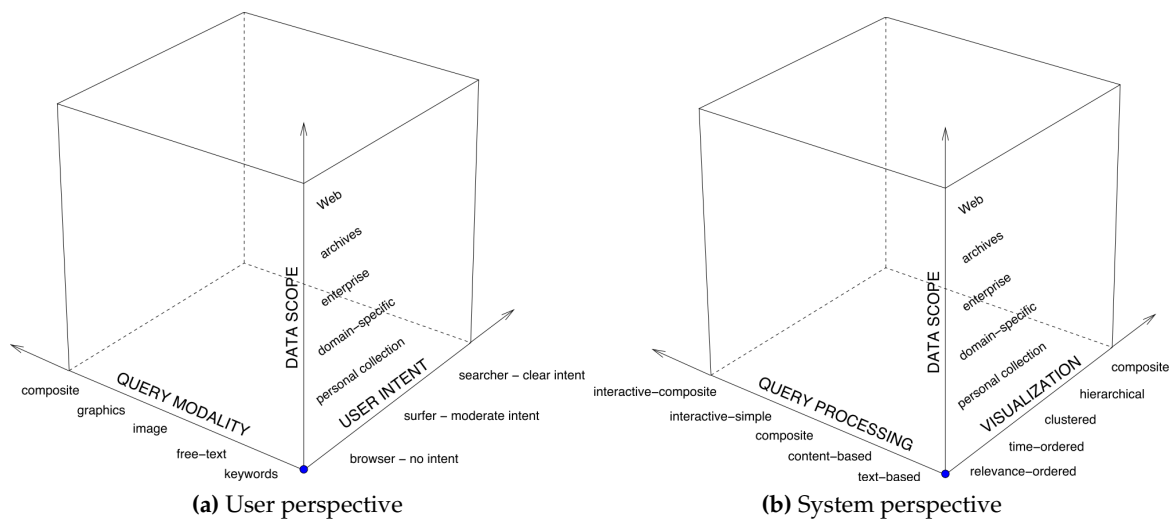
**Relevance feedback:** Approaches utilizing relevance feedback (RF) are based on similar techniques as the category described above. The major difference lies in the online processing which tries to learn the users search intention on the fly. By incrementally learning through the interaction with the user, RF improves the performance of content based retrieval.

**Semantic templates:** Although semantic templates (ST) have not been widely used yet it is a promising approach for retrieval. A semantic template is basically a manually defined mapping between a high level concept and low level features generated from a collection of sample images. To create a ST the user first specifies objects, their spatial and temporal constraints and weights for the different low level features. Since the generation largely depends on the in depth understanding of the high level concept and the low level features these approaches require expert users.

**Complementary sources:** Beside the information that can be extracted from the content, approaches belonging to this category consider also complementary information (text) for the retrieval process. This information can be extracted from different sources including closed captions, meta-data (EXIF, IPTC) and tags.

Many approaches combine several of these techniques to support semantic content retrieval. For example, RF is often combined with object ontologies and machine learning [Mezaris et al., 2003].

According to Datta et al. [2008] the design of a search engine requires understanding and characterizing of user system interaction from a user and a system perspective (see figure 2.7). Combining them leads to several criteria that need to be considered when a developing



**Figure 2.7:** Characterization of the user system interaction for a multimedia search engine from a user and a system perspective [Datta et al., 2008]. Since some of the dimensions are related the 4 most important dimensions are related to the data, the user, the query and the presentation.

multimedia retrieval engine [Datta et al., 2008]:

**User intent:** This describes the clarity of the user about what he wants. This ranges from a user with no clear goal (browser) over a user with a moderate clarity (surfer) to a user who is very clear about what he is looking for (searcher).

**Data scope:** Understanding the scope of the multimedia data plays a key role in the complexity of the search system. Ordered by complexity the following categories can be distinguished: personal collection, domain specific collection, enterprise collection, archives, and the world wide web (WWW).

**Query modality:** Since any retrieval engine relies in some form on the interaction with the user, the complexity of queries supported by the system is an important aspect. Query modalities reach from text based (keywords, free text) over content based (images, graphics) to interactive (relevance feedback) and combinations of them.

**Presentation type:** The presentation of search results is perhaps one of the most important factors in the acceptance and popularity of a retrieval system. Depending on the modality of the data (audio, image, video, text) different presentations are required. Nevertheless, the criteria used for the representation of the documents are similar. They include relevance ordered, time ordered, clustered, hierarchical and combinations of them.

### 2.4.3 Discussion

Multimedia content description and retrieval is a very generic application scenario that can be also part of other scenarios (e.g. forensic search for surveillance). The basic idea is to

analyze the multimedia content and extract a comprehensive description in form of meta-data that can be used to search and retrieve interesting items. Since humans are one of the most interesting objects, the audiovisual analysis of humans is an important step to extract suitable meta-data.

## **2.5 Conclusion**

The looking at people research domain touches several research fields (e.g. surveillance, biometrics, search and retrieval) with different goals and interests. Nevertheless, each of them is at least partly based on the visual analysis of humans. Thereby, they focus on different aspects of humans such as body motion for interpreting activity and behavior (surveillance), face appearance and body motion for recognizing the identity (biometrics), or a combination of different interests (search and retrieval). Furthermore, traditionally independent fields move closer together such as surveillance and biometrics to achieve situation awareness that requires both the location and the identity of a person [Hampapur et al., 2005]. This shows the need for a generic visual analysis and description framework which is related to the recent standardization activities within MPEG-7 [Manjunath et al., 2002].

## Chapter 3

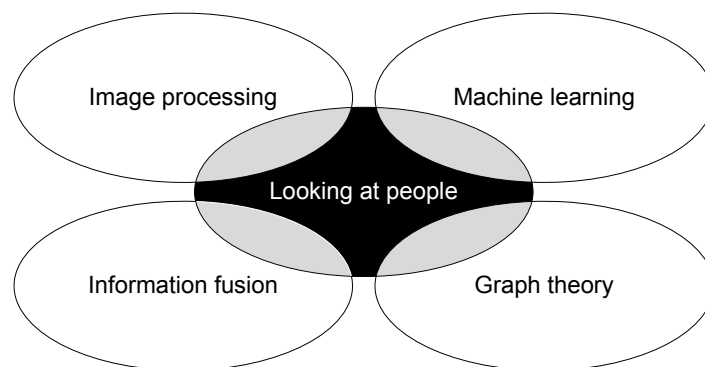
# Fundamental techniques

### 3.1 Introduction

The looking at people domain lies at the crossroads of several other research areas as it has been shown in chapter 2. Furthermore, it is based on techniques from several fields which are shown in figure 3.1. The goal of this chapter is to provide a comprehensive overview of the relevant research fields and describe the techniques considered within this work in more detail. Besides that, the most important references are provided and the techniques are linked to the parts of this thesis where they are used.

### 3.2 Image processing

Image processing deals with low level algorithms applied to images or videos. This section reviews common image processing and analysis technologies which are used throughout this work, including point operations, geometric transformations and image analysis methods.



**Figure 3.1:** Technologies used within the looking at people domain.

### 3.2.1 Point operations

Point operations are a simple method for image enhancement [Fisher, 2004]. A pixel value  $p'$  of the output image depends only on the corresponding pixel of the input image  $p$  and a mapping function  $f$ . The overall process can then be described as

$$p' = f(p) \quad (3.1)$$

Below several point operations used within the thesis will be shortly described.

#### Contrast stretching

Contrast stretching (normalization) attempts to improve the contrast of the pixels  $p$  within an image by stretching the range of its intensity values  $[c, d]$  to the maximum range  $[a, b]$ . This is achieved by applying the following linear mapping function

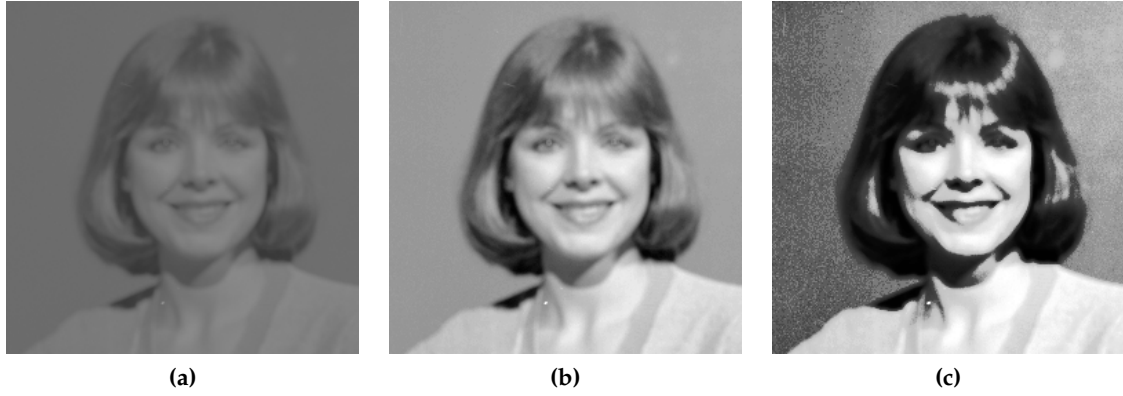
$$p' = (p - c) \frac{b - a}{d - c} + a \quad (3.2)$$

Since outliers may lead to an unrepresentative scaling the intensity histogram is used to determine the input range  $[c, d]$ . One way is to consider a set of percentiles (e.g. 5% and 95%) as the boundaries. Another way is to determine the mode of the histogram and define a cutoff fraction which is the minimum fraction of this peak magnitude below which the data is ignored. Figure 3.2 provides an example of contrast stretching which shows a significant contrast improvement of the output image (b) over the input image (a).

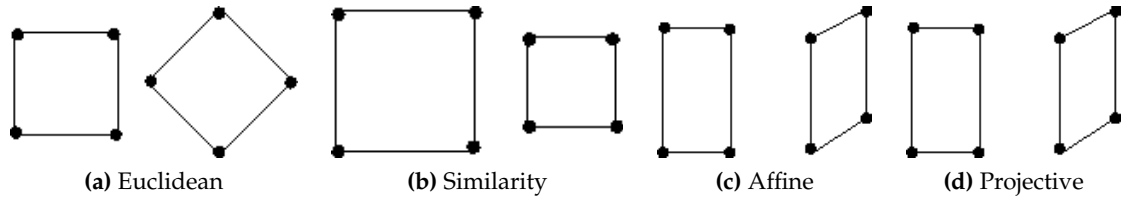
Within this work contrast stretching is used for illumination compensation in the face detection (see chapter 6) and face recognition (see chapter 7) modules.

#### Histogram equalization

Histogram equalization employs a monotonic, non-linear mapping function to the pixels  $p$  of the input image to obtain an output image  $p'$  with a uniform intensity distribution. In the digital domain, the output image may not be fully equalized and may contain unused intensity levels. These effects decrease when the number of pixels and intensity levels in the input image increase. Figure 3.2 provides an example of histogram equalization that shows a dramatic but artificial contrast improvement of the output image (c) in comparison to the input image (a).



**Figure 3.2:** Image enhancement with different point operations: (a) input image, (b) output image after contrast stretching, (c) output image after histogram equalization.



**Figure 3.3:** Hierarchy of geometric 2D transformations from the Euclidean transformation with 3 degrees of freedom (DOF) to the projective transformation with 8 degrees of freedom.

### 3.2.2 Geometric transformations

A geometric 2D transformation maps a pixel from the location  $(x_1, y_1)$  in the input image to another location  $(x_2, y_2)$  in the output image [Fisher, 2004] according to

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \mathbf{A} \times \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \vec{b} \quad (3.3)$$

The whole transformation may be composed of different geometric operations that can be defined through the matrix  $\mathbf{A}$  and the vector  $\vec{b}$ :

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad (\text{translation}) \quad (3.4)$$

$$\mathbf{A} = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (\text{rotation}) \quad (3.5)$$

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (\text{scaling}) \quad (3.6)$$

These operations can be combined which leads to a hierarchy of 2D transformations as shown in figure 3.3. The *euclidean transformation* supports only translation and rotation

which leads to 3 degrees of freedom (DOF) and requires at least 2 point correspondences to be estimated. The *similarity transformation* adds isotropic scaling which leads to 4 DOF (2 points). The *affine transformation* adds shear which leads to 6 DOF (3 points). The projective transformation is the most complex transformation with 8 DOF (4 points).

Within this work geometric transformations are used to normalize faces regarding scale, translation and in plane rotation for both face detection (chapter 6) and face recognition (chapter 7).

### 3.2.3 Image analysis

The goal of image analysis is to extract mid level information from the low level image data. The extracted data may again be represented in the form of images or maps. Again the goal of this section is not review the whole area of image analysis but provide a short overview of the techniques used within this thesis.

#### Connected component labeling

Connected component labeling scans an image and groups similar pixels based on their connectivity. The components are usually described by assigning unique labels to all the pixels within a group.

The connected pixel regions are identified by scanning the image from the top left to the bottom right and labeling connected pixels that share the same set of intensity values. It works both on binary and gray level images and different measures of connectivity can be used. Within this work only binary images and 8 connectivity are considered. An operator is moved across the image in a row major scan until it comes to a pixel  $p$  where  $p = 1$ . It examines the four neighbors of  $p$  that have been previously scanned and labels  $p$  as follows:

- If all neighbors are 0 assign a new label to  $p$
- If only one neighbors is 1 assign its label to  $p$
- If more than one neighbor is 1 assign one of the labels to  $p$  and store equivalences

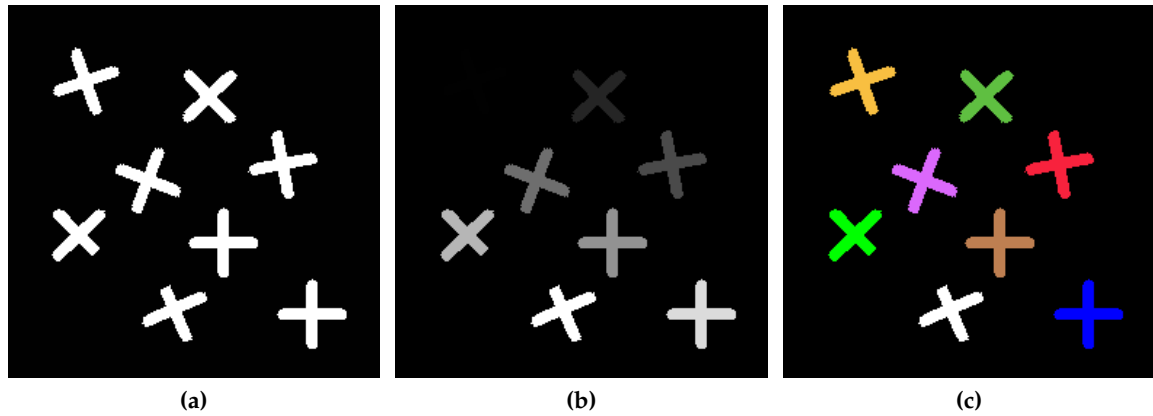
After completing the scan, equivalences are grouped and a unique label is assigned to them. Based on that, each pixel is replaced with its group label. For better visualization colors may be used to represent the groups. The whole process is shown for an example in figure 3.4.

Within this work connected component labeling is used to group pixels into blobs for the body recognition (chapter 5).

## 3.3 Machine learning

Machine learning deals with the assignment of a physical object or event to one of several categories [Duda et al., 2002]. The basic concepts of machine learning are features and patterns [Gutierrez, 2002]. A *feature* can be any distinctive aspect, quality or characteristic and





**Figure 3.4:** Connected component labeling applied to a binary example: (a) binary input image, (b) labeled output image, (c) colored output image.

may be either symbolic (e.g. color) or numeric (e.g. height). The combination of  $D$  individual features  $x_i$  into a vector  $\vec{x}$  is usually called a *feature vector*. The  $D$ -dimensional space defined by the feature vector is called the *feature space*. A *pattern* is a composite of traits or features of a concept or individual. It can be seen as a pair of variable  $(\vec{x}, c)$  with  $\vec{x}$  a feature vector and  $c$  the concept (class) behind an observation.

The large amount of developed machine learning approaches can be grouped into three major categories

**Statistical:** Patterns are classified based on the underlying statistical model of the features.

The model is defined as a family of class conditional probabilities.

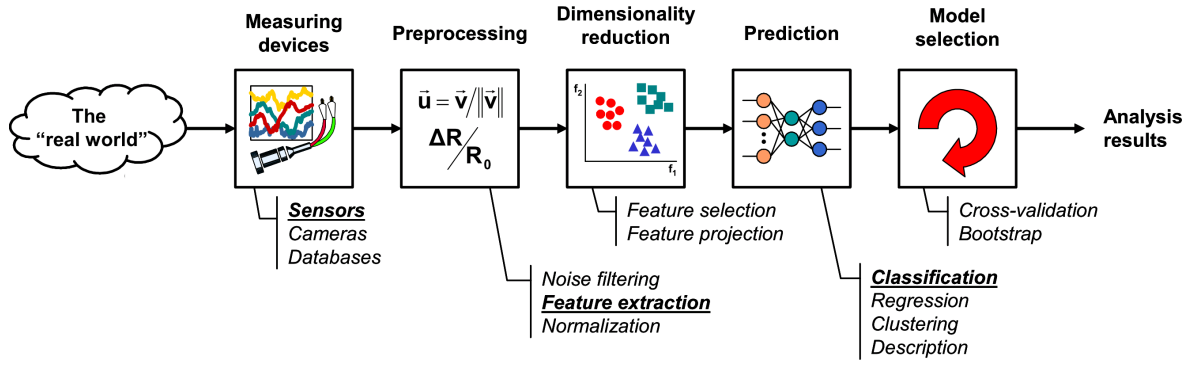
**Structural:** Patterns are classified based on measures of structural similarity. The structure is usually represented as formal grammars or relational descriptions (graphs).

**Neural:** The classification is based on the response of a network of processing units (neurons) to an input stimuli (pattern). The response of the network is determined by the connectivity and strength of synaptic weights between the neurons.

A typical pattern recognition system consists of several modules which are shown in figure 3.5. This work considers techniques for matching, feature reduction, density estimation, clustering and classification, which will be reviewed in the following sections.

### 3.3.1 Matching

Matching refers to the comparison of two feature vectors  $x$  and  $y$  which can be expressed either by a measure of similarity or dissimilarity. Within this work matching is an integral part to compare different feature vectors to each other and is used in each of the chapters.



**Figure 3.5:** Overview of a typical pattern recognition system [Gutierrez, 2002]. Within this work techniques for preprocessing, dimensionality reduction, and prediction are used.

## Dissimilarities

The dissimilarity  $d(\vec{x}, \vec{y})$  measures the discrepancy between two features  $\vec{x}, \vec{y}$ . It might also be viewed as a measure of disorder and usually has a range between  $[0, \infty[$  or  $[0, 1]$ . There are many types of dissimilarity functions which may satisfy some or all of the following conditions [von Luxburg, 2004]:

$$d(\vec{x}, \vec{x}) = 0 \quad (3.7)$$

$$d(\vec{x}, \vec{y}) \geq 0 \quad (\text{non negativity}) \quad (3.8)$$

$$d(\vec{x}, \vec{y}) = d(\vec{y}, \vec{x}) \quad (\text{symmetry}) \quad (3.9)$$

$$d(\vec{x}, \vec{y}) = 0 \implies \vec{x} = \vec{y} \quad (\text{definiteness}) \quad (3.10)$$

$$d(\vec{x}, \vec{y}) + d(\vec{y}, \vec{z}) \leq d(\vec{x}, \vec{z}) \quad (\text{triangle inequality}) \quad (3.11)$$

If a function satisfies the first two conditions it is called a *distance*. On the other hand, if a function satisfies all the conditions it is called a *metric*.

## Similarities

The similarity  $s(\vec{x}, \vec{y})$  between features  $\vec{x}, \vec{y}$  is usually quite difficult to measure. It reflects the strength or relationship between two objects and usually has a range of  $[-1, 1]$  or  $[0, 1]$ . In a similar way to dissimilarities, several properties of similarities can be defined [von Luxburg, 2004]:

$$s(\vec{x}, \vec{x}) > 0 \quad (3.12)$$

$$s(\vec{x}, \vec{y}) = s(\vec{y}, \vec{x}) \quad (\text{symmetry}) \quad (3.13)$$

$$s(\vec{x}, \vec{y}) \geq 0 \quad (\text{non negativity}) \quad (3.14)$$

$$s(\vec{x}, \vec{y}) \leq s(\vec{x}, \vec{x}) \quad (3.15)$$

### Conversion

Usually it is recommended to use an algorithm that can deal directly with the present representation (similarity or dissimilarity). Nevertheless, sometimes it is necessary to convert between different representations. This can be achieved by several heuristics. The general idea is to convert a similarity into a dissimilarity or vice versa by applying a monotonically decreasing function. This is according to the intuition that a distance is small if a similarity is large. The following equations show several ways to convert distances to similarities :

$$s(\vec{x}, \vec{y}) = 1 - d(\vec{x}, \vec{y}) \quad (\text{linear}) \quad (3.16)$$

$$s(\vec{x}, \vec{y}) = \exp(-d(\vec{x}, \vec{y})) \quad (\text{exponential}) \quad (3.17)$$

$$s(\vec{x}, \vec{y}) = 1 / (1 + d(\vec{x}, \vec{y})) \quad (\text{quotient}) \quad (3.18)$$

### Distances

Between two  $D$ -dimensional vectors or points  $\vec{x} = (x_1, \dots, x_D)$  and  $\vec{y} = (y_1, \dots, y_D)$  different distances can be defined. The most common distance measure is the Minkowski metric which leads to several well known distance measures ( $p$ -norm distances) depending on the order  $p$ . In its general form it is given as:

$$d_p(\vec{x}, \vec{y}) = \left( \sum_i |x_i - y_i|^p \right)^{1/p} \quad (3.19)$$

The 1-norm distance, also called the *Manhattan distance* can be illustrated by the distance a car needs to drive within a city laid out in square blocks. It is given as:

$$d_1(\vec{x}, \vec{y}) = \sum_i |x_i - y_i| \quad (3.20)$$

The 2-norm distance, also called *Euclidean distance* is the most intuitive distance, that can be illustrated as the distance between two points measured with a ruler. It is given as:

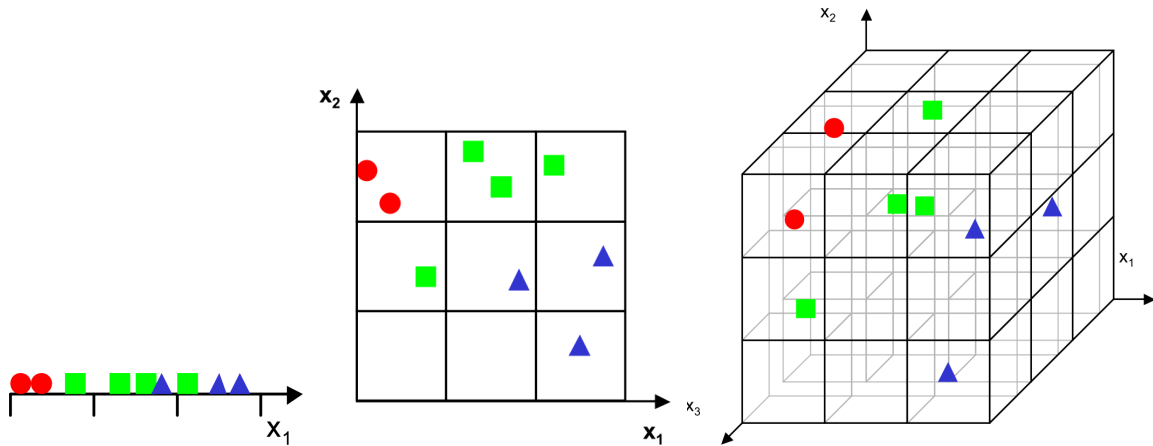
$$d_2(\vec{x}, \vec{y}) = \sqrt{\sum_i (x_i - y_i)^2} \quad (3.21)$$

The  $\infty$ -norm distance, also called *Chebyshev distance* can be illustrated as the largest distance into one direction. It is given as:

$$d_\infty(\vec{x}, \vec{y}) = \max_i |x_i - y_i| \quad (3.22)$$

### 3.3.2 Feature reduction

Feature reduction is the process of reducing the number of dimensions of random variables under consideration. Beside the obvious reasons such as reduced storage space and



**Figure 3.6:** Illustration of the “curse of dimensionality” for a fixed number of samples and different number of dimensions [Gutierrez, 2002]: (a) 1 dimension, (b) 2 dimensions, (c) 3 dimensions. As it can be seen the volume increases exponentially with the addition of extra dimensions.

computational complexity, another reason is to avoid the “*curse of dimensionality*” which is illustrated in figure 3.6. This effect initially mentioned by Bellman [1957] describes the problem caused by the exponential increase in volume associated with adding extra dimensions to a mathematical space. More precisely, traditional machine learning methods do not work very reliably if the number samples  $N$  is small and the number of dimensions  $D$  is large.

The general assumption behind feature reduction techniques is that the required information to achieve a certain task is not distributed along all dimensions. Thus only some features are important and extracting them is the general goal of any feature reduction technique. The difference of the approaches lies mainly in how the reduced feature vector is obtained. Existing approaches can be grouped into two major categories [Jain et al., 2000]

**Feature selection:** Approaches belonging to this category try to find a subset of the original features by filtering or wrapping. Proposed methods include exhaustive search, branch and bound search, sequential forward selection (SFS) and sequential backward selection (SBS).

**Feature projection:** Approaches belonging to this category apply a mapping of the original feature space into another reduced feature space. Proposed methods include principal component analysis (PCA), linear discriminant analysis (LDA), independent component analysis (ICA), kernel PCA, and self organizing maps (SOM).

### Feature selection

Given a set of features  $X = \{x_i | i = 1 \dots D\}$  the goal of feature selection is to find a subset  $Y = \{x_j | j = 1 \dots M\}$  with  $M < D$  that optimizes an objective function.

In order to find the best feature subset a *search strategy* is needed to reduce the number of evaluated combinations and direct the search process [Gutierrez, 2002]. The large number

of search approaches can be grouped into three major categories:

**Sequential:** These algorithms add or remove features sequentially, but have the tendency to become trapped in local minima. Representative methods include sequential forward selection, sequential backward selection and bidirectional search.

**Exponential:** These algorithms evaluate a number of subsets that grows exponentially with the dimensionality of the search space. Representative methods are exhaustive search, and branch and bound.

**Randomized:** These algorithms incorporate randomness into their search procedure to escape local minima. Representative methods include simulated annealing and genetic algorithms.

The *objective function* evaluates candidate subsets and returns a quality measure that is used by the search strategy to select candidates [Gutierrez, 2002]. Objective functions can be divided into two groups:

**Filters:** The objective function evaluates features subsets by their information content such as interclass distance, statistical dependence and information theoretic measures. The advantages of filters are a lower complexity and the better generality while the major disadvantages are the difficulty to select a suitable criteria and the tendency to select large feature subsets.

**Wrappers:** The objective function is based on a classifier and compares features subsets based on an evaluation measure such as accuracy or recognition rate. The advantages of wrappers are better performance and less over-fitting since they consider not only the data but also the used classifier. On the other hand, this leads to a much larger complexity and a lack of generality.

Within this work two feature selection methods are considered for body recognition (chapters 5), namely best individual feature (BIF) and sequential forward selection (SFS). Wrappers are used as objective function.

**Best individual feature (BIF)** The best individual feature selection is by far the least complex method [Gutierrez, 2002], since it evaluates each feature  $\{x_i | i = 1 \dots D\}$  individually and then selects those  $M$  features with the highest performance. If  $M$  is not predefined it is found by sorting the features based on their performance and sequentially adding them to the subset until the overall performance measure does not increase anymore.

Since this strategy ignores the interaction between the different features, it may not find an optimal subset of features.

**Sequential forward selection (SFS)** Sequential forward selection is the simplest greedy search algorithm [Gutierrez, 2002]. It starts from an empty set and sequentially adds the feature  $x_i$  that results in the best objective measure when combined with the features  $Y_k$  that have been already selected until iteration  $k$ . This process continues until the objective performance measure does not increase anymore.

SFS performs best when the optimal subset has a small number of features. The major disadvantage of SFS is that features are not removed after becoming obsolete by the addition of other features.

### Feature projection

Given a set of feature vectors  $\vec{x}_k \in \mathbb{R}^D, k = 1, \dots, N$  within the original  $D$ -dimensional feature space, both approaches project them into a reduced  $M$ -dimensional feature space. The new feature vectors  $\vec{y}_k \in \mathbb{R}^M, k = 1, \dots, N$  are then defined by the following linear transformation

$$\vec{y}_k = \mathbf{W}^T \vec{x}_k, k = 1, 2, \dots, N \quad (3.23)$$

where  $\mathbf{W} \in \mathbb{R}^{D \times M}$  is the transformation matrix with the orthonormal base functions  $\vec{w}_j$  along the columns. Finding the transformation matrix is guided by an objective function that is optimized. Depending on the criteria measured by the objective function, existing approaches can be grouped into two categories which are illustrated for a 2D example in figure 3.7:

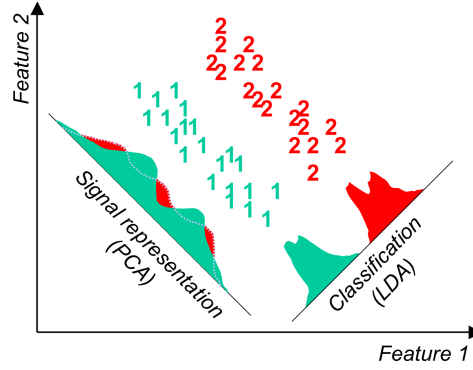
**Representation:** The goal of this approach is to represent the data samples as accurately as possible in the lower-dimensional space. Therefore no class information is considered. The most prominent method is the principal component analysis (PCA).

**Classification:** The goal of this approach is to enhance the class-discriminatory information within the lower-dimensional space. Therefore the corresponding class information is considered beside the data samples. The most important method is the linear discriminant analysis (LDA).

Since the class information required for the LDA may not be available within certain application scenarios such as search and retrieval, only the PCA is considered within this work for the face recognition (chapter 7).

**Principal component analysis (PCA)** The principal component analysis (PCA) finds a linear projection that maximizes the overall scatter of the data. Since it does not use any additional information beside the raw data, it can be used in conjunction with supervised and unsupervised learning approaches.

Given a set of  $D$ -dimensional feature vectors the *total scatter matrix*  $\mathbf{S}_T$  which is equiva-



**Figure 3.7:** Illustration of the different feature projection approaches for a two dimensional sample [Gutierrez, 2002]. For supervised learning with a representative data set LDA usually achieves a better class separation than PCA. On the other hand, LDA is not applicable for unsupervised problems or if the data set is not representative enough.

lent to the covariance matrix is computed as

$$\mathbf{S}_T = \sum_{i=1}^N (\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu})^T \quad (3.24)$$

where  $N$  is the number of sample images and  $\vec{\mu} \in \mathbb{R}^D$  is the sample mean of the feature vectors. Since the goal is to maximize the total scatter the projection matrix  $\mathbf{W}_{PCA}$  is chosen to maximize the determinant of the total scatter matrix

$$\mathbf{W}_{PCA} = \underset{\mathbf{W}}{\operatorname{argmax}} |\mathbf{W}^T \mathbf{S}_T \mathbf{W}| \quad (3.25)$$

It can be shown, that the optimal solution is found by computing the eigenvectors of the total scatter matrix  $\mathbf{S}_T$ . The corresponding eigenvalues provide a measure on how much variance each dimension contains. Thus the eigenvalues and their corresponding eigenvectors are sorted in a descending way and a subset of  $M$  eigenvectors is chosen as the reduced basis. The dimensionality of the projected feature space  $M$  can be predefined or determined automatically by computing the normalized cumulative sum of the sorted eigenvalues and thresholding it.

**Linear discriminant analysis (LDA)** The linear discriminant analysis (LDA) finds a linear projection that is optimal for discriminating between the classes rather than best describing the data. More formally it finds a linear combination of the independent features which yields the largest mean differences between the given classes.

For a given set of  $D$ -dimensional feature vectors two measures are defined: The *within class scatter matrix* given as

$$\mathbf{S}_W = \sum_{j=1}^C \sum_{i=1}^{N_j} (\vec{x}_i^j - \vec{\mu}_j)(\vec{x}_i^j - \vec{\mu}_j)^T \quad (3.26)$$

where  $\vec{x}_i^j$  is a sample belonging to class  $j$ ,  $\vec{\mu}_j$  is the mean and  $N_j$  the number of samples belonging to class  $j$ . The *between class scatter matrix* is defined as

$$\mathbf{S}_B = \sum_{j=1}^C N_j (\vec{\mu}_j - \vec{\mu})(\vec{\mu}_j - \vec{\mu})^T \quad (3.27)$$

where  $\vec{\mu}$  represents the overall mean. The goal is to maximize the between class scatter, while minimizing the within class scatter which is known as fisher linear discriminants (FLD). Thus the optimal projection matrix  $\mathbf{W}_{FLD}$  is chosen to maximize the ratio between the determinant of  $\mathbf{S}_B$  and the determinant of  $\mathbf{S}_W$  as

$$\mathbf{W}_{FLD} = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|} \quad (3.28)$$

It can be shown that if  $\mathbf{S}_W$  is a non singular matrix the optimal solution is found by computing the eigenvectors of  $\mathbf{S}_W^{-1} \mathbf{S}_B$ . The corresponding eigenvalues provide a measure on how much discriminative information each dimension provides. Furthermore it should be noted, that there are at most  $C - 1$  nonzero generalized eigenvectors and that at least  $N + C$  samples are required to guarantee that  $\mathbf{S}_w$  does not become singular. To solve this problem PCA is used to create an intermediate space on which the FLD is applied. That way, the original  $D$ -dimensional space is projected into an  $K$ -dimensional space using PCA and finally into a  $M$ -dimensional space using FLD

$$\mathbf{W}_{LDA} = \mathbf{W}_{PCA} \mathbf{W}_{FLD} \quad (3.29)$$

with  $\mathbf{W}_{LDA} \in \mathbb{R}^{D \times M}$ ,  $\mathbf{W}_{PCA} \in \mathbb{R}^{D \times K}$  and  $\mathbf{W}_{FLD} \in \mathbb{R}^{K \times M}$  which is commonly referred to as linear discriminant analysis (LDA).

### 3.3.3 Density estimation

The goal of density estimation is derive a density model  $P(X)$  from a finite number of data points  $X$ . These models form the basis for the different classifiers described in section 3.3.5.

In general two groups of approaches for density estimation can be distinguished

**Parametric:** A given form of the density function is assumed (e.g. Gaussian) and the parameters of the function (e.g. mean and variance) are then estimated by fitting the model to the given data set. Well known examples include the mean model (MM), the single Gaussian model (SGM), and the Gaussian mixture model (GMM).

**Nonparametric:** No form of the density function is assumed and the density estimation is entirely driven by the data. Well known examples include the instance model used within the  $k$  nearest neighbor (kNN) classifier, and the naive and joint Bayes models (BM).



### Instance model

The instance model is not really a density model and usually only used in combination with the  $k$  nearest neighbor decision rule (see section 3.3.5). The model simply consists of all the stored samples within a dataset.

The distance  $d(\vec{x})$  of an observation  $\vec{x}$  from the model is typically computed as the minimum distance of the observation to all the instances  $\vec{x}'_i$  with  $i = 1, \dots, N$  stored in the model

$$d(x) = \min_i d(\vec{x}, \vec{x}'_i) \quad (3.30)$$

with  $d(\vec{x}, \vec{x}'_i)$  being either a generic distance defined in section 3.3.1 or a data specific distance.

Using the  $k$  nearest neighbor rule the density estimate becomes [Gutierrez, 2002]

$$p(\vec{x}) = \frac{k}{NV} \quad (3.31)$$

with the number of instances  $k$  within a growing hypersphere of volume  $V$  and the overall number of instances  $N$ .

### Bayes model

The Bayes model (usually in the form of a histogram) is the simplest form of non parametric density estimation. The data space is quantized into a predefined number of bins and the density at the center of each bin is approximated by the fraction of samples that fall into the corresponding bin  $k$  and the number of overall samples  $N$  which can be written as

$$p(\vec{x}) = \frac{k}{N} \quad (3.32)$$

Regarding the different dimensions, two different cases can be distinguished. The joint Bayes model (usually in the form of one multidimensional histogram) considers the dependence between the individual dimensions and the corresponding probability is defined as

$$p(x_1, \dots, x_N) = p(x_1)p(x_2, \dots, x_N|x_1) = p(x_1)p(x_2|x_1)p(x_3, \dots, x_N|x_1, x_2) \quad (3.33)$$

On the other hand, the naive Bayes model (usually in the form of several one dimensional histograms) assumes independence between the different dimensions. Based on the Bayes rule and the independence assumption  $p(x_i|x_j) = p(x_i) \quad \forall (i, j) : i \neq j$  the probability is simplified to

$$p(x_1, \dots, x_N) = \prod_{i=1}^N p(x_i) \quad (3.34)$$

The advantages of a Bayes model are that it is easy to implement, has a low complex-

ity and provides a good density estimate if the number of samples is large enough. The disadvantages are the discontinuities between the bins and the exponential growth of the required number of bins which leads to the curse of dimensionality.

Within this work Bayes models have been used for modeling the color of individual body parts in the body recognition module (chapter 5).

### Single Gaussian

The assumption behind the single Gaussian model is that the data distribution can be approximated by a Gaussian distribution. It can be shown that the maximum likelihood (ML) parameter estimates for a univariate Gaussian distribution defined as

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x-\mu}{2\sigma^2}\right) \quad (3.35)$$

are given by the sample mean  $\mu$  and the sample variance  $\sigma$ .

Similarly, the parameters for a multivariate Gaussian distribution, defined as

$$p(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right) \quad (3.36)$$

are provided by the sample mean vector  $\vec{\mu}$  and the sample covariance matrix  $\Sigma$ . Both can be estimated directly from the data based on the following equations

$$\vec{\mu} = \frac{1}{N} \sum_{i=1}^N \vec{x}_i \quad (3.37)$$

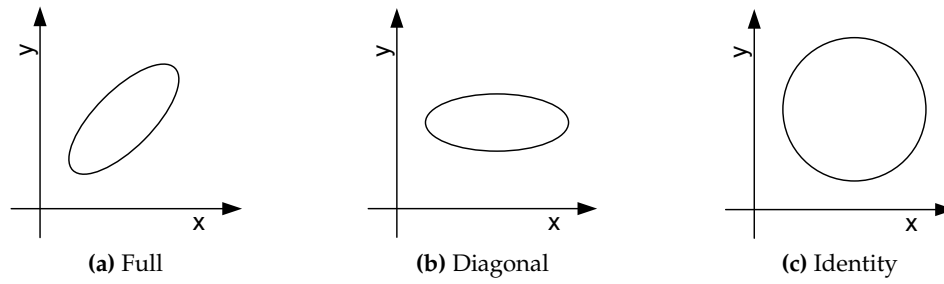
$$\Sigma = \frac{1}{N} \sum_{i=1}^N (\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu})^T \quad (3.38)$$

Different types of the covariance matrix  $\Sigma$  can be distinguished. While a full covariance matrix describes the dependence of the different features, a diagonal matrix assumes independence of them. If the elements of the diagonal are restricted to have the same values, equal variance of the individual features is further assumed. This is illustrated in figure 3.8 for 2 dimensional data.

Within this work single Gaussian models are used for relevance feedback in the multi-modal person search system (chapter 8).

### Mean model

The mean model is the simplest parametric model and is mostly used in conjunction with a minimum distance classifier. It describes only the location of the data distribution based on its mean vector  $\vec{\mu}$  (as defined in equation 3.37) and discards any variation information. The



**Figure 3.8:** Illustration of a 2D multivariate Gaussian distribution with different covariance matrix types. While the full covariance matrix usually offers the best description of the data, it requires a large number of samples to obtain a robust estimate.

distance of a sample to the model is simply defined as

$$d(\vec{x}) = d(\vec{x}, \vec{\mu}) \quad (3.39)$$

with  $d(\vec{x}, \vec{\mu})$  being either a generic distance (defined in section 3.3.1) or a data specific distance.

Within this work mean models are used in conjunction with a minimum distance classifier for the classification in the body recognition module (chapter 5).

### 3.3.4 Clustering

Clustering can be considered as an unsupervised learning problem, where the goal is to find a structure within unlabeled data. Objects are grouped into clusters according to some similarity criterion, which can be based on distances or concepts. The criterion depends largely on the application and on the goal of the clustering including:

- Finding representatives for homogeneous groups for data reduction
- Finding natural clusters and describe their unknown properties
- Finding useful and suitable groupings
- Finding unusual objects for outlier detection

Existing clustering approaches can be categorized based on several criteria [Jain et al., 1999]:

**Exclusive vs. fuzzy:** Exclusive (hard) clustering approaches (e.g. k-means, agglomerative clustering) group data in an exclusive way, which means that an object belongs to a single cluster and can not be in another cluster. On the other hand, fuzzy (overlapping) clustering approaches (e.g. fuzzy c-means) are based on fuzzy sets, where each object belongs to all clusters with a certain degree of membership. A fuzzy clustering can be converted into an exclusive clustering by assigning each object to the cluster with the largest membership value.

**Partitional vs. hierarchical:** Partitional clustering (e.g. k-means, fuzzy c-means) leads to single level clustering with a fixed number of clusters. In contrast, hierarchical clustering (e.g. agglomerative clustering) generates a multi level grouping that combines clusters iteratively. A hierarchical clustering can be converted into a partitional clustering by cutting or pruning the resulting dendrogram.

**Deterministic vs. probabilistic:** Deterministic clustering approaches (e.g. agglomerative clustering) generate always the same clusterings for a given set of objects, while probabilistic clustering approaches (e.g. expectation maximization) are usually based on probability distributions to describe clusters and random initialization.

The different clustering methods described below have been considered for creating the visual thesaurus in the visual person search system (section 9).

### k-means

The k-means approach [MacQueen, 1967] is one of the simplest clustering approaches, that follows an easy way to assign the objects within a given dataset to a certain number of clusters  $k$ . For each of the clusters, an initial cluster centroid is chosen randomly. The clustering itself is an iterative process, that consists of two steps:

1. Assign data points to closest cluster based on the distance to the centroid
2. Update cluster centroids by computing the centroid of the assigned points

These steps are repeated until convergence of the centroid positions or until a predefined number of iterations is reached.

Although the iterative procedure will always terminate, the k-means approach does not necessarily find a global optimum. It is very sensitive to the initially selected cluster centers. One way to deal with this, is to run the clustering multiple times. Furthermore, it is difficult to know the number of clusters a priori. This can be handled by repeating the clustering with different number of clusters and choose the optimal number of clusters based on an internal evaluation criteria (as described in section B.6).

### fuzzy c-means

Fuzzy c-means clustering [Bezdek, 1981] can be seen as an extension of the k-means clustering towards fuzzy sets. The main idea is to allow each feature vector  $\vec{x}_i$  to belong to different clusters centers  $\vec{c}_j$  with a certain degree of membership  $u_{ij}$ . The approach is based on minimizing the following objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|\vec{x}_i - \vec{c}_j\|^2, \quad 1 < m \leq \infty \quad (3.40)$$

The clustering is carried out by iteratively optimizing the objective function and updating the memberships  $u_{ij}$  and cluster centroids  $c_j$  in the following way:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left( \frac{\|\vec{x}_i - \vec{c}_j\|}{\|\vec{x}_i - \vec{c}_k\|} \right)^{\frac{2}{m-1}}} \quad (3.41)$$

$$\vec{c}_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot \vec{x}_i}{\sum_{i=1}^N u_{ij}^m} \quad (3.42)$$

The approach iterates until the maximum difference of the membership values falls below a certain threshold:  $\max_{ij} |u_{ij}^{k+1} - u_{ij}^k| < \epsilon$ , where  $k$  is the iteration step. Like the k-means approach, the fuzzy c-means approach may converge to a local optimum.

### Agglomerative clustering

The basic process of any agglomerative clustering approach [Johnson, 1967] consists of the following steps:

1. Assign each item to a single cluster and compute the pairwise distance between them to build a distance matrix.
2. Find closest pair of clusters and merge them into a single cluster.
3. Compute distances between the new cluster and all other clusters and update the distance matrix.
4. Repeat steps 2 and 3 until all clusters have been merged into a single cluster.

When clusters contain only single items, the distance between the clusters is simply the distance between the corresponding items. However, once a cluster contains multiple items, a *linkage or amalgamation rule* is required to measure the distance between two clusters [Sturn, 2000]. Some of the most commonly used linkage rules are described below:

**Single linkage:** The distance between two clusters is defined as the smallest distance between objects of the different clusters. Its greatest drawback is the tendency to produce long chain like clusters.

**Complete linkage:** The distance between two clusters is defined as the largest distance between objects of the different clusters. This rule usually performs quite well, if objects actually form naturally distinct clouds. The method is inappropriate for chain like data.

**Average linkage:** The distance between two clusters is defined as the average distance between the objects of the different clusters.

Any agglomerative clustering approach leads to a hierarchical representation of the cluster mergings, which can be represented as a dendrogram. A *dendrogram* is a tree, where each node corresponds to a certain cluster and leafs correspond to individual objects. Several methods have been proposed to create a partition from a dendrogram [Solomonoff et al., 1998]. These techniques are either based on *cutting* the dendrogram at a given height or *pruning* the dendrogram by selecting clusters at different heights.

### 3.3.5 Classification

Classification can be seen as a supervised learning problem with the goal to learn a model from labeled objects to predict the label of previously unseen objects as accurately as possible [Tan et al., 2005]. It usually consists of two individual steps. During the *training* stage a model is learned from features  $\vec{x}$  with known class labels  $c$ . In the testing stage this model is used to predict the class labels of unknown objects.

According to Jain et al. [Jain et al., 2000] classification approaches can be grouped into 3 major categories:

**Matching approaches:** These intuitive approaches are based on the concept of similarity which means that similar patterns should be assigned to the same class. Usually a few prototypes per class in combination with a suitable dissimilarity metric are used for the classification. The appropriate choice of the prototype and the metric is crucial for the classification performance. Well known approaches include minimum distance classifier (MDC) and k nearest neighbor (kNN) classifier.

**Probabilistic approaches:** Approaches belonging to this category use the Bayes decision rule to assign a pattern to the class with the highest a posteriori probability. Besides the a priori class probabilities, costs for different types of misclassifications based on a loss function can also be taken into account. If a 0/1 loss function and equal error probabilities are considered the maximum a posteriori (MAP) classifier is equivalent to the maximum likelihood (ML) classifier.

**Discriminant approaches:** Instead of modeling the different classes, these algorithms construct a decision boundary explicitly by optimizing a certain error criteria. Typically these approaches are applied for two class classification problems but can be adapted to single or multi class problems. The most prominent approach are support vector machines (SVM).

Each of the classification approaches typically represents a whole family of classifiers with several parameters and criteria that can be tuned to reach the optimal performance for a certain classification task. As it has been shown in several evaluations [Michie et al., 1994], that there is no optimal classification approach for all possible tasks due to the large variability of the data. In general, the performance and suitability of an approach depends a

lot on the actual classification problem, including aspects such as number of classes, dimensionality of the data, number of training samples, characteristics of the data distribution and complexity constraints.

The following sections briefly review the classification approaches considered in this work.

### **k nearest neighbor classifier**

The k nearest neighbor (kNN) classifier [Duda et al., 2002] is one of the simplest supervised learning approaches. Objects are classified based on a majority vote of its closest neighbors. It is based on a set of instance models, that represent the different classes.

The *training* phase simply consists of storing the training samples and the corresponding class labels in several instance models. During the *testing* phase the distances between the testing sample and all the training samples are computed. After ordering the training samples with increasing distance to the testing sample, the  $k$  closest samples are selected. Different ways have been proposed to use these neighbors to classify the unknown sample. The usual way is to assign the most common class among the neighbors, which is equivalent to majority voting. Another way is to consider the individual distance of the neighbors for the decision.

The standard parameters of a kNN include the number of neighbors considered within the majority voting and the distance metric used for the matching. Any generic distance described in section 3.3.1, or a special distance for a certain feature can be used. While a larger number of neighbors makes the kNN more robust to noise, it also makes the boundaries less distinct.

The k nearest neighbor is implicitly considered for the body matching described in chapter 5.

### **Minimum distance classifier**

The minimum distance classifier (MDC) [Duda et al., 2002] is another relatively simple classifier, that is typically used in conjunction with a set of mean models. In that case it is often called minimum distance to means (MDM) classifier.

Samples are classified based on their distance to their different class models, which can be written as

$$c = \underset{j}{\operatorname{argmin}} d(\vec{\mu}_j, \vec{x}) \quad (3.43)$$

where  $\vec{\mu}_j$  is the mean model (prototype) of class  $c_j$ .

The only parameter of a MDC is the distance metric used for computing the dissimilarity between the mean models and the unknown sample. Similar to the kNN any generic distance described in section 3.3.1 can be used.

The minimum distance classifier is used for the face recognition described in chapter 7.

### Bayes classifier

Classification can also be seen from a probabilistic point of view [Duda et al., 2002]. Given an observation the goal of the classification is then to compute the posteriori probabilities of the different classes and choose the class with the maximum a posteriori probability.

The basis for this approaches is provided by the Bayes rule, which adapted for a class  $c_j$  is given by

$$p(c_j|\vec{x}) = \frac{p(\vec{x}|c_j)p(c_j)}{p(\vec{x})} \quad (3.44)$$

with the posterior  $p(c_j|\vec{x})$ , the likelihood  $p(\vec{x}|c_j)$ , and the a priori probability  $p(c_j)$ . The evidence  $p(\vec{x})$  of the observation is constant for all classes and can be ignored for the classification task. Given that, the classification is achieved using the *maximum a posteriori* (MAP) criteria.

$$c = \underset{j}{\operatorname{argmax}} p(c_j|\vec{x}) = \underset{j}{\operatorname{argmax}} p(\vec{x}|c_j)p(c_j) \quad (3.45)$$

If the a priori probabilities  $p(c_j)$  of the different classes  $c_j$  are equal, the criteria can be simplified to the maximum likelihood criteria

$$c = \underset{j}{\operatorname{argmax}} p(\vec{x}|c_j) \quad (3.46)$$

Both the a priori and the likelihood probabilities of a class can be estimated from training data. The likelihood can be estimated using a subset of the models described in section 3.3.3. More specifically nonparametric Bayes models (naive or joint), single Gaussian models can be used.

The Bayes classifier depends on several parameters, some of which are generic to the classification criteria or others are specific to the used model. For the classification either the MAP or the ML criteria can be used. For the different models, various parameters exist which are described in section 3.3.3.

A Bayes classifier based several Bayes models is used for extracting the individual body parts within chapter 5.

### Support vector machines

Support vector machines invented by [Vapnik, 2000] are linear classifiers that find the optimal separating hyperplane between two classes. In order to support nonlinear decision boundaries, different kernel functions can be used.

Given a set of training samples with feature vectors  $\vec{x}_i$  and class labels  $y_i \in \{1, -1\}$  the SVM finds a linear function of the form

$$f(x) = \vec{w}^T \vec{x} + b \quad (3.47)$$



$$y = \begin{cases} 1 & \text{if } \vec{w}^T \vec{x} + b \geq 0 \\ -1 & \text{if } \vec{w}^T \vec{x} + b < 0 \end{cases} \quad (3.48)$$

with  $w$  being the weight vector. The hyperplane (also called decision boundary) that separates the positive and the negative training samples is then given by

$$\vec{w}^T \vec{x} + b = 0 \quad (3.49)$$

Out of the infinite number of possible hyperplanes the SVM looks for the one that maximizes the margin between the two different classes, which is equal to solve the following optimization problem

$$\frac{\vec{w}^T \vec{w}}{2} \quad (3.50)$$

Since the linear separation may not be possible on real data, due to noise and data distribution, the margin constraints are relaxed, by introducing a penalty term in the optimization function which leads to the so called soft margin SVM

$$\frac{\vec{w}^T \vec{w}}{2} + C \sum_{i=1}^n \xi_i \quad (3.51)$$

Furthermore, to deal with non linear separation the data is transformed from the input space to a higher dimensional space, in which a linear separation is possible. For the mapping different kernel functions are used that can be directly applied in the optimization function. Commonly used kernels include

$$K(\vec{x}, \vec{z}) = \vec{x}^T \vec{z} \quad (\text{Linear}) \quad (3.52)$$

$$K(\vec{x}, \vec{z}) = (\vec{x}^T \vec{z} + \theta)^d \quad (\text{Polynomial}) \quad (3.53)$$

$$K(\vec{x}, \vec{z}) = \exp(-\|\vec{x} - \vec{z}\|^2 / 2\sigma) \quad (\text{Radial basis function}) \quad (3.54)$$

$$K(\vec{x}, \vec{z}) = \tanh(k\vec{x}^T \vec{z} - \delta) \quad (\text{Sigmoidal}) \quad (3.55)$$

Since the SVM is a discriminant classifier it natively supports only two class classification. For multi class classification problems, different strategies can be applied, including one against all, one against one and error correcting codes.

Beside the penalty coefficient  $C$ , which is a general parameter of the SVM, several kernel specific parameters can be modified to influence the classification performance. Generally it is quite difficult to predict the optimal parameter set for a given task, which requires a grid search to choose suitable parameter values.

Support vector machines are used for the two-class relevance feedback within chapter 8.

### 3.4 Information fusion

The general goal of information fusion is to combine the data provided by different sources [Kittler et al., 1998] in order to improve the overall performance of the system or to generate a new data representation. More specifically the key idea is to exploit the correlation between sources to improve the performance in comparison to the individual sources. The use of information fusion can be justified by several advantages [Sanderson, 2002]:

- Utilizing complementary information may reduce error rates.
- Complexity can be reduced by using multiple simple instead of a single complex classifier.
- Sensors can be physically separated to support the acquisition from different points of view.

By considering information fusion within a system several interesting questions arise:

- Which are the most efficient classifiers for a certain application?
- Which features are the most appropriate ones for the task?
- How many classifiers are required?
- Given a set of classifiers which combination scheme improves the performance?
- How should the information provided by the different experts be fused?

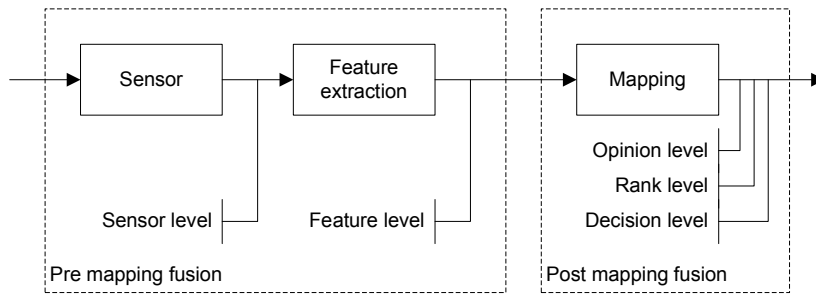
The first two questions are classical pattern recognition issues. The third question refers to the number and suitability of different experts. The fourth question deals with how the different experts are combined and the fifth question is related to the output of the different classifiers and how to fuse the information.

Existing approaches can be broadly categorized depending on their position in the processing chain relative to the mapping (learning) into *premapping* and *postmapping* fusion. These two categories can be further subdivided based on the type of information that is fused into sensor, feature, opinion, rank, and decision level fusion. Figure 3.9 explains this categorization visually, by showing the position of the different categories in a complete pattern recognition chain.

Within this work postmapping fusion methods are used for body recognition (chapter 5), face recognition (chapter 7) and multimodal person search (chapter 8).

#### 3.4.1 Premapping fusion

Premapping fusion combines the data *before* any learning step. Approaches can be subdivided into sensor and feature level fusion, depending on the type of data that is fused.



**Figure 3.9:** Position of the different fusion approaches in the overall processing chain. While the sensor level fusion operates directly on the input data (images, audio streams), feature level fusion usually combines low level features. The choice of a suitable post mapping fusion method usually depends on the output which is provided by the mapping step.

### Sensor level fusion

In sensor level fusion [Hall and Llinas, 2001] the acquired data of the individual sensors is directly combined, to obtain a joint representation. *Weighted summation* is the most common approach which requires the data to have the same dimensionality and range. For example, it can be employed to combine two images to reduce noise.

### Feature level fusion

Instead of combining the raw data, feature level fusion [Hall and Llinas, 2001] combines either similar features from multiple sensors or different features from a single sensor. Again *weighted summation* can be used to combine features with the same dimensionality and range. Another way is to *concatenate* the different features into a combined feature vector, which is used within the further analysis steps.

#### 3.4.2 Postmapping fusion

Postmapping fusion is applied *after* a learning/classification step. It can be subdivided into opinion, rank and decision level fusion, depending on the output of the different experts that are fused.

### Opinion level fusion

For opinion level fusion each expert  $e$  provides an opinion  $o_{ec}$  on each possible decision  $c$ . Opinions can be either in the form of probabilities (similarities) or distances (dissimilarities). Since different types of experts may be used, opinions need to be commensurate before further processing. Therefore, different conversion and normalization techniques can be applied (see section 3.3.1). The main advantage of opinion level fusion with respect to rank and decision level fusion, is that information regarding the goodness of a decision is considered. The opinions can be fused in two different ways [Jain et al., 2005a], either by combination or classification.

The *combination* approach fuses the opinions  $o_{ec}$  of the different experts  $e$  into a joint set of opinions  $o_c$  and applies a decision rule. Several combination rules have been proposed [Jain et al., 2005a].

The *product rule* assumes statistical independence of the different experts  $e$ . In general different biometric traits (face, voice) are mutually independent. The joint opinions are given by

$$o_c = \prod_e o_{ec} \quad (3.56)$$

Apart from statistical independence the *sum rule* also assumes that the posterior probabilities do not deviate much from the prior probabilities. Thus it is applicable if a high level of noise leads to ambiguity in the classification problem. The joint probabilities are obtained by

$$o_c = \sum_e o_{ec} \quad (3.57)$$

The *min rule* is derived by bounding the product of posterior probabilities and computes the joint probabilities as the minimum:

$$o_c = \min_e o_{ec} \quad (3.58)$$

The *max rule* approximates the mean of the posteriori probabilities and fuses the probabilities by taking the maximum:

$$o_c = \max_e o_{ec} \quad (3.59)$$

Alternatively to the combination approach a *post classifier* can be used to reach the a joint decision based on the opinions provided by the experts. The opinions of  $E$  experts regarding  $C$  classes form a  $E \times C$  dimensional feature vector which is used by the post classifier. An important advantage of the classification approach is that the opinions not necessarily need to be commensurate as in the combination approach. While any classification approach is suitable as post classifier, the most common ones are linear discriminant analysis and Bayesian classifier (see section 3.3).

## Decision fusion

The decision fusion combines the information of the different sources at the latest stage by considering decisions. Using this approach a joined decision  $d$  is reached by combining the decisions  $d_e$  of the individual experts  $e$ . Depending on the classification problem different methods can be used. While majority voting can be used for any classification problem (unary, binary, n-ary), AND and OR fusion can only be used for binary classification problems.

In *majority voting* a consensus on the decision is reached by taking the decision for which the majority of the experts agree. This can be interpreted as taking the mode of the histogram

over all possible decisions:

$$d = \text{mode}_e d_e \quad (3.60)$$

Using *AND fusion* the joint decision is only positive if all experts make a positive decision:

$$d = \bigcap_e d_e \quad (3.61)$$

This rule is quite restrictive and is usually used for applications where a low number of false positives is required.

On the other hand the *OR fusion* leads to a positive decision if at least one expert makes a positive decision:

$$d = \bigcup_e d_e \quad (3.62)$$

This is a quite relaxed criteria, which is usually used for application where a low number of false negatives is required.

### 3.4.3 Score normalization

Since the data (features, opinions) of the different source is usually heterogeneous, normalization is required to transform them into a common domain before combining them [Jain et al., 2005a; Montague and Aslam, 2001]. Score normalization achieves that by changing the location and the variation parameters of a score distribution. The required parameters can be obtained in different ways:

**Predefined:** Parameters are known a priori.

**Fixed:** Parameters are estimated offline given a fixed training set

**Adaptive:** Parameters are estimated online given an actual testing set

The good normalization should be both robust and efficient regarding estimating the location and scale parameters [Jain et al., 2005a]. *Robustness* refers to the insensitivity in the presence of outliers and *efficiency* refers to the proximity of the obtained estimate to the optimal estimate.

*Min max normalization* is the simplest normalization scheme and maps the scores into the range  $[0, 1]$ . It is defined as

$$s'_k = \frac{s_k - \min s_k}{\max s_k - \min s_k} \quad (3.63)$$

and quite suitable if the bounds are known a priori. When estimated from a given set of scores it is highly sensitive to outliers. It retains the original distributions except for scaling factor.

The *z-score normalization* is the most commonly used technique and maps 68% of the scores into the range  $[-1, 1]$ . It is based on the mean  $\mu$  and the standard deviation  $\sigma$  of a

given set of scores and is defined as

$$s'_k = \frac{s_k - \mu}{\sigma} \quad (3.64)$$

It can be expected to perform well if prior knowledge about the mean and standard deviation is available. Otherwise it needs to be estimated from a given set of scores. It is quite sensitive to outliers and assumes that the scores are distributed according to a Gaussian distribution.

The *3-sigma normalization* is an adaptation of the z-score normalization which maps 97% of the scores into the range  $[-1, 1]$ . It is defined as

$$s'_k = \frac{s_k - \mu}{3\sigma} \quad (3.65)$$

The *median absolute deviation (MAD) normalization* utilizes the median and the median absolute distance instead of the mean and the standard deviation as parameters. It is defined as

$$s'_k = \frac{s_k - \text{median } s_k}{\text{median}(s_k - \text{median } s_k)} \quad (3.66)$$

In comparison to the z-score normalization it is more robust but less efficient. It does not retain the original distributions and may not transform the scores into a common range.

### 3.5 Graph theory

As already mentioned in section 3.3 the field of pattern recognition can be divided into statistical, neural and structural approaches. While statistical and neural pattern recognition use feature vectors to represent patterns, structural pattern recognition uses symbolic data structures such as graphs, trees and strings. This makes it much more powerful in terms of representational capabilities, because any feature vector can be represented by a graph, tree or string, but not vice versa [Bunke et al., 2002]. Furthermore, symbolic data structures are able to model structural relationships between the various parts of a complex pattern, while feature vectors are limited to the joint statistical representation of individual features.

Graphs are a very powerful data structure for the representation of objects and concepts. In a graph representation the nodes typically represent objects or object parts while the edges describe their relationships [Bunke, 2000].

The following sections will summarize the basic concepts of graphs and describe how graphs can be compared to each other via graph matching.

Within this work graph matching is used for verifying the topology in the face detection (chapter 6).

### 3.5.1 Graph concepts

Formally, a *graph* is a 2-tuple  $G = (V, E)$  where  $V$  denotes a set of vertices (nodes) and  $E \subseteq V \times V$  a set of edges. An *edge*  $e = \{u, v\} \subset V^2$  is a pair of vertices  $u, v \in V$  which are connected.

Two vertices  $u, v$  are *adjacent* if they are connected by an edge  $e = \{u, v\}$ . Two edges  $e, f$  are *incident* if they have at least a common node  $v$  which can be written as  $e = \{u, v\}$  and  $f = \{v, w\}$ .

The number of vertices  $|V|$  is called the *order* and the number of edges  $|E|$  is called the *size* of a graph  $G$ . Based on the number of edges and nodes different types of graphs are defined. A *null graph* is a graph without any vertices and edges while a graph with nodes but without any edges is called a *empty graph*.

If several edges share the same vertices they are called *multiple edges* and the *multiplicity of an edge* is the number of edges that share the same vertices. A *loop* is an edge that has the same vertex at both ends. Graphs without multiple edges and loops are called *simple graphs*, graphs with multiple edges are called *multigraphs* and graphs with both are called *pseudographs*. In this work only simple graphs are considered.

Edges may have a direction. *Directed edges* are an ordered pair of vertices  $e = (u, v) \subset V^2$  with the head  $u$  and the tail  $v$ . An *undirected edge* discards any directional information and treats both vertices as an unordered pair  $e = \{u, v\} \subset V^2$ . Based on these *directed* and *undirected graphs* can be distinguished. Within this work only undirected graphs are considered.

Labels (weights) may be used on vertices and edges to identify them or indicate a meaning. Graphs with labeled vertices or edges are known as *labeled graphs*, otherwise as *unlabeled graphs*. Within this work labeled graphs are used to represent object parts and their relationships.

A graph  $G' = (V', E')$  is a *subgraph*  $G' \subset G$  of another graph  $G = (V, E)$  if its vertex set  $V' \subseteq V$  and its edge set  $E' \subseteq E$  are subsets of the other graph.

A *path* of a graph consists of a sequence of incident edges and their vertices where the terminating vertices are distinct. In contrast to that, a *cycle* refers to a sequence of incident edges where the terminating vertices are the same. The *length* of a path or a cycle is defined as the number of its incident edges.

If it is possible to establish a path between any two vertices of the graph, the graph is said to be *connected*, otherwise *disconnected*. If a graph is disconnected it consists of several *connected components* which are maximally connected subgraphs.

### 3.5.2 Graph matching

For many applications measuring the similarity between objects is an important step. If graphs are used for the object representation the problem turns into determining the similarity of graphs, which is commonly referred to as graph matching [Bunke, 2000].

Two types of graph matching can be distinguished based on the way they handle vertex and edge labels. While exact graph matching techniques are based on an exact mapping of these labels, inexact graph matching techniques provide a certain error tolerance.

### Exact graph matching

Exact graph matching provides several concepts that can be used for different matching tasks:

**Graph isomorphism:** A graph isomorphism from a graph  $G$  to a graph  $G'$  exists if there is a bijective mapping from the vertices of  $G$  to the vertices of  $G'$  that preserves all the labels of the vertices and the structure of the edges. It is useful concept to find out if two objects are the same up to the invariance properties of the underlying graph representation.

**Subgraph isomorphism:** A subgraph isomorphism from a graph  $G$  to a graph  $G'$  exists if there is a graph isomorphism between the graph  $G$  and a subgraph  $G'' \subseteq G'$ . It can be used to determine if an object is part of another object, or if an object is present within a group of objects.

**Maximum common subgraph:** A maximum common subgraph of two graphs  $G$  and  $G'$  is a graph  $G''$  that is the subgraph of both  $G$  and  $G'$  that has the maximum number of nodes among all possible subgraphs. It can be used to measure the similarity of objects even without any graphy or subgraph isomorphism. The larger the maximum common subgraph of the two graphs, the greater their similarity.

### Inexact graph matching

In the real world objects are usually affected by noise such that the graph representation of objects may not match exactly. This is especially true for labels that correspond to real valued measurements. Therefore, it is necessary to integrate some error tolerance into the graph matching process [Bunke, 2001].

Inexact (error tolerant) graph matching based on the graph edit distance provides a powerful alternative to the exact graph matching concepts introduced before [Bunke, 2001]. In its most general form a *graph edit operation* can be either a insertion, deletion, or substitution applied to vertices or edges. Edit operations are used to model the errors that change the distorted graph into an ideal one. For enhanced modeling, different costs can be assigned to the operations. They are application-dependent and must be defined based on the a priori knowledge of the underlying domain.

Inexact graph matching can be understood as a sequence  $S$  of edit operations that transform a graph  $G$  into  $G'$  such that the accumulated cost  $c(S)$  of all edit operations is minimized. The cost associated with such a sequence of edit operations is called the *graph edit*



*distance* and can be written as

$$d(G, G') = \min_S c(S) \quad (3.67)$$

Clearly, if  $G = G'$  no edit operations are needed for the transformation and  $d(G, G') = 0$ . On the other hand, the more  $G$  and  $G'$  differ from each other, the larger  $d(G, G')$ .



## Chapter 4

# Visual person description framework

### 4.1 Introduction

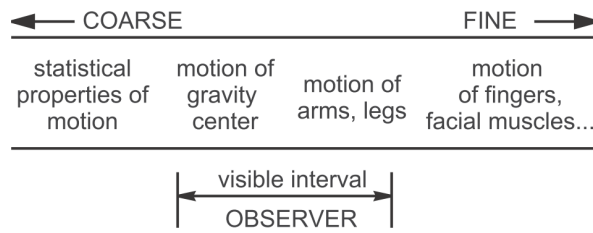
As already mentioned before, looking at people research deals with the visual analysis of humans within an environment by a machine. It involves several tasks including the detection, tracking and recognition of humans and interpreting their behavior.

Naturally the interesting information comes in form of different *channels*, such as body, face and hands. These channels are described with several visual *features* including color, texture, shape and motion. Depending on the interest and the application certain channels and features are more relevant than others.

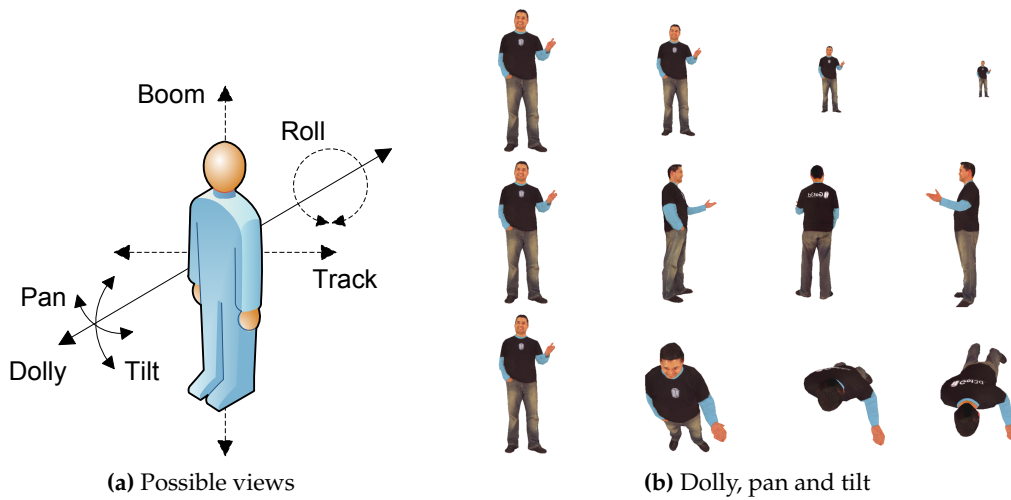
The visual analysis of humans based on channels and features has some analogy to the classical *scale space theory* [Sporring et al., 1997] which describes the fact that certain properties of an observed object appear only when observed at a proper scale. With respect to humans the scale may reach from analyzing the whole body (coarse) to individual body parts (fine). Furthermore, visual features with different complexity may be used to describe the same channel. This has been discussed for the analysis of human motion by Pers et al. [2003] and is illustrated in figure 4.1. Within a typical surveillance application the motion is usually observed at the coarsest scale in form of statistical properties. Sports analysis operates in a slightly finer scale by considering positions and velocities of the human centroid. Human computer interaction requires an even finer scale that considers the motion of hands and face. In a similar way, the human appearance can be observed at different levels, reaching from the whole body (clothes) over the face to hair and eyes.

Given a set of possible scales the choice of an appropriate scale or level of detail depends on two criteria

**External criteria:** Within the real world it is impossible to observe the full scale of information. The visible interval of the scale is determined by the camera setup (resolution, speed, field of view) and the environmental conditions (illumination, occlusions). For example, depending on the view not all scales of information may be available as illustrated in figure 4.2. While the body is more or less visible for all views the face can



**Figure 4.1:** Illustration of the scale space character of human motion analysis [Pers et al., 2003]. Even if a certain scale is available, it may not be suitable for the current task. Scale in this thesis refers to both channels and features.



**Figure 4.2:** Views as an external criterion that determines the visible interval of all possible scales. Depending on the view (distance, angle) only some channels (face, body, hands) can be observed. While the body is visible all the time, the face is only visible in certain views.



only be seen in some views.

**Internal criteria:** Depending on the interest or task a certain scale of information is required.

While a coarser scale may decrease the performance due to the lack of relevant information, a finer scale may decrease the performance as well by adding irrelevant information. For example, while facial appearance is required to verify the identity of a person for access control, it is not suitable to determine the team of a soccer player where the appearance of the body (clothes) is required.

The goal of this chapter is to derive a hierarchical framework that analyzes persons at different levels depending on external and internal criteria similar to the human visual perception. The following sections will derive a hierarchical person description based on the way how humans describe each other, provide a general system overview with the individual modules, and review anthropometrical models commonly used within the analysis modules.


PLEASE RECORD AS MUCH INFORMATION AS POSSIBLE

SEX	RACE	AGE	HEIGHT	WEIGHT	WEAPON TYPE
HAIR		 		HAT (color, type)	
GLASSES TYPE				TIE	
COMPLEXION				SHIRT	
SCARS/MARKS				COAT	
TATTOO				TROUSERS	
JEWELRY				SHOES	
AUTO LICENSE, MAKE, COLOR			DIRECTION OF TRAVEL		
ADDITIONAL INFORMATION					

Make additional copies of this page and keep them in areas that are readily available to employees.

(a) USA

YDRE KRAKTERISTIKA

KØN	ALDER	HØJDE	VÆGT
Mand <input type="checkbox"/> Kvinde <input type="checkbox"/>			
HÅR Sort <input type="checkbox"/> Brunt <input type="checkbox"/> Lyst <input type="checkbox"/>			HOVED-BEKLÆDNING (hat, hue, farve)
ØJNE			SLIPS
BRILLER			FRAKKE
TATOVERINGER			SKJORTE
AR/MÆRKER			BUKSER
UDSEENDE			SKO
BESKRIVELSE AF FLUGTBILEN (registreringsnr., mærke, farve osv.)			

(b) Denmark

**Figure 4.3:** Questionnaires for describing suspicious persons in a surveillance scenario from two different countries. Both contain 3 categories of person descriptions: semantics, head/face and body/clothes.

4.2 Human description

The major goal of the developed framework is to detect and describe humans at different levels to provide a flexible visual description that is close to the human visual perception. In order to allow the machine to extract a suitable description a closer look at how humans describe each other is necessary.

One can find a very good example on how humans describe each other in surveillance applications. The forensic analysis usually starts with a victim describing the incident together with the involved humans. Figure 4.3 shows two typical questionnaires from different countries (USA, Denmark) for describing a suspicious person. Although the languages differ it is evident that both descriptions are quite similar. Besides additional information (weapon, car) they focus mainly on 3 different information channels with respect to humans:

**Semantics:** This group includes high level descriptions such as identity, ethnicity, age and gender of the person. It is usually determined by a joint audiovisual analysis (face, speech) together with available a priori knowledge.

**Head/face:** In contrast to the previous group it combines low and mid level descriptions with respect to the head or face of a human. It is usually based on describing the color, texture and shape of facial components (eyes, nose, mouth), special features (scars,

beard, hair) and accessories (glasses, jewelry).

**Body/clothes:** Similar to the previous group it combines low and mid level descriptions with respect to the body or clothes. It may consider color, texture and shape to describe the appearance and type of clothes (hat, shirt, coat, trousers, shoes) as well as the built of the body (height, weight).

Given that, a typical description of a suspect in a *surveillance* scenario could be:

“The person of interest is a male European between 30 and 35 years old. He has bright skin, short red hair and was wearing sun glasses. He is of average height and built and was wearing a blue shirt with a white logo and dark trousers.”

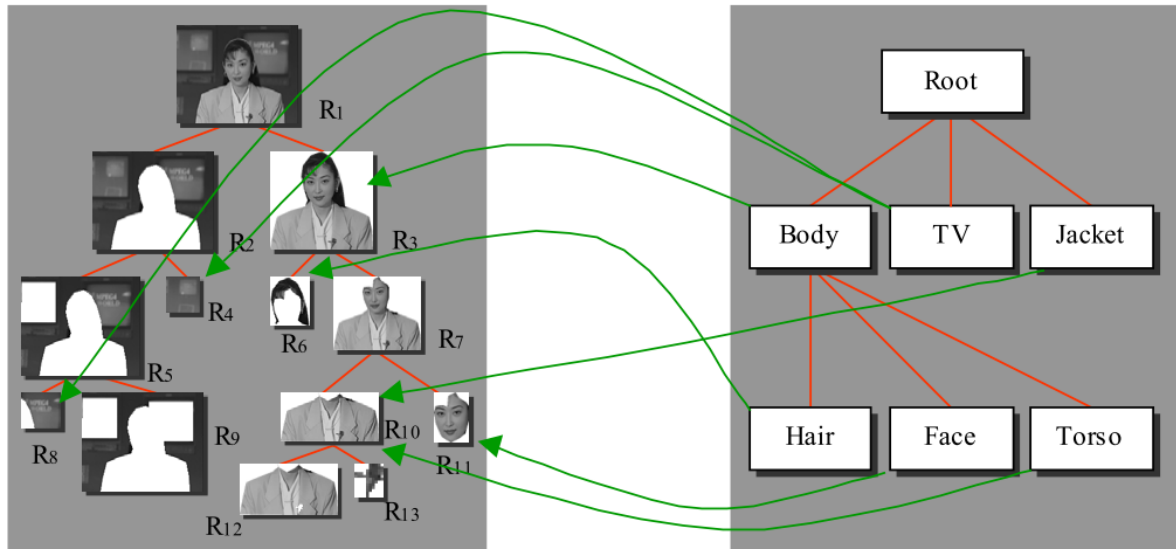
Although other scenarios (sports, human computer interaction, smart room technologies) may require different descriptions, these are usually a subset of available channels (face, body, hands) and features (color, texture, shape, motion). Given a complete set of information the choice of the appropriate scale of information depends then on *internal criteria*. For example, while face appearance is typically used to identify humans. it is not appropriate for recognizing that the person belongs to the security company. Therefore, the appearance of the body (clothes) is the suitable criterion. This examples illustrate how humans choose the most appropriate scale based on internal criteria.

Not all of the above mentioned information may be available due to *external criteria*. Humans usually focus on the most reliable information that is available. For example, if one wants to identify another person visually, he will start with the face since this is the most reliable cue. If the face is partially covered, we usually concentrate on the visible parts. If the face is fully covered or the person is too far away, we may look for other traits such as gait or body built if we have not seen the person recently. On the other hand, if we just met the person some minutes ago, we may concentrate on the clothes since they are quite reliable for a short time span. These examples illustrate how humans choose the most appropriate information scale depending on external criteria.

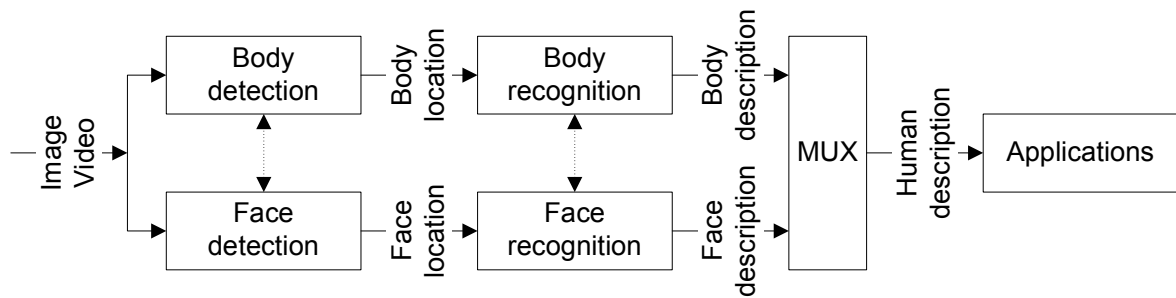
### 4.3 System overview

Within this work a hierarchical framework is proposed that describes humans at different levels which are inspired by human visual perception. By doing this it supports a large variety of applications with different interests and environmental conditions.

The idea can be interpreted as a special case of the hierarchical multimedia description scheme proposed by Salembier et al. [1999] and integrated in the recent MPEG-7 standard [Manjunath et al., 2002]. The basic idea is to decompose a multimedia document (image, video, audio) hierarchically into a set of syntactic regions and corresponding semantic objects that are described individually with suitable low and high level features. This idea is illustrated in figure 4.4 for a single video frame. The region tree on the left side decomposes



**Figure 4.4:** Sample of a generic hierarchical visual description with the low level region tree (left) and the corresponding high level object tree (right) [Salembier et al., 1999]. The hierarchical person description can be seen as a way for automatic extraction of the body branch.

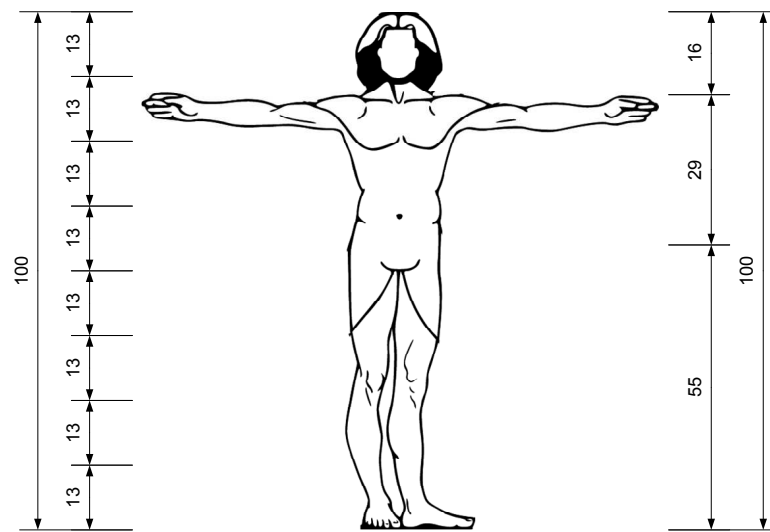


**Figure 4.5:** Overview of the proposed hierarchical human analysis framework with the channels along the vertical axis and the tasks along the horizontal axis. The extracted human description can be used for different applications and conditions.

the overall image into low level regions of finer granularity while the object tree provides a decomposition of the scene into the corresponding high level objects. In contrast to the work of Salembier et al. [1999] where the links between the region and the object tree are established manually, the idea is to consider the special case of decomposing a human into interesting regions (parts) and develop ways to extract the description automatically.

Figure 4.5 provides an overview of the developed system that considers body and face analysis. Given a video frame or still image the detection stages detect and segment the body and the face of visible humans and provide their locations to the corresponding recognition stages. Each of these recognition stages extracts a rich appearance description of the corresponding part which are combined into a hierarchical human description. Finally, the extracted descriptions can be used for a large variety of tasks and applications.

Since body detection was outside the scope of this thesis, the background segmentation approach developed by Mustafa Karaman was adopted throughout this work. It is



**Figure 4.6:** Idealized (left side) and simplified (right side) body anthropometry imposed on the “Vitruvian Man” [da Vinci, 1487]. While the left model is inspired by the human anatomy, the right model is based on the human appearance which is largely dominated by clothes.

not described within this thesis, but parts of it are described within several joint publications including [Karaman et al., 2006, 2009] and a complete description can be found in his dissertation [Karaman, 2009].

## 4.4 Human anthropometry and modeling

Anthropometry refers to the measurement of the human body across individuals for understanding physical similarities and variations. Nowadays, it plays an important role in industrial and clothing design, ergonomics, and architecture where body dimensions are used to optimize products. In the looking at people research domain average human measures are commonly used to initialize human models such as face and body models.

### 4.4.1 Body anthropometry

The first description of human body proportions goes back to the “Vitruvian Man” created by da Vinci [1487]. According to his work an idealized human body can be split vertically into 8 equally sized parts as shown on the left side in figure 4.6 limited by the following body features: (a) chin (b) chest (c) cavel (d) pubis (e) thigh (f) calf (g) ankles (h) feet.

For describing the appearance of the human body a simpler body anthropometry as shown on the right side in figure 4.6 is derived that splits the human body vertically into 3 parts: (a) head (b) upper body (c) lower body. This is motivated by the fact that each of these body parts usually has a uniform appearance which not necessarily means a uniform color or texture. This body anthropometry has been widely used for body detection [Dalal, 2006; Wu, 2008] and human motion analysis [Park and Aggarwal, 2002].





**Figure 4.7:** Body model with holistic representation (red) and component based representation (blue). Both representations are considered for the body recognition.

Based on this body anthropometry a hierarchical body model is derived that describes the human body at two levels:

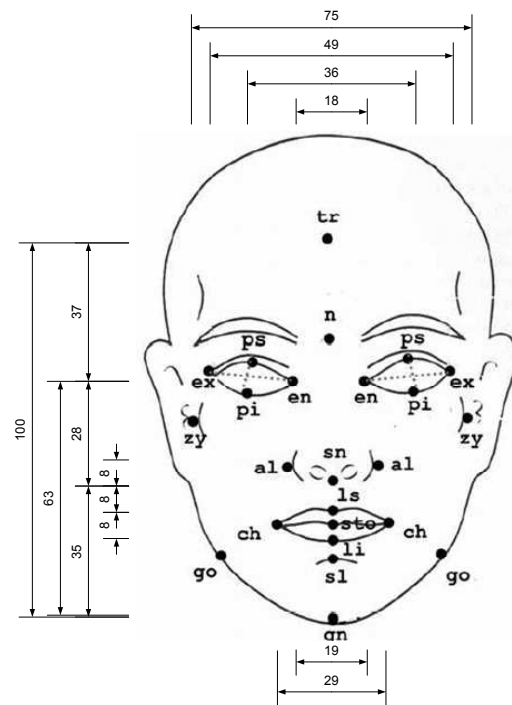
**Holistic representation:** The human body is described as a single arbitrarily shaped region with equal importance given to all pixels inside the binary support map. While this representation is naturally less flexible it is supposed to be less complex than the component based representation. Since the variation across the whole body is usually much larger than within individual body parts, more complex features may be required for the description of the whole body.

**Component based representation:** The human body is described as a set of 3 arbitrarily shaped regions that correspond to individual body parts (head, upper body, lower body). In that way it is more flexible and can be adapted based on internal and external criteria. An example for an internal criteria might be that a victim remembers only the color of the burglars jacket. In that case only the upper body provides enough information for the search task. In the presence of occlusions due to other objects which can be seen as an external criteria only the non-occluded body parts maybe considered for the further analysis.

Figure 4.7 shows a body sample with the holistic representation (red) and the component based representation (blue).

#### 4.4.2 Face anthropometry

The most influential work regarding face anthropometry has been done by Farkas [1994]. It consists of several distances measured between a set of well-defined facial features (shown



**Figure 4.8:** Face anthropometry with most important features and average distances [Farkas, 1994]. These distances are commonly used to initialize face models.

in figure 4.8) across a large population. The average vertical and horizontal distances between these points are commonly used to initialize face models within the face analysis domain.

Based on this face anthropometry a hierarchical face model is derived that describes a face on two levels:

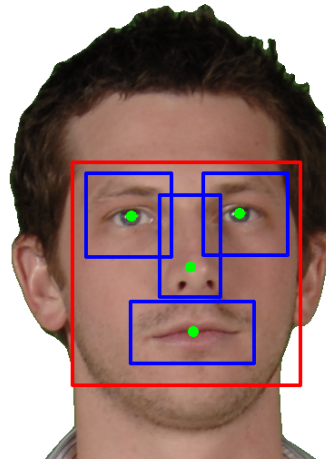
**Holistic representation:** The face is described as a single rectangular patches with equal importance given to any pixel within it. Thus, it can not be adapted according to internal or external criteria which makes it less flexible. On the other hand the detection and recognition based on this representation is naturally less complex. Furthermore, it can be applied for the detection and recognition of faces at lower resolutions.

**Component based representation:** The face is described as a set of four rectangular patches corresponding to the most important facial components (right eye, left eye, nose and mouth). The possibility to assign weights to each of the components makes it more flexible with respect to internal or external criteria. As an example for an external criteria occlusions have been considered within the developed face detection (see section 6) and face recognition (see section 7) approaches. On the other hand this representation is more complex and requires a higher resolution.

The exact definition of the rectangular patches corresponding to the holistic and the component based representation are provided in table 4.1. The dimensions are relative to the

Part	xmin	ymin	xmax	ymax	xdim	ydim	size
Face	0	0	75	100	75	100	7500
Whole	0	12	75	87	75	75	5625
Right eye	5	20	35	50	30	30	900
Left eye	40	20	70	50	30	30	900
Nose	25	30	50	70	25	40	1000
Mouth	10	65	65	90	55	25	1375

**Table 4.1:** Definition of the rectangular patches for the holistic and the component based representation relative to the anthropometric face region.



**Figure 4.9:** Face model with facial features (green), holistic representation (red) and component based representation (blue). Both representations are used for the face detection and the face recognition.

anthropometric face region with the width equal to “zyzy” and the height equal to “trgn” as shown on figure 4.8.

Figure 4.9 shows a face samples with the considered facial features (green), the compact holistic representation (red) and the flexible component based representation (blue).

## 4.5 Conclusion

### 4.5.1 Summary

This chapter describes the proposed hierarchical human analysis framework and its motivation. It starts by reviewing the different channels and visual features commonly used for human analysis. Based on that, it discusses the analogy to the classical scale space theory, where the appropriate choice of scale is based on internal and external criteria. A hierarchical human analysis framework is proposed, that combines visual analysis at different levels to support a large variety of interests and applications. This thesis focuses only on two channels (body, face) and their appearance (color, texture) and describes the modules

for the detection and recognition and their use within different application scenarios. The person model and visual features are chosen in way that they mimic the person description by humans which has been discussed for a surveillance scenario. Finally, commonly used body and face models are derived from human anthropometric models.

#### **4.5.2 Future work**

Although this work considers only two body parts (body, face) and their appearance (color, texture) the idea can easily applied to other body parts (hands) and visual features (shape, motion). Therefore, tracking modules are required for the different channels beside the detection and recognition modules. Furthermore, additional channels with finer granularity may be required. Two examples that consider other channels and features beside the appearance of the body and the face are the personalized human computer interface described in chapter 10 which analyzes also the motion of the hands and the multimodal person search described in chapter 8 that adds voice characteristics as another channel.

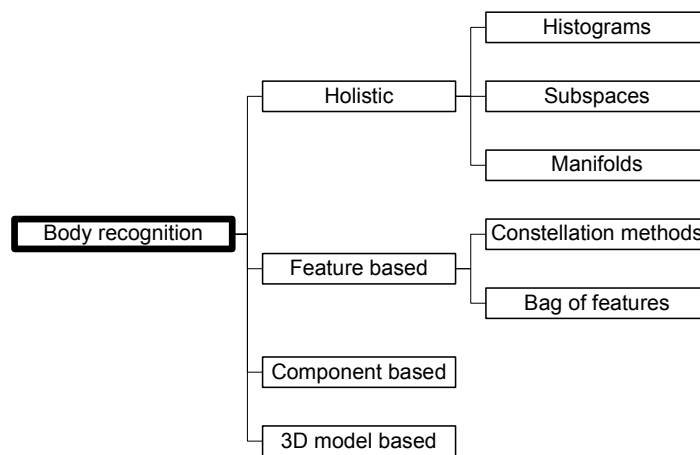
## Chapter 5

# Body recognition

### 5.1 Introduction

The human body provides several traits that can be analyzed visually including shape, motion, and appearance. Each of these traits provides information that can be used for different tasks. *Motion* can be analyzed at different granularities. The global motion trajectory of the human body can be used to analyze the activity of a human within an environment which is useful for surveillance applications. The local motion of different body parts (e.g. arms) can be interpreted to recognize the behavior of humans. Finally, analyzing the gait of a human can be used to recognize the identity. The shape of the human body provides information that can be used for recognizing the identity (height, build) and the posture. In contrast to these task, the focus within this chapter is laid on appearance based body analysis which in contrast to motion based body analysis is applicable for both images and videos. Since the appearance of a persons body is largely dominated by clothes which can be exchanged, it can only be used in short time scales for identifying humans. On the other hand, clothes may carry additional information such as a group or a team the person belongs to.

Appearance models of humans can be used for several tasks and applications. Within a single camera view they can be used to assist object tracking through and after occlusions by providing additional information apart from the motion and shape characteristics usually considered for tracking [Balcells Capilades, 2004]. For multiple cameras they can be used to match objects between non-overlapping camera views and merge intra camera tracks into inter camera tracks [Teixeira and Corte-Real, 2008]. Within a surveillance scenario the appearance information can support forensic search by prefiltering present persons based on the description provided by witnesses [Hansen et al., 2007]. Finally another application is to group persons based on the appearance of their clothes into certain groups or teams, which is of major interest for sport and entertainment applications [Ekin and Tekalp, 2003].



**Figure 5.1:** Taxonomy of recent body recognition approaches based on the used body representation. The developed approach combines a holistic and a component based representation with each other.

### 5.1.1 Related work

In contrast to face recognition, body recognition is a rather unexplored field. Nevertheless, it is related to the more generic field of appearance modeling, that deals with the description and recognition of various objects. The current state of the art can be divided into three major categories (see figure 5.1) depending on how the humans are represented [Gray et al., 2007]:

**Holistic approaches:** These approaches consider a person as a whole and include methods such as templates [Stauffer and Grimson, 2001], subspace methods [Black and Jepson, 1998], and manifolds [Murase and Nayar, 1995]. Furthermore, visual features commonly used for content based image retrieval are adopted for body description and recognition [Nakajima et al., 2000; Hähnel et al., 2004].

**Feature based:** These approaches consider a set of local descriptions, which are extracted at certain interest points (e.g. corners, blobs, and edges) that can be robustly detected across view and illumination changes [Tuytelaars and Mikolajczyk, 2006]. The description of these feature points is usually based on invariant features such as Haar features [Viola and Jones, 2001a], image patches [Bart et al., 2004] or features based on the scale invariant feature transform (SIFT) [Lowe, 2004]. While methods based on the bag of words [Teixeira and Corte-Real, 2008] paradigm discard any spatial information regarding these features points, constellation methods [Fergus et al., 2003] consider the relative spatial arrangement of them.

**Component based approaches:** These approaches usually exploit the anthropometry of the human body (see section 4.4.1) to extract several body parts and describe them independently [Annesley et al., 2005]. This is motivated by the fact that individual pieces of clothes (shirt, trousers) are usually homogeneous in color and texture [Hansen et al.,

Key	Representation	Description	Mapping	Fusion
Nakajima et al. [2003]	Holistic	Color histogram, local shape features	Support vector machine	None
Jaffre and Joly [2004]	Holistic	Color histogram	Bhattacharyya coefficient	None
Hähnel et al. [2004]	Holistic	Color histogram, color structure, quadrature mirror filters, oriented Gaussian derivatives, homogeneous texture, edge histogram	Neural network, k nearest neighbor	Feature fusion
Annesley et al. [2005]	Holistic, components	Dominant color, color layout, scalable color, color structure	Several distance metrics	Part and feature fusion
Gheissari et al. [2006]	Holistic, components, features	Color histogram, edgel histogram	Several distance metric	None
Gray et al. [2007]	Holistic, components	Color histogram, color correlogram, template, subspace	Bhattacharyya coefficient	Part fusion
Hansen et al. [2007]	Components	Color names	None	None
Teixeira and Cortes-Real [2008]	Features	SIFT features, bag of words	Support vector machine	None

**Table 5.1:** Overview of recent body recognition approaches based on certain criteria with older methods (upper part) and recently proposed methods (lower part). It can be seen that the research trend moved towards component or feature based methods.

2007].

An overview of selected body recognition approaches is given in table 5.1. It compares them based on several criteria, including the human representation, visual description and fusion method. While the upper part contains older methods proposed before this work, the lower part contains newer approaches that have been published recently.

### 5.1.2 Challenges

Appearance based body recognition methods have to deal with several challenges, including different camera views, varying illuminations and changing pose. Furthermore, the articulated motion of the human body and the elastic motion of the clothes [Aggarwal et al., 1994] may lead to rapid appearance changes and partial occlusions. Below the individual challenges are discussed in more detail:

**Illumination:** These variations are usually caused by the intensity and color of the light within an environment. To obtain a robust description one can either try to compensate for these variations by applying contrast stretching [Gonzalez and Woods, 2007] and color constancy methods [Funt and Finlayson, 1995] or extract a coarse description that is invariant to different illuminations.

**Views:** Viewpoint invariance can be interpreted differently and mean anything from imperfect alignment to invariant regarding radical view and scale changes. One way to handle different views is to use features that discard any spatial information such as

color histograms. Nevertheless this may decrease the discriminability. Another way is to utilize a more robust representation such as components and features that can be extracted under a large variety of views.

**Occlusions:** In the case of the human body occlusions can be caused by the body parts itself or by other objects present in the environment. In any case they typically change the appearance of the human body leading to a drop in the recognition performance. There are basically two ways to deal with occlusions. The first one ignores them based on the assumption that occlusions happen only occasionally. Within a video the performance can then be increased by simply combining multiple samples by majority voting (see section 3.4). Another strategy is to build an occlusion aware body analysis framework that allows to detect and localize occlusions of a human body. Based on this information only non-occluded bodies (sample selection) or body parts (partial sample) can be used for description and recognition.

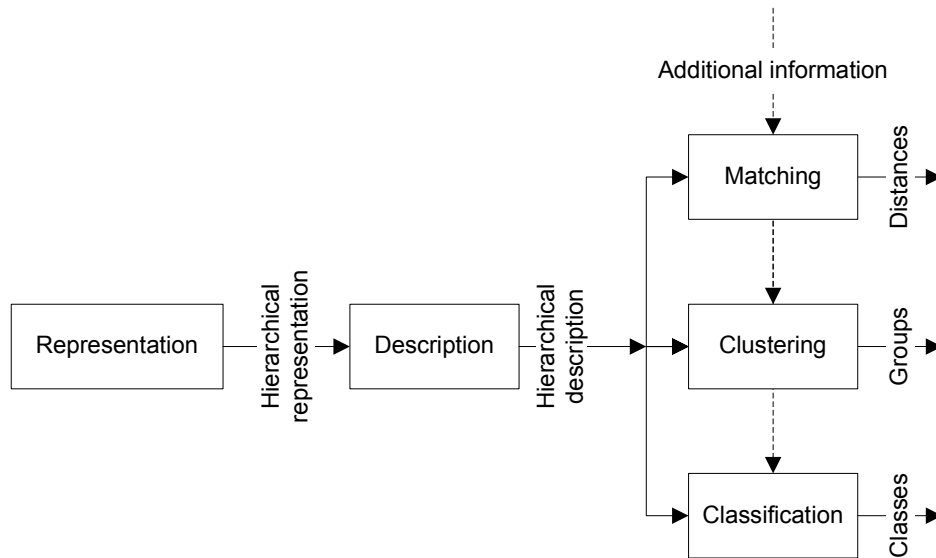
### 5.1.3 Objective

The major objective of this work is to develop a generic and robust approach for describing and recognizing the appearance of the human body at different levels. Therefore, a region based description of the human body is proposed, that allows to combine holistic and component based approaches in a straightforward way. This is again motivated (see section 4 for a more general discussion) by the idea of hierarchical visual description schemes [Salembier et al., 1999] incorporated in the MPEG-7 standard [Manjunath et al., 2002] which correspond well to the coarse-to-fine human visual perception.

## 5.2 Approach

Figure 5.2 provides an overview of the proposed body description and recognition module. Given an image or video frame and the segmentation mask provided by the body detection module a hierarchical body representation is extracted by splitting the whole body into a set of predefined body parts (head, upper, lower). Several methods have been developed for that purpose which are described in section 5.2.1. Together with the corresponding image this body model is given to the description stage, which extracts a set of visual features for each of the body parts individually. The considered visual low-level features are described in section 5.2.2. The resulting hierarchical body description is stored in a database and may be used for various applications. The recognition stage described in section 5.2.3 supports different functionalities including matching, clustering and classification of persons based on the extracted body descriptions.





**Figure 5.2:** Overview of the body recognition module. The representation step takes a binary segmentation mask from the body detection and extracts a hierarchical model that consists of the whole body split into head, upper and lower body. The description step extracts a set of suitable low-level features for these regions. Depending on the application the recognition may be based on matching, clustering or classification techniques. Finally, parts and features may be fused to improve the overall performance.

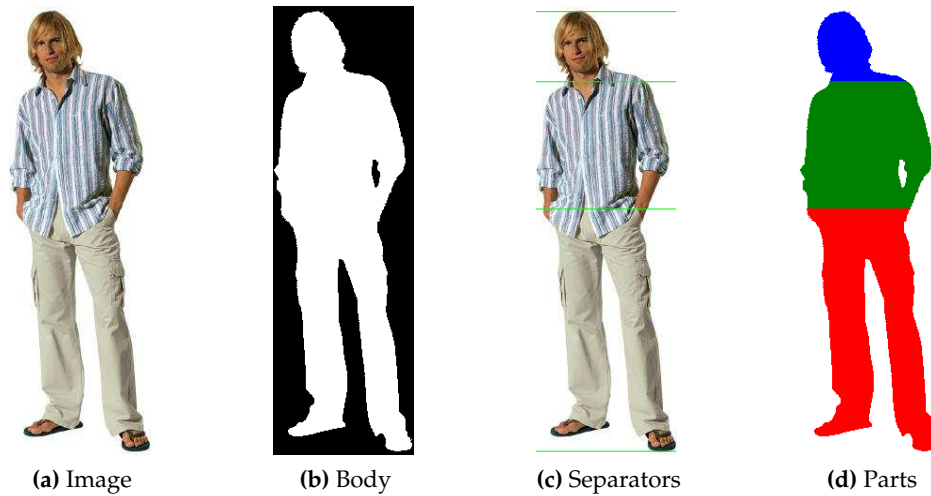
### 5.2.1 Representation

The goal of the representation stage is to extract a hierarchical body model for the detected person based on the segmentation mask  $M_s$  describing the whole body and the image  $I$ . This model is simply a combination of the holistic model with a component based model to provide a variable and adaptive description of the bodies appearance. In contrast to motion based body recognition where an articulated body model with several rigid segments connected by joints is used a much simpler body model is considered here that divides the body in head (H), upper body (U) and lower body (L). The extraction of these two representations is described in the following sections.

#### Holistic

The holistic representation describes the human body as a whole without considering any structure. It basically ignores that the object of interest is a human and can thus be used for any other object as well. While it provides less structural information than the component based representation both the extraction and the description are less complex.

The extraction of the holistic representation is straightforward. With the assumption that the segmentation mask  $M_s$  describes the whole person with only some minor segmentation faults, the whole body mask is simply defined as  $M_w = M_s$ . To avoid unreliable descriptions of partially occluded humans, segmentation masks touching the boundaries of the cameras field of view are not considered.



**Figure 5.3:** Illustration of the top-down body modeling approach. Based on the binary body mask (b) and the body anthropometry horizontal separators (c) are computed that split the body mask into individual parts (d).

### Components

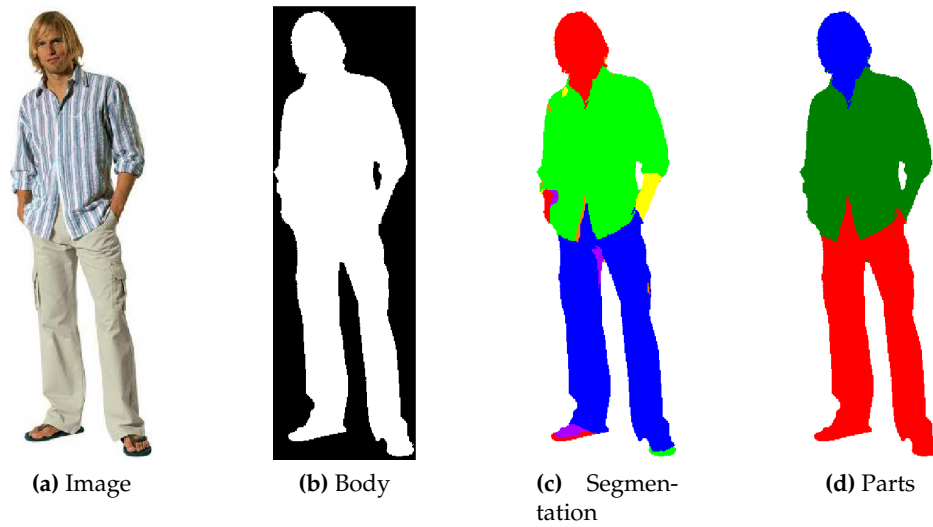
The component based representation describes the human body based on three individual including head (H), upper body (U), lower body (L). This is done by segmenting the whole body mask  $M_w$  into three individual masks  $M_h$ ,  $M_u$  and  $M_l$ .

Three different methods have been developed for this segmentation that differ mainly in how they consider the body anthropometry (described in section 4.4). Each of these methods has different characteristics including segmentation accuracy and reliability under the large variety of costumes and conditions.

**Top-down extraction** The so called top-down approach, is solely based on the *body anthropometry* (see section 4.4) and does not consider any visual information. Thus it uses only the binary segmentation masks that describes the whole body to extract the individual body parts. This approach is similar to the one used by Park and Aggarwal [2000, 2002].

Given the binary segmentation mask of the whole body, provided by the body detection module, the bounding box is extracted. Together with a set of vertical ratios that define the location of the individual body parts in relation to this bounding box, a set of bounding boxes that correspond to the individual parts is computed. The binary segmentation masks of the individual parts is then obtained by simply cropping these regions from the binary masks of the whole body. This process is illustrated in figure 5.3 for a sample of the Neckermann Database (described in section A.2.1).

As it can be seen already from the example, this approach is not very precise in segmenting the different body parts with respect to the clothes. This is not very surprising since it does not consider any appearance information for the segmentation and relies only on the



**Figure 5.4:** Illustration of the bottom-up body modeling approach. Given the body mask (b) the image is segmented into homogeneous regions (c). Based on the body anthropometry these regions are merged into the corresponding parts (d).

shape of the whole body described by the binary segmentation mask. On the other hand, this makes it quite robust to color and texture similarities between the different body parts.

**Bottom-up extraction** The bottom-up approach starts by segmenting the image into a set of coherent regions and groups them based on the whole body mask and the body anthropometry into individual body parts which is illustrated in figure 5.4 for a sample of the Neckermann Database. This approach is comparable to the one proposed by Hansen et al. [2007], but differs both in the segmentation and in the grouping stage.

The extraction starts by partitioning the image  $I$  into a set of coherent regions  $R$  using an *image segmentation* method. Several approaches have been considered including optimized mean shift (MS) [Bailer et al., 2005], region based automatic segmentation (RBAS) [Adamek et al., 2005], modified recursive shortest spanning tree (MRSST) [Adamek and O'Connor, 2007] and spatio-temporal video segmentation (SEG2DT) [Galmar and Huet, 2006]. Based on a comparison published recently [Goldmann et al., 2008a] the MRSST approach was chosen for this task. The MRSST approach is an extension of the well known recursive spanning tree (RSST) algorithm [Alatan and Onural, 1998] that considers additional homogeneity criteria (global and local shape complexity, region adjacency and total inclusion) beside the color homogeneity criteria. The merging order of the regions is determined by fusing the different criteria using Dempster Shafer (DS) theory [Smets et al., 1988], which takes into account the reliability of the different sources of information. The merges are recorded in a binary partition tree (BPT) which can be interpreted as a hierarchy of image partitions. In order to select a single partition (set of non-overlapping regions) that reflects meaningful image content a suitable stopping point is chosen by locating a corner in the accumulative

merging cost curve. The final partition  $R$  is then obtained by selecting the corresponding regions from the BPT.

The next step selects a subset  $R' \subset R$  of this partition based on the whole body mask  $M_w$  and groups them in individual body parts  $p$  using the body anthropometry. Two methods have been developed for this task. The *centroid based* method assigns the remaining regions  $r \in R'$  to the different body parts  $p$  by comparing their centroids  $c_r$  to the bounding boxes  $b_p$  of the body parts. More specifically a region  $r$  is assigned to a body part  $p$  if the centroid  $c_r$  lies inside the bounding box  $b_p$ . In the *overlap based* method the assignment is based on the pixel-wise overlap between the masks of the regions  $M_r$  and the masks of the body parts  $M_p$  derived from the whole body mask  $M_w$  as described in the top-down approach. This is similar to the figure segmentation approach proposed by Ge et al. [2005] for image segmentation evaluation. For a region  $r$  and a part  $p$  the matching criteria  $\rho(r, p)$  is computed as

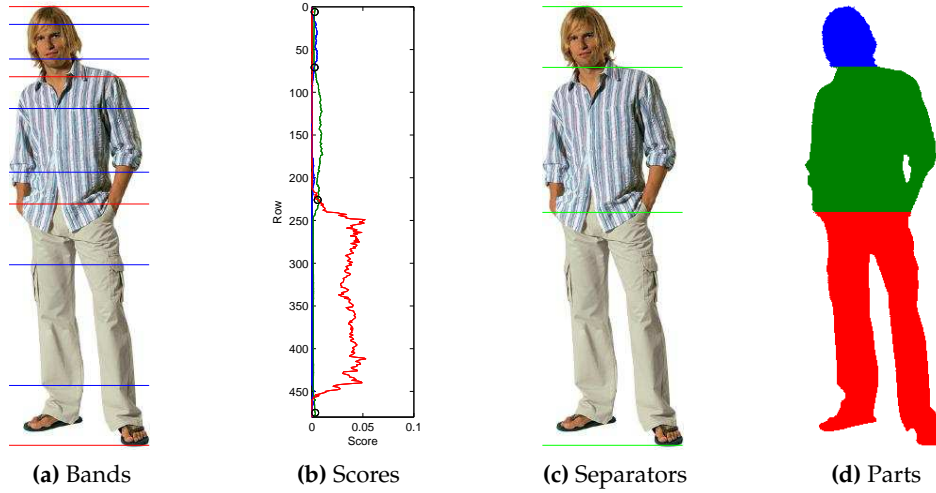
$$\rho(r, p) = \max \left\{ \frac{\text{area}(M_r \cap M_p)}{\text{area}(M_p)}, \frac{\text{area}(M_r \cap M_p)}{\text{area}(M_r)} \right\} \quad (5.1)$$

If  $\rho(r, p) > 0.5$  the region is assigned to the body part by including it into the set  $R'_p \subset R'$ . The threshold ensures that every region is only assigned to a single body part. No matter what assignment approach has been used the masks of the different body parts  $M_b$  are then computed by combining the masks  $M_r$  corresponding to the assigned regions  $R'_p$  with an OR operator

$$M_p = \bigcup_{r \in R'_p} M_r \quad (5.2)$$

As it can be seen from the sample in figure 5.4 the bottom-up approach may achieve a very precise segmentation of the body parts since it considers both the shape and the appearance for the extraction. Furthermore, it utilizes nonlinear boundaries between the different body parts, which decreases the confusion between them. The major problem, as it will be shown in the experiments (section 5.3) is its decreased robustness if body parts are visually similar. This causes the image segmentation step to merge regions belonging to different body parts.

**Hybrid extraction** The idea behind the hybrid approach is to combine the two previous approaches, by integrating high-level body anthropometry and low-level image segmentation. This is achieved by estimating individual appearance models for each body part from confidence zones and predict the body part for each pixel within the transition zones. Based on the predictions the location of the horizontal separators from the top-down approach are shifted vertically to reduce the confusion between the different body parts. This idea is illustrated in figure 5.4 for a sample of the Neckermann Database. This approach is inspired by the work of Elgammal and Davis [2001] which considers a similar approach for segmenting persons under occlusions. Nevertheless, our approach differs both in the modeling and the localization step.



**Figure 5.5:** Illustration of the hybrid body modeling approach. Based on the body mask and the anthropometry confidence and ambiguity zones are defined (a). For each of the body parts a color model is learned from the confidence zones and applied to the transition zones. Based on the derived row scores (b) for each part the horizontal separators (3) are adjusted vertically to minimize the confusion between neighboring body parts. Finally, the separators are used to split the body into the individual parts (d).

Similar to the bottom-up approach, the hybrid approach considers both the image  $I$  and the whole body mask  $M_w$  for extracting the body part masks  $M_p$ . It starts by computing the initial separators  $s$  (red horizontal lines in figure 5.4(a)) in the same way as the top-down approach. These initial separators are used to derive the confidence zones  $z_p$  for each body part  $p$  (areas between two blue horizontal lines in figure 5.4(a)) defined as the sub areas centered between separator pairs  $(s_p, s_{p+1})$ . Based on the corresponding  $z_p$  the color distribution of each part is modeled using a joint Bayes model  $m_p$  (see section 3.3.3). These models are used to predict the probabilities  $s_p(i, j)$  of the different body parts for the pixels within the transition zones  $\bar{z}_p$ . Each pixel is then assigned to a body part using the ML decision rule defined as

$$d(i, j) = \arg \max_p s_p(i, j) \quad (5.3)$$

In order to compute the refined separators  $s'$  (green lines in figure 5.4(c)) row scores  $s_p(i)$  are computed for each of the body parts by combining the pixel-wise scores or decisions over the columns. The *soft decision method* computes the row scores  $s_p(i)$  by averaging the pixel scores  $s_p(i, j)$  over the columns  $j$ . The *hard decision method* computes the row scores  $s_p(i)$  for each part  $p$  as the ratio of pixels with the decision  $d(i, j) = p$  and all the pixels in the row  $i$ . By analyzing the row scores  $s_p(i)$  over the rows  $i$  (see figure 5.4(b)) for adjacent body parts several crossing points can be determined, where the curve of one body part crosses the curve of the body part below it. These crossing points determine the location of the refined separators  $s'$ . Noise and visual similarities between the body parts may cause several crossing points around the optimal separator position. For increased stability the

crossing point corresponding to the median of all crossing points is chosen as the separator position. Based on refined separator pairs  $(s'_p, s'_{p+1})$  the body part masks  $M_p$  are extracted by cropping the corresponding regions from the whole body mask  $M_w$  similar to the top-down approach.

As expected, the precision of the hybrid approach is somewhere between the top-down and the bottom-up approach (compare figure 5.5 with 5.3 and 5.4). Although the refinement improves the segmentation accuracy, it is still based on horizontal boundaries that cause some confusion between the body parts. Nevertheless this confusion is minimized by analyzing the appearance of the individual body parts. The hybrid approach retains the robustness of the top-down approach, since it considers the body anthropometry as the primary criterion.

### 5.2.2 Description

The appearance of a human body is usually dominated by the color and texture of the worn clothes. The variety of color is usually quite large and may include different shades and tones. Furthermore, clothes may have very different textures depending on the material and design. One usually distinguishes between homogeneous textures such as stripes and non-homogeneous textures such as a logos or appliques. The shape or size of a piece of clothes may allow to distinguish between different types of clothes. But this is not considered here.

A large number of visual features has been proposed within the context of content based multimedia retrieval which can be grouped into color, texture, shape and motion features. Only features belonging to the former two groups are considered here. Within the scope of the MPEG-7 standard some of the proposed features have been compared to each other to determine a rich and robust set of visual descriptors for content based multimedia retrieval. For the description of the human body both standard and non-standard features are considered.

In general, visual features can be extracted from the whole image or parts of it. Spatial regions can be either rectangular patches or regions with an arbitrary shape. Within the field of content based image retrieval features are usually extracted from the whole image or in predefined rectangular regions due to the difficulties of reliable image segmentation. Body description and recognition is more restricted and allows to utilize a priori knowledge for the segmentation. For describing arbitrary shaped regions most of the visual features need to be extended. While the extension of pixel-wise features (e.g. color moments, color histograms) is straightforward, the extension of other features (e.g. homogeneous texture descriptor) is not as easy. Nevertheless, all features have been extracted for the arbitrary shaped regions described by the binary masks.

The features that have been considered for the description are summarized in table 5.2. The major idea was to have a representative set of standard and non-standard color and texture features with different complexity and characteristics. The extraction of these fea-

Category	Feature	Abbrev.	Standard	Dimensions	Matching
Color	Average color	AC	None	3	L2
Color	Color moments	CM	None	12	L2
Color	Color histogram	CH	None	512	L1
Color	Color coherence vector	CCV	None	1024	L1
Color	Color spatiogram	CS	None	2560	L1
Color	Scalable color descriptor	SCD	MPEG-7	64	L2
Color	Color structure descriptor	CSD	MPEG-7	32	L1
Color	Color layout descriptor	CLD	MPEG-7	18	L1
Texture	Intensity moments	IM	None	4	L2
Texture	Cooccurrence matrix	CM	None	48	L2
Texture	Tamura feature	TF	None	3	L1
Texture	Homogeneous texture descriptor	HTD	MPEG-7	62	L1
Texture	Edge histogram descriptor	EHD	MPEG-7	80	L1

**Table 5.2:** Overview of the considered visual low level features grouped into color and texture types as well as standard (MPEG-7) and non-standard ones.

tures is concisely described in the following sections and the corresponding references are provided.

### Average color (AC)

The average color (AC) is the simplest color feature, which corresponds to the mean of the color distribution within an image. While it is very efficient for uniform regions it usually provides a rather poor description of regions with complementary colors.

Given an image  $P$  with its pixels  $p_{ijk}$  the first statistical moment (mean)  $m_k^1$  for an individual channel  $k \in [1, K]$  is defined as

$$m_k^1 = \frac{1}{W \times H} \sum_{i=1}^H \sum_{j=1}^W p_{ijk} \quad (5.4)$$

The resulting feature vector  $f_{AC}$  is simply the concatenation of the means of the individual channels given as

$$\vec{f}_{AC} = (m_1^1, m_2^1, \dots, m_K^1) \quad (5.5)$$

with the dimensionality  $D = K$ .

The two parameters that influence the extraction are the color space and the number of channels used to represent the image.

Since the average color describes the color distribution only by the mean it is only suitable to describe regions with homogeneous colors or gradients. It is a rather rough description and does not consider any spatial information.

### Color moments (CM)

The major idea of color moments [Stricker and Orengo, 1995] is to consider the color distribution of an image region as a probability distribution. Since probability distributions can be characterized by statistical moments, the colors of an image region can also be described by these moments.

Given an image  $P$  with its pixels  $p_{ijk}$  and the mean  $m_k^1$  (see equation (5.4)) for the channel  $k$ , the other statistical moments  $m_k^h$  of order  $h$  can be computed as

$$m_k^h = \left( \frac{1}{W \times H} \sum_{i=1}^H \sum_{j=1}^W (p_{ijk} - m_k^1)^h \right)^{\frac{1}{h}} \quad (5.6)$$

Usually only the first 4 moments are considered, which are mean  $m_k^1$ , standard deviation  $m_k^2$ , skewness  $m_k^3$  and kurtosis  $m_k^4$ . These moments are extracted for each of the  $K$  color channels and concatenated into a feature vector  $\vec{f}$  represented as

$$\vec{f}_{CM} = (m_1^1, m_1^2, m_1^3, m_1^4, \dots, m_K^1, m_K^2, m_K^3, m_K^4) \quad (5.7)$$

with the dimensionality  $D = K \times H$ .

The parameters for the color moment extraction are the used color space, the number of channels  $K$ , and the number of computed moments  $H$ .

Color moments provide a robust and compact representation of the color distribution of an image. Since they do not consider any spatial information, they only describe the present colors without considering their spatial arrangement.

### Color histogram (CH)

Color histograms [Swain and Ballard, 1991] are one of the basic approaches for modeling the color distribution within an image.

To create a color histogram the color space is quantized into a set of predefined bins  $B$  for each color component. That can be either done *individually* for each channel which does not consider the correlation between them or *jointly* over all color channels. After partitioning the feature space, the probabilities for each bin are computed by counting the number of pixels that fall within each of the bins. Since each of the pixels belongs only to a single bin, this is often referred to as *crisp* histogram. Another way is to use *fuzzy* histograms [Siggelkow, 2002] where each pixel contributes to each bin with a certain amount that is computed based on a membership function. The nature and the dimensionality of the feature vector  $\vec{f}_{CH}$  differs for the different channel combination schemes. Individual histograms  $\vec{h}_k$  are concatenated in the following way

$$\vec{f}_{CH} = (\vec{h}_1, \vec{h}_2, \dots, \vec{h}_K) \quad (5.8)$$





**Figure 5.6:** Two images with similar color distributions but different spatial structure. While the left one contains only very small homogeneous regions, the right one consists of a few large homogeneous regions. While the color histogram is not able to distinguish between these two spatial distributions, the color coherence vector allows this by splitting the histogram into a coherent and an incoherent part.

which leads to a feature vector dimensionality  $D = K \times B$ . In the case of a joint histogram  $\vec{h}$ , the feature vector is simply defined as

$$\vec{f}_{CH} = \vec{h} \quad (5.9)$$

with the dimensionality  $D = B^K$ .

The extraction of a color histogram can be influenced by several parameters including the used color spaces, the number of bins per channel, the combination of the channels (either independently or jointly) and the assignment of the pixels to the bins (crisp or fuzzy). For the body description only histograms with joint channels and crisp memberships are considered.

Depending on the number of bins, color histograms are a quite precise way of approximating the color distribution within an image. Color channels can be either modeled independently or jointly by considering correlations between the different channels. Like the average color and the color moments, the color histogram does not consider any spatial information.

### Color coherence vector (CCV)

Since color histograms lack any spatial information regarding the color distribution, Pass and Zabih [1996] have proposed the color coherence vector (CCV) which incorporates spatial information into a color histogram. This is done by classifying pixels as either coherent or incoherent and creating an individual color histogram for each of the classes. This allows to distinguish between images with similar colors but different spatial distributions, as it can be seen in figure 5.6.

The extraction is similar to that of a color histogram. First of all the colors are quantized

into a predefined number of bins  $B$ . The next step classifies the pixels within each bin into coherent or incoherent pixels. Therefore, adjacent pixels corresponding to an individual bin  $b$  are grouped together using connected component labeling (see section 3.2.3) which leads to a set of regions  $r$ . Each pixel  $i$  is then classified as coherent or incoherent based on the area  $a_i$  of its region  $r_i$  as

$$c_i = \begin{cases} \text{coherent} & \text{if } a_i > \hat{a} \\ \text{incoherent} & \text{else} \end{cases} \quad (5.10)$$

for each of the bins  $b$  this leads to a pair of coherent pixels  $c_b$  and incoherent pixels  $\bar{c}_b$  which is called the *coherence pair*. Clearly, the total number of bins  $h_b = c_b + \bar{c}_b$  is equal to a bin in the corresponding color histogram. The final feature vector  $\vec{f}$  is simply the concatenation of the coherence pairs of all the bins  $b$ , defined as

$$\vec{f}_{CCV} = (c_1, \bar{c}_1, c_2, \dots, c_B, \bar{c}_B) \quad (5.11)$$

with the dimensionality  $D = 2 \times B$  for a joint color channel quantization with  $B$  bins.

The extraction of a color coherence vector can be influenced by the same parameters as the color histogram. Furthermore, the neighborhood for the connected component labeling (4 or 8) and the threshold for coherence classification can be set.

The color coherence vector is a simple extension to the color histogram that includes spatial coherency information by splitting the histogram into two parts. Thus, it provides a rather rough description of the spatial characteristics.

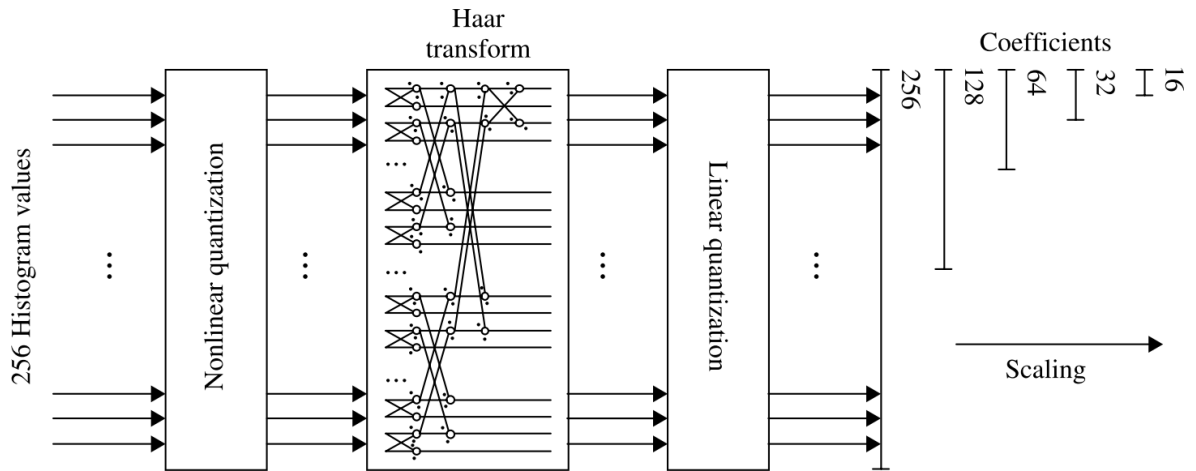
### Color spatiogram (CS)

A color spatiogram (CS) [Birchfield and Rangarajan, 2005] can be seen as a generalization of a color histogram which includes spatial information for each histogram bin in form of spatial mean and covariances. Thus it can be seen as a compromise between a color histogram that does not capture any spatial information and a color template which links color values and locations directly.

Again, the extraction is similar to that of a color histogram. First of all the colors are quantized into a predefined number of bins  $B$  which leads to the zero order spatiogram (histogram)  $n_b$ . Furthermore, the mean vector  $\vec{\mu}_b$  and the covariance matrix  $\Sigma_b$  of the corresponding pixel coordinates  $(x, y)$  are computed for each bin  $b$ . Instead of the covariance matrix  $\Sigma_b$  only the standard deviation vector  $\vec{\sigma}_b$  can be used, which treats the coordinates independently. Each bin  $b$  is then represented by a 5-tuple  $(n_b, \vec{\mu}_b, \vec{\sigma}_b)$ . These tuples are combined into a feature vector  $\vec{f}$  defined as

$$\vec{f}_{CS} = (n_1, n_2, \dots, n_B, \vec{\mu}_1, \vec{\mu}_2, \dots, \vec{\mu}_B, \vec{\sigma}_1, \vec{\sigma}_2, \dots, \vec{\sigma}_B) \quad (5.12)$$

with the dimensionality  $D = 5 \times B$ .



**Figure 5.7:** Extraction of the MPEG-7 scalable color descriptor (SCD) [Manjunath et al., 2002]. A 256 bin color histogram is nonlinearly quantized and the scalability is achieved through Haar encoding.

The extraction of a color coherence vector can be influenced by the same parameters as the color histogram. Furthermore, different types (full, diagonal, equal) of the covariance matrix can be used.

The color spatiogram combines the description of the color distribution with the spatial distribution by modeling the spatial characteristics of each histogram bin with a multivariate Gaussian distribution. Hence, it does not only provide information regarding the coherency of each quantized color but also about the location.

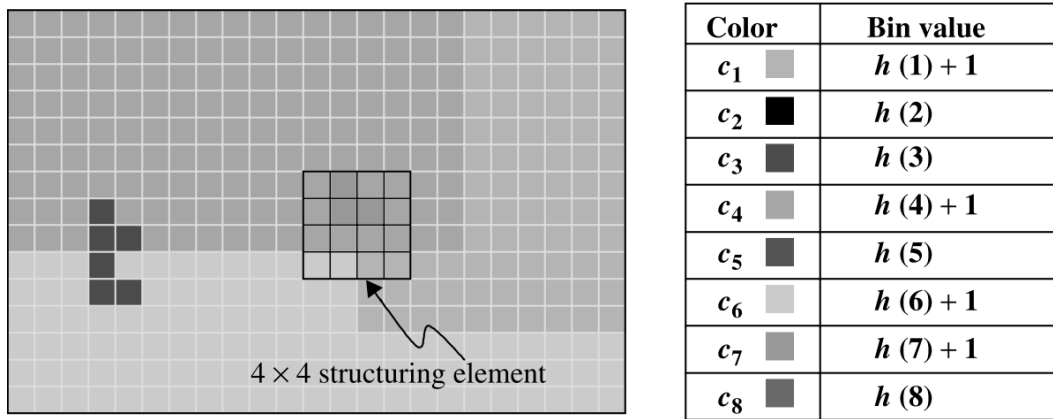
### Scalable color descriptor (SCD)

The scalable color descriptor [Manjunath et al., 2001] defined in the MPEG-7 standard combines a color histogram with a color space and a color quantization descriptor.

For the extraction, the HSV color space is uniformly quantized into 256 bins with (16/4/4) levels for the individual components. The histogram values are truncated into an 11 bit integer representation which is further compressed into a non linear 4 bit representation. In order to make the representation scalable the 256 bin representation is encoded using a Haar transform, where each subset of Haar coefficients corresponds to a histogram of (128,64,32,16) bins (see figure 5.7). Furthermore, also the magnitude of the different components can be scaled by considering different number of bits reaching from 1–8 bits. Depending on the number of coefficients  $C$  this results in a feature vector  $\vec{f}_{SCD}$  with the dimensionality  $D = C$ .

While the extraction itself does not require any parameters, the representation allows to choose between different number of Haar coefficients  $C$ .

The scalable color descriptor is merely a color histogram in the HSV color space where the number of bins is scalable due to the Haar transform based encoding. Like the color histogram it does not contain any spatial information.



**Figure 5.8:** Extraction of the MPEG-7 color structure descriptor (CSD) [Manjunath et al., 2002]. The color structure is a generalization of the color histogram that applies a structuring element to distinguish between different spatial color distributions.

### Color structure descriptor (CSD)

The MPEG-7 color structure descriptor (CSD) [Manjunath et al., 2001] aims at describing local structure within an image using a structuring element. In contrast to a color histogram, it represents both the color distribution and the local spatial structure of the color within an image. This is achieved by counting the number of times a particular color is present within the structuring element while scanning the image.

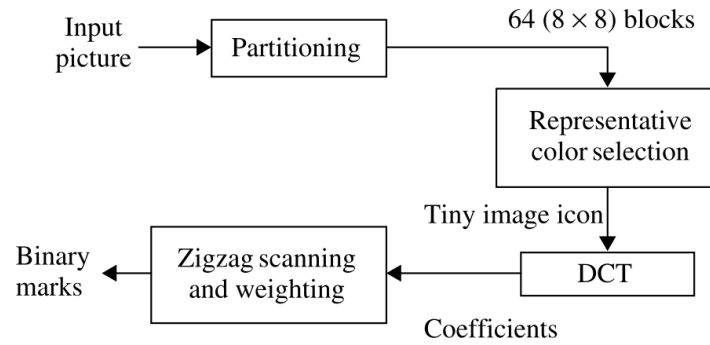
Colors are represented in the HMMD color space, which is non-uniformly quantized into different number of bins (184,120,64,32). In order to compute the CSD an  $s \times s$  structuring element scans the image in the way that it visits every position in the pixel grid and always lies entirely within the image. At each position a histogram is updated on the basis of which colors are present within the structuring element. Thereby the actual number of pixels of color does not matter, since only present and not present are distinguished which is illustrated in figure 5.8. It is interesting to note that the CSD may be viewed as a generalized color histogram since it is reduced to that when a  $1 \times 1$  structuring element is used. The resulting feature vector  $\vec{f}_{\text{CSD}}$  is built from the normalized histogram values with the dimensionality  $D = B$  varying with the number of bins.

Since the HMMD color space is chosen by default and the size of the structuring element is set to  $8 \times 8$  only one parameter is remaining. The number of histogram bins is set to 32.

The color structure descriptor is comparable to the color coherence vector but uses a structuring element to analyze the spatial coherency of colors within local neighborhoods.

### Color layout descriptor (CLD)

The color layout descriptor (CLD) [Manjunath et al., 2001, 2002] is a very compact and efficient representation of the spatial color distribution of an image. It is especially useful for fast sketch based retrieval, image filtering and visualization.



**Figure 5.9:** Extraction of the MPEG-7 color layout descriptor (CLD) [Manjunath et al., 2002]. The color layout descriptor can be seen as a color template within the frequency domain.

The extraction, illustrated in figure 5.9, starts by converting the image into YCbCr color space and subsampling the image into a 8x8 grid. Thereby, each block is represented with the average color of its pixels. The discrete cosine transform (DCT) is applied individually to each of the color channels, which leads to a series of DCT coefficients  $\vec{c}$ . After applying a zig-zag scan a few low frequency coefficients from each channel are selected. The final feature vector  $\vec{f}$  is composed of the coefficients in the following way

$$\vec{f}_{\text{CLD}} = (\vec{c}_Y, \vec{c}_{\text{Cb}}, \vec{c}_{\text{Cr}}) \quad (5.13)$$

The color layout descriptor can be seen as a color template in the YCbCr color space that directly combines color and spatial information. The transformation into the frequency domain allows to discard higher order frequencies which are usually not important.

### Intensity moments (IM)

Intensity moments are an adaptation of the color moments to grayscale images. The extraction is similar to that of a the color moment with a single color channel. While the first order moment (mean)  $m^1$  is computed according to equation (5.4), the higher order moments  $m^h$  are computed as defined in equation (5.6). The resulting feature vector with the dimensionality  $D = 4$  is defined as

$$\vec{f}_{\text{IM}} = (m^1, m^2, m^3, m^4) \quad (5.14)$$

Intensity moments are comparable to color moments extracted from grayscale images. While they describe the brightness and the contrast, they provide no information regarding the directionality or coarseness of a texture.

### Grayscale cooccurrence matrix (GCM)

In contrast to the first order statistics of the intensity histogram, Haralick [1979] proposed to extract second order statistics based on so called grayscale cooccurrence matrices (GCM).

A GCM is defined as a matrix of probabilities  $p_v(i, j)$  at which two quantized gray level

values  $i$  and  $j$  are separated by a vector  $v(\phi, d)$  defined by the angle  $\phi$  and the distance  $d$  as it is shown in figure 5.10. For each of the  $v$  displacement vectors another matrix is built which allows to capture different texture characteristics. From each GCM various features can be computed, that can be categorized into texture characteristics, statistics, information theoretic and correlation measures [Haralick, 1979; Gotlieb and Kreyszig, 1990]. The most commonly used features are

$$f_v^1 = \sum_{i,j} p_v(i,j)^2 \quad (\text{Energy}) \quad (5.15)$$

$$f_v^2 = \sum_{i,j} p_v(i,j) \log p_v(i,j) \quad (\text{Entropy}) \quad (5.16)$$

$$f_v^3 = \sum_{i,j} (i-j)^2 p_v(i,j) \quad (\text{Contrast}) \quad (5.17)$$

$$f_v^4 = \sum_{i,j} \frac{p_v(i,j)}{1 + |i-j|} \quad (\text{Homogeneity}) \quad (5.18)$$

which are extracted for each of the GCMs defined by a displacement vector  $v$ . All computed measures are combined into a feature vector in the following way

$$\vec{f}_{\text{GCM}} = (f_1^1, f_2^1, \dots, f_V^1, \dots, f_1^4, f_2^4, \dots, f_V^4) \quad (5.19)$$

with the dimensionality  $D = 4 \times V$ .

The features derived from the cooccurrence matrix are influenced by certain parameters including the number of quantization levels  $B$ , the number of displacement vectors  $V$  defined by varying the angle  $\phi$  and the distances  $d$ .

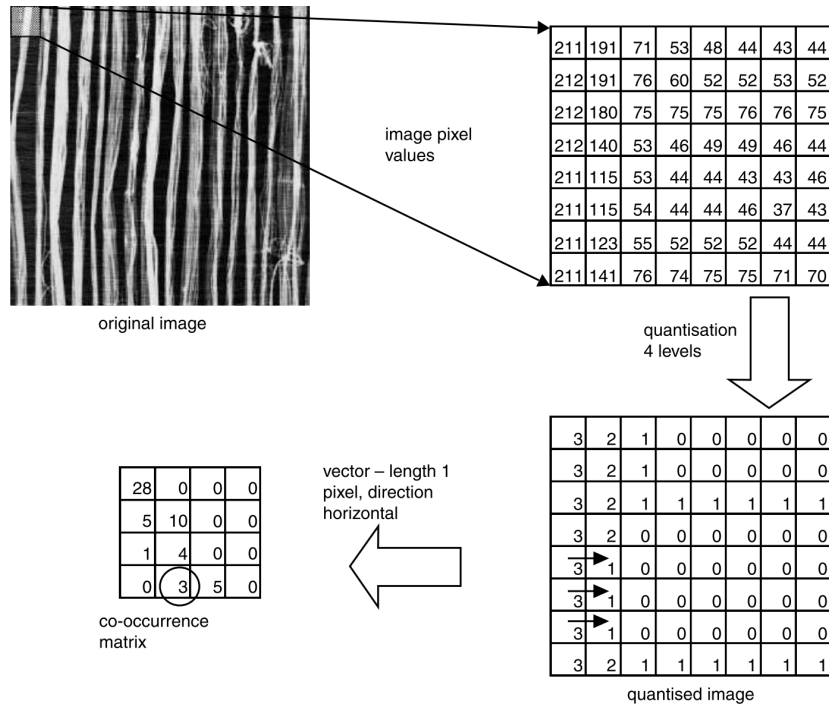
GCM derived features describe the second order statistics between different intensity values.

### Tamura features (TF)

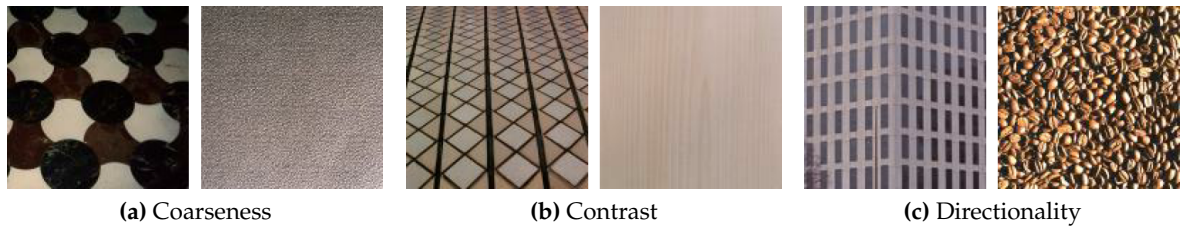
Tamura et al. [1978] proposed texture features that correspond to the human visual perception. They defined six textural features (coarseness, contrast, directionality, line likeness, regularity, and roughness) and compared them to psychological measurements of humans. They found that especially the first three correlate strongly with the human perception.

*Coarseness* has a direct relationship to the scale and repetition of a texture and was seen as the most fundamental texture feature by Tamura et al. [1978]. If an image contains textures at different scales, coarseness aims to describe the largest macro texture, even if smaller micro textures exist. The extraction starts by averaging the image at every point  $(x, y)$  over neighborhoods of different sizes  $2^k \times 2^k$  which is written as

$$A_k(x, y) = \sum_{i=x-2^{k-1}}^{x+2^{k-1}-1} \sum_{j=y-2^{k-1}}^{y+2^{k-1}-1} p(i, j) / 2^{2k} \quad (5.20)$$



**Figure 5.10:** Extraction of the grey level cooccurrence matrix (GCM). Given a quantized image a grayscale cooccurrence matrix measures the probability that two intensity levels within an image are separated by a certain angle and distance.



**Figure 5.11:** Illustration of the different Tamura features for some samples. The perceptually inspired measures describe certain aspects of a texture. The coarseness measures the scale of the largest texture within an image. The contrast measures the intensity difference within an image. The directionality allows to distinguish between directed and undirected textures.

Then at each point the absolute horizontal and vertical difference  $E_k^h$  and  $E_k^v$  between averages of non overlapping neighborhoods on opposite sides of the point  $(x, y)$  are computed as

$$E_k^h(x, y) = |A_k(x + 2^{k-1}, y) - A_k(x - 2^{k-1}, y)| \quad (5.21)$$

$$E_k^v(x, y) = |A_k(x, y + 2^{k-1}) - A_k(x, y - 2^{k-1})| \quad (5.22)$$

For each point one then picks the largest from these two differences and determines the best

fitting scale  $S(x,y)$  defined as

$$S(x,y) = \operatorname{argmax}_k \max_{d=\{h,v\}} E_k^d(x,y) \quad (5.23)$$

Finally the coarseness measure is the average over  $2^{S(x,y)}$  computed as

$$f_1 = \frac{1}{W \cdot H} \sum_{x=1}^W \sum_{y=1}^H 2^{S(x,y)} \quad (5.24)$$

*Contrast* aims to capture the dynamic range of the gray levels within an image and the polarisation of this distribution regarding black or white. The first is measured using the standard deviation  $\sigma$  of the gray levels and the second is described by the kurtosis  $\alpha_4$ . Both are combined into the contrast measure defined as

$$f_2 = \frac{\sigma}{\alpha_4^z} \quad (5.25)$$

with the factor  $z$  experimentally determined to be  $1/4$  [Howarth and Rüger, 2004].

*Directionality* measures not the orientation itself but the presence of orientation in the texture. That is, two textures differing only in the angle are considered to have the same directionality [Deselaers, 2003]. To extract the directionality the horizontal and vertical derivatives  $\Delta_h$  and  $\Delta_v$  are computed by convolution of the input image with the two kernels

$$k_h = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad k_v = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \quad (5.26)$$

from which the magnitude  $|\Delta|$  and the angle  $\phi$  are computed as

$$|\Delta| = (|\Delta_h| + |\Delta_v|)/2 \quad (5.27)$$

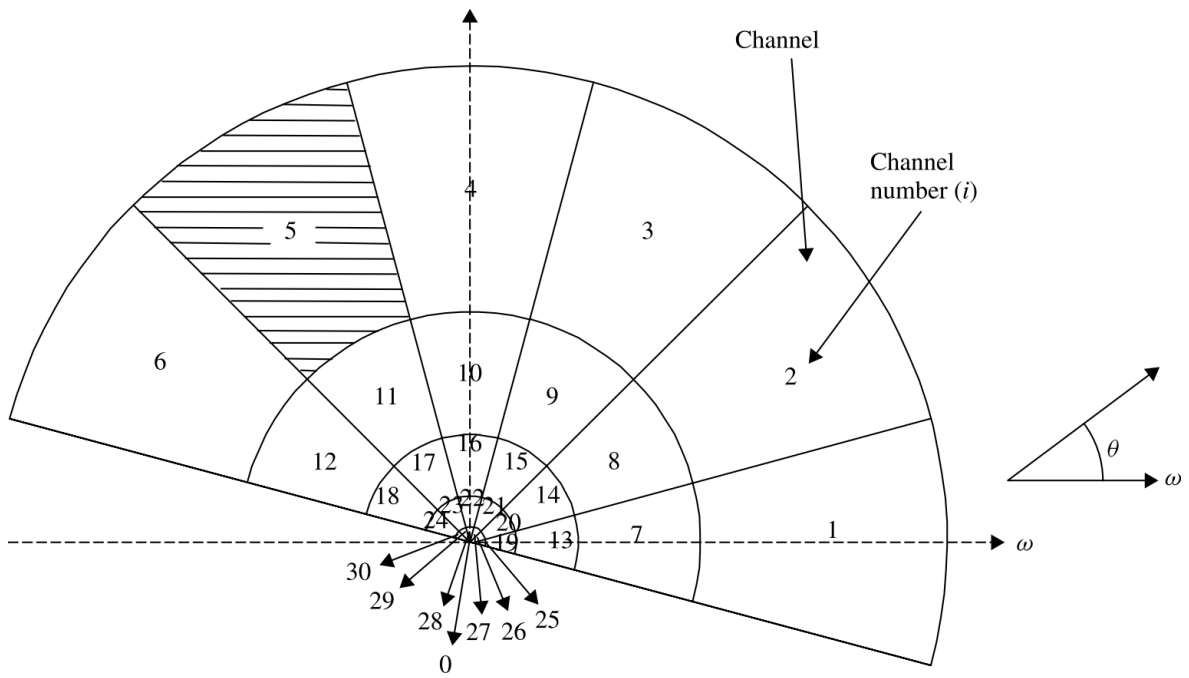
$$\phi = \arctan(\Delta_v / \Delta_h) + \pi/2 \quad (5.28)$$

The distribution of edge angles  $\phi$  is described with a histogram  $h_D$  by considering only points with edge magnitudes  $|\Delta|$  greater than a threshold and quantizing the corresponding angles. The directionality can be computed from this histogram by measuring the sharpness of its peaks  $p$  as

$$f_3 = 1 - r \cdot n_p \sum_p \sum_{\phi \in w_p} (\phi - \phi_p)^2 \cdot h_D(\phi) \quad (5.29)$$

with the number of peaks  $n_p$ , the position of the  $p$ th peak  $\phi_p$  and its range  $w_p$ , and the normalization factor  $r$ .





**Figure 5.12:** Illustration of the different Gabor filters used for the extraction of the MPEG-7 homogeneous texture descriptor (HTD) [Manjunath et al., 2002]. It distinguishes between 6 linearly quantized angles and 5 logarithmically quantized scales.

These features are combined into a feature vector of dimensionality  $D = 3$  defined as

$$\vec{f}_{\text{TF}} = (f_1, f_2, f_3) \quad (5.30)$$

Several parameters influence the extraction of the Tamura features. The coarseness is influenced by the neighborhood sizes which are defined by the factor  $k$ . The contrast is influenced by the exponent  $z$  which is set to  $1/4$ . The directionality is influenced by 3 parameters including the number of histogram bins  $B$ , the threshold  $|\Delta|$  and the normalization factor  $r$ .

The Tamura features are a small set of features that correspond very well to the human visual perception. Each of these features describes another aspect of a texture including coarseness, contrast and directionality. Nevertheless they do not consider information regarding the spatial layout of multiple textures.

### Homogeneous texture descriptor (HTD)

The homogeneous texture (HTD) [Manjunath et al., 2001, 2002] describes the texture of a region using the mean and deviation of the energy from a set of frequency channels. Therefore, the 2D frequency plane is partitioned into 30 channels as shown in figure 5.12. While the frequency partitioning is uniform along the angular direction it is non uniform along the radial direction.

The first two features of the homogeneous texture descriptor are the mean  $\mu$  and the stan-

standard deviation  $\sigma$  in the image space which correspond to the first two intensity moments. The other features are computed in the frequency space. Therefore, each of the individual channels shown in figure 5.12 is modeled as a Gabor function [Manjunath and Ma, 1996]. Based on the frequency layout and the different Gabor functions, the energy mean  $e_{sr}$  of the channel with the angular index  $r \in \{0, 1, \dots, 5\}$  and the radial index  $s \in \{0, 1, \dots, 4\}$  is defined as the log scaled sum of the squared Gabor filter responses of the image defined as

$$e_{sr} = \log(1 + p_{sr}) \quad (5.31)$$

$$p_{sr} = \sum_{\omega} \sum_{\phi} (G_{sr}(\omega, \phi) |\omega| P(\omega, \phi))^2 \quad (5.32)$$

with the Gabor filter function  $G_{sr}$  and the Fourier transform of the image in the polar frequency domain  $P$ . In a similar way the energy deviation  $d_{sr}$  is computed for each channel as

$$d_{sr} = \log(1 + q_{sr}) \quad (5.33)$$

$$q_{sr} = \sqrt{\sum_{\omega} \sum_{\phi} \left( (G_{sr}(\omega, \phi) |\omega| P(\omega, \phi))^2 - p_{sr} \right)^2} \quad (5.34)$$

This computation can be efficiently performed using the Radon transform [Jain, 1989], which is defined as the integral along a line with the angle  $\phi$  and the distance  $\omega$  to the origin. Given that, the 1D Fourier transform of the image at angle  $\phi$  is equal to the slice at angle  $\phi$  in the 2D Fourier transform of the image, the Radon transform allows to reduce the computational complexity considerably. The feature vector has a dimensionality  $D = 2 + 2 \times 6 \times 5 = 62$  and is defined as

$$\vec{f}_{\text{HTD}} = (\mu, \sigma, e_{0,0}, e_{0,1}, \dots, d_{0,0}, d_{0,1}, \dots) \quad (5.35)$$

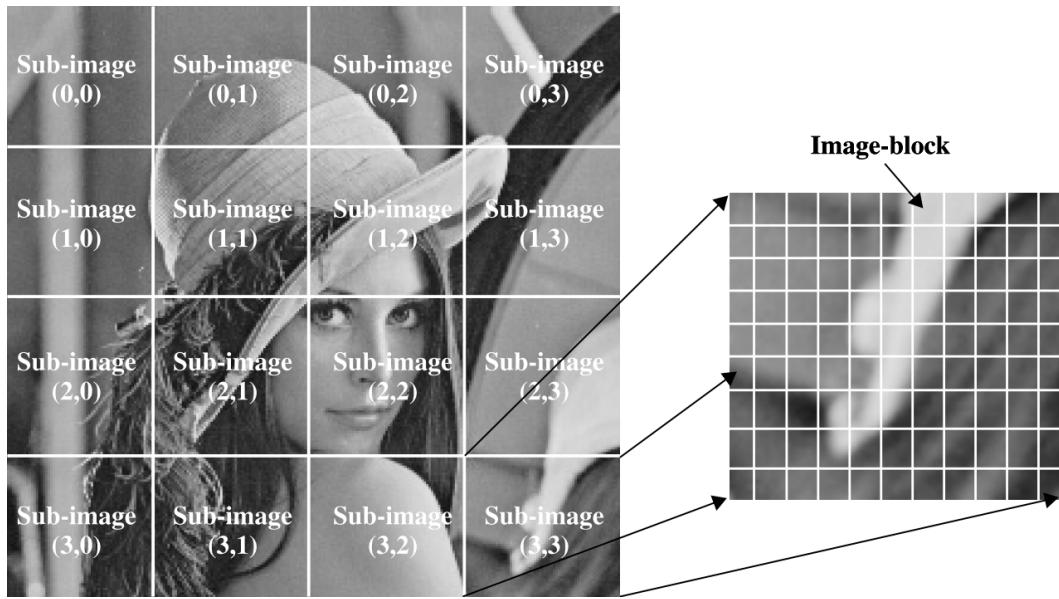
Since the frequency layout (angles and distances) is already predefined there are no free parameters remaining.

The HTD describes the direction and the scale of homogeneous textures, which allows to distinguish between textures of different orientation or coarseness. In the case of inhomogeneous textures the results are difficult to interpret.

### Edge histogram descriptor (EHD)

The edge histogram descriptor (EHD) [Won and Park, 1997; Park et al., 2000] describes the local edge distribution by building a histogram over different edge types (vertical, horizontal,  $45^\circ$ ,  $135^\circ$  and non-directional) individually for each of the blocks within a predefined grid.

The extraction of the edge histogram descriptor is straightforward. First of all, the image is subdivided into subimages  $i$  for which individual histograms are extracted. Each of these subimages is further subdivided into a predefined number of blocks  $j$  which are



**Figure 5.13:** Extraction of the MPEG-7 edge histogram descriptor (EHD) [Manjunath et al., 2002]. The images is subdivided into 16 blocks and for each block a histogram of different edge types is extracted.

classified into one of the 5 edge types and combined into a histogram  $h_i(e)$  of edge types  $e \in v, h, u, d, n$ . For the classification each block is subsampled into a  $2 \times 2$  macro block and several edge detectors corresponding to the different edge types are applied. Given the filter responses of the different edge types for a single block, the strongest one is considered to be dominating edge. If the corresponding strength exceeds a predefined threshold this edge type is considered within the histogram. Otherwise, if the block contains no edges it does not have an influence on the histogram. The individual local edge histograms are concatenated into the feature vector defined as

$$\vec{f}_{\text{EHD}} = (h_1, h_2, \dots, h_{16}) \quad (5.36)$$

with the dimensionality  $D = 16 \times 5 = 80$ .

The extraction of the edge histogram descriptor can be influenced by several parameters. The first one is the number of subimages which is predefined to  $4 \times 4 = 16$ . The number of blocks for the edge classification is set to 1100 which is equal to 33 blocks per dimension. The last parameter is the threshold for classifying a block as edge or non edge which is set to.

The EHD describes textures based on the local analysis of edges and gradients. It allows to distinguish between different spatial layouts of multiple textures.

### 5.2.3 Recognition

As already discussed in chapter 4 recognition in this work may refer to several related tasks including matching, clustering and classification.

*Body matching* refers to measuring the dissimilarity between two body descriptions. For a feature extracted from a certain body part the distance between two descriptions can be computed by several distance measures described in section 3.3.1. Depending on the type of feature certain distance metrics are more suitable than others. Table 5.2 provides an overview of the recommended distance measure for each feature.

*Body clustering* describes the process of grouping similar body descriptions together. This can be used to automatically determine the number of distinct costumes from a set of images or videos that may correspond to certain teams or groups in sports or surveillance. Any of the clustering methods described in section 3.3.4 is suitable for this task and the optimal choice depends largely on the task and the data. Clustering is used to create a human visual thesaurus for visual person search in chapter 9.

*Body classification* predicts a category for a body description based on trained category models and a classification rule. This can be used for several tasks, including occlusion handling, camera handover and determining the group (team) of persons. Any of the classification approaches described in section 3.3.5 can be applied.

### 5.2.4 Fusion

Information fusion maybe used to improve the performance of the different body recognition tasks by combining complementary information available within the body descriptions.

This includes mainly two different types of information: a set of different representations (whole, head, upper, lower) and a set of visual low-level features (color and texture).

From this set of information channels (parts, features) a subset can be selected based on different criteria. The first way is to combine all or a predefined set of representations, which may improve the results. The second way is to apply feature selection techniques (BIS, SFS) described in section 3.3.2 to select an optimal subset of the available visual features based on a performance criteria. The third way is to select the reliable representation based on some a priori knowledge provided by the body detection or the application itself.

Given the selection the final question is how to fuse the information channels. Therefore any fusion method described in section 5.2.4 is applicable. Nevertheless, only score level fusion in form of score combination rules (min, max, sum, product) will be considered here.

## 5.3 Experiments

The goal of the following experiments is to assess the performance of the body recognition module in terms of the different representations and features. The first step is to compare the different methods for the component extraction with each other regarding their accuracy



**Figure 5.14:** Samples of the Neckermann Database with a large variety of costumes with different colors, textures and shapes.

and reliability for a large variety of costumes. The second step is to evaluate the different representations and features for the recognition tasks in order to select the most appropriate channels for this task. Furthermore, the fusion of both parts and features will be explored.

### 5.3.1 Dataset

#### Neckermann database

The Neckermann Database (see section A.2.1 for more details) contains 42 images of humans with as many different costumes from an online fashion shop. Since it contains only one sample per costume it is not suitable for the evaluation of the body recognition performance. On the other hand, the large variety of costumes with different colors, textures and shapes (see figure 5.14 for some samples) allows to comprehensively evaluate the different body representation approaches.

#### Free Character Database

The Free Character Database (see section A.2.2 for more details) contains 216 images of humans with 54 costumes in 4 views (frontal, left, back, right) with a uniform background. Since each of the 4 views can be treated as a sample it may be used for evaluating body recognition approaches. It also contains a large variety of different costumes reaching from sports over casual to business (see figure 5.15 for some samples).

### 5.3.2 Evaluation

#### Representation

The extraction of the individual body parts can be evaluated as a recognition problem (see section B.4 for more details), where each pixel is classified as head, upper or lower body. Given a set of pixel-wise ground truth and predicted region maps, shown in figure 5.16, a confusion matrix is derived from which the accuracy (recognition rate) is computed.



**Figure 5.15:** Samples of the Free Character Database with a large variety of costumes reaching from sports over casual to business.

## Recognition

The body recognition has been evaluated as recognition problem (see section B.4 for more details). Therefore the cumulative match characteristic (CMC) curve is computed based on the pair wise match scores between a training and a testing set. The top rank score of the CMC curve is equal to the recognition rate of a 1 nearest neighbor classifier (see section 3.3.5) trained on the training set.

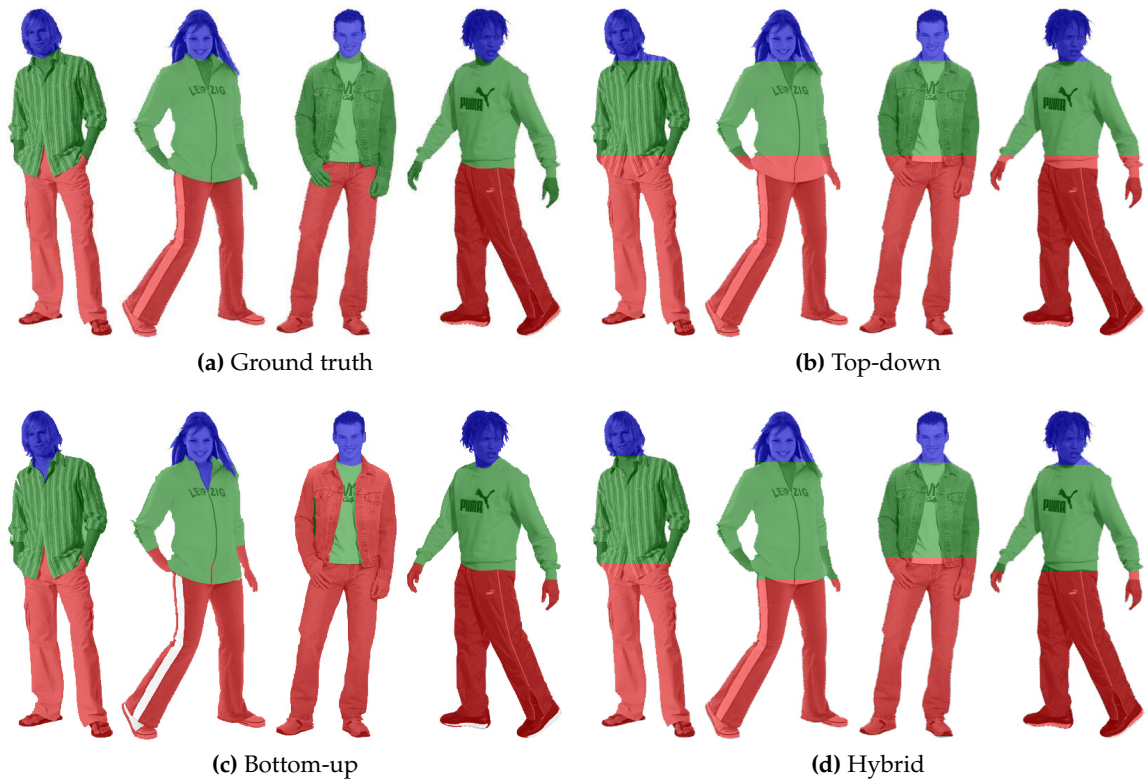
### 5.3.3 Results

#### Representation

The goal of this set of experiments is to assess the performance of the different segmentation approaches for the component based representation (see section 5.2.1). All the experiments are based on the Neckermann Database since it contains a large variety of costumes with different colors, textures and shapes. The performance has been assessed subjectively through visual samples and objectively by considering the extraction as a segmentation problem.

Figure 5.16 provides a subjective comparison between the different approaches by showing representative segmentation results along with the manually defined ground truth (figure 5.16a) used for the objective evaluation. The *top-down* approach (figure 5.16b) is based on horizontal boundaries at fixed vertical positions. Thus the segmentation accuracy depends largely on the size of the individual components. The *bottom-up* approach (figure 5.16c) is not based on any predefined boundary shape and thus more precise for a large number of cases. Nevertheless it is not very robust in cases where two body parts have a similar color or texture. In these cases two components may be partially merged which results in a very poor segmentation accuracy. The *hybrid* approach (figure 5.16d) is also based on horizontal boundaries but the vertical position is adjusted based on the appearance of the body parts. Thus it achieves a precision that is between the bottom-up and the top-down approach. On the other hand, it inherits the reliability of the top-down approach.

The subjective results are supported by the objective evaluation based on the comparison of the ground truth and the predicted segmentation maps. Figure 5.17 provides a comparison of the different approaches based on a boxplot of the accuracy. The highest median



**Figure 5.16:** Visual samples of the different body modeling approaches. While the bottom-up approach provides the best segmentation accuracy it is not very reliable in the presence of visually similar clothes. The best tradeoff between accuracy and reliability is achieved by the hybrid approach.

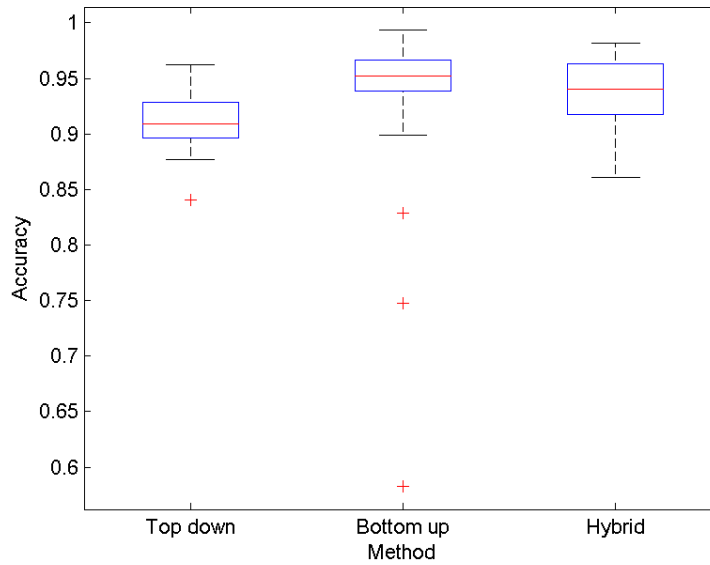
accuracy is achieved by the bottom-up approach (95%) closely followed by the hybrid approach (94%) and the top-down approach (91%). On the other hand, the number of outliers for the bottom-up approach is much higher than for the two other approaches which shows the decreased reliability of this approach. Based on these results the hybrid approach was chosen since it provides the best tradeoff between accuracy and reliability.

## Recognition

The goal of this set of experiments is to assess the performance of the different representations and the different visual descriptors for body recognition. Furthermore, the fusion of several representations and descriptors is explored. All the experiments are based on the Free Character Database since it provides multiple samples for each costume in various views.

The goal of the first experiment is to assess the performance of the individual descriptors and representations for the body matching task. Table 5.3 provides an overview of the performance across the different descriptors grouped by type (color, texture) and standard (non, MPEG-7) over the different representations. It is important to note that the extraction





**Figure 5.17:** Objective evaluation of the different extraction approaches for the component based body representation. The objective evaluation confirms the subjective results. While the bottom up approach achieves the highest accuracy it is not very reliable which is documented by the outliers.

of the MPEG-7 descriptors is only defined for rectangular images or regions and the extension to arbitrarily shaped regions is not straightforward. The current system is based on the aceToolbox developed within the aceMedia<sup>1</sup> project, which extends the MPEG-7 XM<sup>2</sup> software to arbitrarily shaped regions described by a binary segmentation mask. Unfortunately, the extraction of the texture descriptors does not work for all possible region shapes and sizes, which prevents a reliable evaluation. These cases are excluded. In general, the performance of the color features is much higher than that of the texture features. Considering the best features of each group for the upper body, the color coherence vector and the edge histogram descriptor achieve a recognition rate of 0.975 and 0.198, respectively. Although the recognition rate of the texture descriptors is quite low it is well above a random guess of  $1/54 = 0.018$ . The performance of the standard and non standard features is comparable. For some body parts the non standard descriptors even outperform the standard ones, e.g. for the whole body the CSD and the CCV achieve a recognition rate of 0.753 and 0.957, respectively. A deeper analysis of the color features shows that those considering spatial information (CCV, CS, CLD) achieve a higher performance for the whole body than for the individual body parts while it is the opposite for those without spatial information (AC, CM). This is caused by the fact, that individual body parts usually have a more uniform color than the whole body. The performance varies quite considerably over the different representations which can be ranked in the following way: whole body, upper body, lower body and head. While the performance of the whole, upper and lower body is quite com-

<sup>1</sup><http://www.acemedia.org/>

<sup>2</sup><http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/mpeg7.html>



Feature	Whole	Head	Upper	Lower
Average color	0.537	0.043	0.617	0.420
Color moments	0.512	0.043	0.272	0.228
Color histogram	0.938	0.389	0.833	0.821
Color coherence vector	0.957	0.352	0.827	0.802
Color spatiogram	0.691	0.333	0.605	0.617
Intensity moments	0.296	0.031	0.179	0.130
Coocurrence matrix	0.117	0.043	0.123	0.062
Tamura features	0.043	0.019	0.080	0.062
Color layout	0.735	0.117	0.574	0.586
Color structure	0.753	0.216	0.870	0.710
Scalable color	0.173	0.086	0.302	0.099
Edge histogram	–	0.105	0.198	0.210
Homogeneous texture	–	–	–	–

**Table 5.3:** Body matching performance of the individual features across the different body parts. The highlighted values correspond to the best performance of the individual feature.

parable it drops considerably for the head due to the lower discriminability. The difference between the upper and the lower body can be explained by the different amount of color and texture variations among these two body parts.

The goal of the second experiment is to explore the fusion of complementary features as described in section 5.2.4 and its influence on the body recognition performance. Due to the extraction problems mentioned above only the non standard descriptors have been considered. Since the number of available descriptor and thus the number of possible combinations is quit large different feature selection have been used to determine the set of descriptors that are fused. Table 5.4 provides the cumulative top rank score for the different feature selection (All, BIF, SFS) and score fusion (min, max, prod, sum) methods across the individual body parts. By comparing it to the best features in table 5.2 several observations can be made. While the performance actually decreases when all features are fused (e.g. from 0.957 to 0.932 for the whole body) a performance gain is achieved with the different feature selection methods (e.g. from 0.957 to 0.963 and 0.975 for BIF and SFS respectively). The performance gain is larger for SFS than for the BIS since it considers also the complementarity of the features. The ranking of the different body parts remains the same as for the single features. Depending on the body part the best performance is achieved either by the sum (W, U, L) or the max rule (H).

The goal of the third experiment is to investigate the fusion of the different representations as described in section 5.2.4 and its influence on the body matching performance. Table 5.5 offers a comparison of different part sets and fusion methods across the standard visual descriptors. By comparing it to the performance if the features across the individual parts in table 5.2 several things can be observed. The influence of the part fusion on the per-

Selection	Fusion	Whole	Head	Upper	Lower
All	min	0.370	0.037	0.253	0.222
All	max	0.796	0.383	0.747	0.802
All	prod	0.840	0.130	0.735	0.636
All	sum	0.932	0.309	0.784	0.827
BIF	min	0.957	0.389	0.833	0.821
BIF	max	0.957	0.420	0.833	0.846
BIF	prod	0.957	0.395	0.833	0.852
BIF	sum	0.963	0.414	0.864	0.901
SFS	min	0.957	0.389	0.833	0.821
SFS	max	0.957	0.426	0.833	0.846
SFS	prod	0.957	0.395	0.864	0.870
SFS	sum	0.975	0.414	0.864	0.901

**Table 5.4:** Body matching performance for different feature fusion methods across the different body parts. The highlighted values correspond to the best performance of each body part.

Parts	Fusion	AC	CM	CH	CCV	CS	IM	CM	TF
WHUL	min	0.568	0.438	0.821	0.778	0.710	0.148	0.086	0.056
WHUL	max	0.488	0.154	0.926	0.932	0.586	0.148	0.142	0.111
WHUL	prod	0.852	0.525	0.951	0.963	0.821	0.148	0.198	0.111
WHUL	sum	0.790	0.321	0.963	0.975	0.833	0.148	0.222	0.117
HUL	min	0.488	0.290	0.728	0.716	0.654	0.148	0.080	0.049
HUL	max	0.475	0.117	0.926	0.932	0.580	0.148	0.142	0.117
HUL	prod	0.796	0.358	0.951	0.957	0.796	0.148	0.173	0.086
HUL	sum	0.728	0.185	0.963	0.969	0.809	0.148	0.167	0.111
UL	min	0.568	0.321	0.784	0.772	0.691	0.148	0.099	0.049
UL	max	0.759	0.259	0.957	0.963	0.673	0.148	0.130	0.123
UL	prod	0.802	0.377	0.944	0.944	0.747	0.148	0.130	0.080
UL	sum	0.796	0.315	0.957	0.957	0.784	0.148	0.117	0.111

**Table 5.5:** Body matching performance for different part fusion methods across the different features. The highlighted values correspond to the best performance of each feature.

formance varies considerably between the different features. While a significant gain can be achieved for some features, e.g. for the AC the cumulative match score increases from 0.617 (U) over 0.796 (HUL) and 0.802 (UL) to 0.852 (WHUL) by 0.235, the gain is much smaller for other features, e.g. for the CCV the CMS changes from 0.957 (W) over 0.957 (UL) and 0.963 (HUL) to 0.975 (WHUL) by 0.018. The ranking of the features is comparable to the individual body parts although the quantitative difference between the complex and simple features decreases. Depending on the feature and the parts either the sum or the product are the optimal fusion methods.

## 5.4 Conclusion

### 5.4.1 Summary

An original method for appearance based body recognition has been developed, that describes the human body at different levels (holistic, components). Given a binary segmentation mask that describes the whole body, different methods (top-down, bottom-up, hybrid) for segmenting it into the individual components (head, upper body, lower body) have been developed. For the visual description of these body parts a large number of color and texture features with different characteristics and complexity have been considered. The use of the resulting hierarchical body model has been discussed for typical recognition tasks, such as matching, clustering and classification. In order to increase the robustness of the body recognition the fusion of features and parts has been explored. While an optimal subset of features is found using feature selection techniques, predefined sets of parts are fused.

Within an extensive set of experiments the performance of both the component extraction and the body recognition has been evaluated on representative databases. For the body representation the hybrid method provides the best tradeoff between segmentation accuracy and reliability. The body matching experiments have shown that very good performance can be achieved by either complex features (e.g. CCV) applied to the whole body or simple features (e.g. AC) applied to the individual body parts and fusing them. That corresponds quite well to the human visual perception, which has been shown to process information both holistically as well as part based.

### 5.4.2 Future work

Although the developed body recognition module allows to distinguish persons quite reliably based on the body appearance, it may be extended in several directions.

One of the limitations so far is that the body representation can only be extracted for persons in a standing *pose*, which is by far the most usual one. Nevertheless, it is possible to include a more sophisticated motion and shape based body model to extract the individual body parts. On the other hand, the application of these models is typically very complex and may not influence the overall body recognition performance that much.

Another challenge that has not been explored so far, is the influence of *illumination* changes onto the body description and thus on the body recognition performance. Although several features that have been considered are insensitive to moderate illumination changes illumination compensation methods (color constancy, contrast stretching) may be considered as a preprocessing step to decrease the influence of illumination variations.

So far only two levels of the body representation have been considered, the first being only the whole body and the second that consists of head, upper body and lower body. The parts in both representations may contain a varying amount of skin which influences the body description. One possible way to handle this would be to apply *skin* detection

techniques to detect skin colored pixels and consider them as a separate parts.

As already discussed the fusion step may consider a priori information for an appropriate feature or part selection. This can be supplied by the application itself or by the body detection module which may provide additional occlusion information. This may used in a similar way as in the face recognition module (see section 7). In any case the major idea is to improve the performance of the body recognition by selecting the most interesting or most reliable body parts.

## Chapter 6

# Face detection

### 6.1 Introduction

Face detection as a special case of object detection deals with finding and localizing an unknown number of faces within an still image or video frame. While humans can fulfill this task effortlessly, it is not an easy task for a machine since faces are dynamic objects with a high degree of variability in their appearance [Hjelmas and Low, 2001].

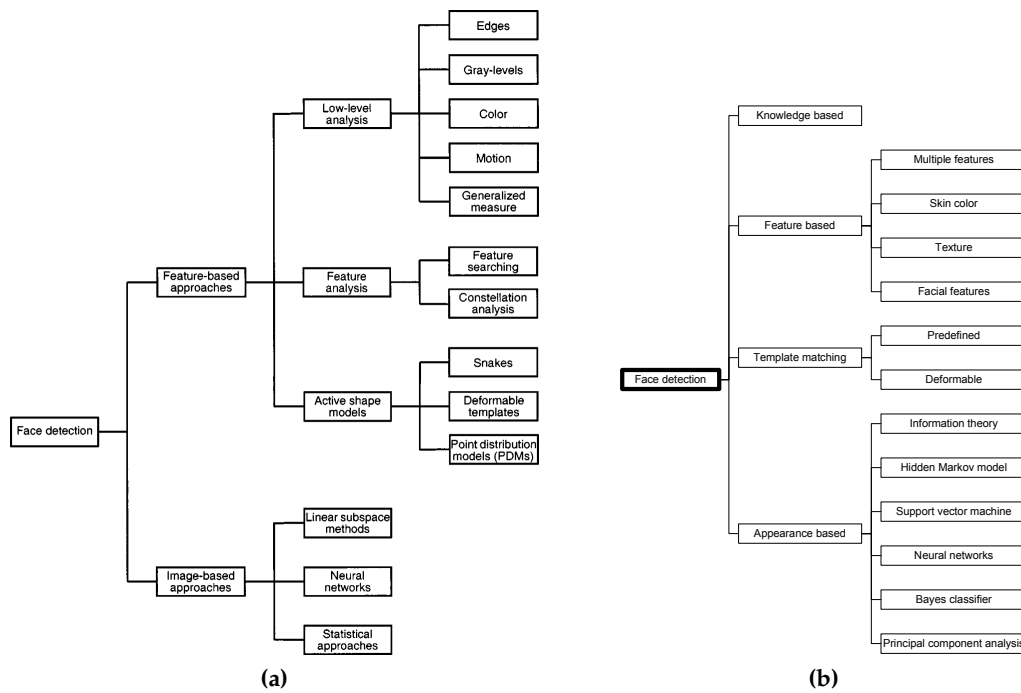
Reliable and precise face detection is even more important because it is the preliminary step in any face analysis system. Related tasks include facial feature detection, face tracking, face recognition and facial expression recognition. Furthermore it is used in a large variety of applications ranging from visual surveillance over human computer interaction to biometrics [Hjelmas and Low, 2001].

#### 6.1.1 Related work

Face detection has been a very active research field in the last decades and a large variety of approaches have been proposed. Several surveys [Hjelmas and Low, 2001; Yang et al., 2002] provide comprehensive overviews and of this research field.

Hjelmas and Low [2001] review approaches until 2001 and categorize them based on how a priori face knowledge is used:

**Feature based approaches:** These approaches incorporate face knowledge explicitly by applying a knowledge based analysis to previously extracted low level features. Methods based on *low level analysis* deal with the segmentation of faces based on pixel properties such as edges [Craw et al., 1987], intensity [Yang and Huang, 1994], color [Terrillon et al., 2000] and motion [Lee et al., 1996]. In order to resolve ambiguities *feature analysis* methods such as feature search [Jeng et al., 1998] and constellation analysis [Yow and Cipolla, 1997] exploit facial geometry to organize these low level features into more global concepts. Finally, *active shape models* including snakes [Kass et al., 1987], deformable templates [Yuille et al., 1992] and point distribution models [Lanitis



**Figure 6.1:** Taxonomies of face detection approaches according to (a) Hjelm and Low [2001] and (b) Yang et al. [2002].

et al., 1994] support face deformations due to facial expressions or varying pose.

**Image based approaches:** Taking advantage of the advances in pattern recognition, image based approaches address face detection as a classical recognition problem. Unlike the feature based approach, face knowledge is implicitly considered within the training stages. *Linear subspace* methods exploit the fact that faces lie in a small subspace within the overall image space by applying principal component analysis (PCA) [Turk and Pentland, 1991a], linear discriminant analysis (LDA) [Hotta et al., 1998], factor analysis (FA) [Yang et al., 2000], and self organizing maps (SOM) [Takacs and Wechsler, 1997]. *Neural networks* as a popular pattern recognition technique have been applied to face detection. Several architectures have been proposed including retinally connected neural networks [Rowley et al., 1998b], constrained generative models (CGM) [Feraud et al., 1997], and probabilistic decision based neural networks (PDBNN) [Lin et al., 1997]. Apart from these two groups, *statistical approaches* have been applied for face detection such as support vector machines [Osuna et al., 1997], Bayes decision theory [Schneiderman and Kanade, 2000], and information theory [Colmenarez and Huang, 1997].

In another survey Yang et al. [2002] reviewed face detection methods for single images and grouped them into 4 major categories (see figure 6.1):

**Knowledge based approaches:** They are based on rules which are derived from common

knowledge about human faces [Yang and Huang, 1994; Lv et al., 2000]. Typically these rules describe the appearance and the relationship between facial features. The major problem with this approach is the difficulty to translate human knowledge into a set of well defined rules, that provide a good tradeoff between generalization and specialization. Moreover it is difficult to extend these approaches to detect faces in different views.

**Feature based approaches:** can be seen as a bottom up approach that is based on invariant facial features that can be robustly detected in the presence of varying pose and illumination. Spatial relationships between these features are used to locate faces. One problem of these methods is that the feature localization may be largely influenced by illumination, noise and occlusions which makes the later grouping of the features unreliable. Methods are usually based on edges [Leung et al., 1995], texture [Dai and Nakano, 1996], skin color [McKenna et al., 1998] or combinations of them [Kjeldsen and Kender, 1996].

**Template matching approaches:** They are based on measuring the correlation between pre-defined face templates and the image. The templates can be either rigid [Sinha, 1994] or deformable [Lanitis et al., 1995]. While this approach is simple to implement it is very difficult to deal with variation in scale, pose and shape.

**Appearance based methods:** They rely on models that are not predefined but learned from a set of training images, which should capture possible variations of facial appearance. In general these approaches use techniques from statistical pattern recognition to find relevant characteristics to distinguish between faces and non faces. These characteristics are usually modelled in the form of distribution models or discriminant functions. Several machine learning methods have been applied for face detection, including principal component analysis (PCA) [Turk and Pentland, 1991a], neural networks [Rowley et al., 1996, 1998b], support vector machines [Osuna et al., 1997], sparse network of winnows [Roth et al., 2000], naive Bayes classifier [Schneiderman and Kanade, 1998], and hidden Markov models (HMM) [Samaria, 1994].

The two taxonomies summarized above group existing methods based on how they describe and classify faces. Another way is to distinguish existing approaches based on how faces are represented [Hsu et al., 2002] into

**Holistic approaches:** They try to detect faces as a whole and usually have the advantage of finding smaller faces or faces in poor quality images.

**Component based approaches:** They are based in individual facial components or features and their spatial configuration (topology) with the advantage that they can cope better with different views and partial occlusions.

Method	Representation	Features	Classifier	Topology
[Rowley et al., 1998a]	Holistic	Pixel values	Neural networks	
[Viola and Jones, 2001a]	Holistic	Haar like	AdaBoost	–
[Schneiderman and Kanade, 2000]	Holistic	PCA coefficients, wavelets	Bayes	–
[Lin et al., 2004]	Holistic	Haar like	AdaBoost	–
[Heisele et al., 2003]	Components	Pixel values	SVM	SVM
[Felzenszwalb and Huttenlocher, 2005]	Components	Derivations of Gaussians	Energy function	Energy function

**Table 6.1:** Comparison of selected face detection approaches according to the used representation, features and classification approach. For component based approaches the topology verification is also considered.

Table 6.1 provides an overview of selected face detection approaches. It compares them based on several criteria including face representation, visual description and classification approach. For component based approaches it further considers the method used for topology verification.

### 6.1.2 Challenges

Face detection is a rather difficult task due to the variability of faces itself and the environment. A robust face detection approach needs to consider the following challenges:

**Size:** A face detector should be able to detect faces in different sizes. This is usually achieved by either scaling the input image or the object model. Nevertheless, the size of the object usually influences the reliability of the detection, since smaller faces are more difficult to detect than larger ones.

**Position:** Besides detecting faces in different sizes a face detector should be also able to detect faces at different positions within the image. This is usually achieved by sliding a window over the image and applying the detector at each window position. The choice of the step size directly influences the detection speed and precision.

**Number:** Most of the face detection approaches are able to detect multiple faces within a single image. An important issue here is to handle partially overlapping faces. The standard way to solve this problem is to apply a post filter to remove multiple overlapping faces and derive a single representative face.

**Orientation:** Faces can appear in different orientations within the image plane depending on the angle of the camera and the face. For 2D data such as images two different



rotations can be distinguished: An in-plane rotation (roll) is a rotation along the axis perpendicular to the image plane and leads to a frontal non-upright face. This type of rotation can be easily handled by rotating the image and applying the frontal face detector at different angles. Out-of-plane rotations (pan, tilt) are rotations along the image axes and lead to a non-frontal (half-profile, profile) face. These rotations are usually handled using multiple detectors, e.g. one detector for frontal and one for profile faces.

**Expressions:** The appearance of a face changes considerably for different facial expressions and thus makes the face detection more difficult. These changes are usually considered within the training process of the face detector or by using an adaptable model.

**Illumination:** Varying illumination can be a big problem for face detection since it changes the appearance of the face depending on the color and the direction of the light. The performance of skin color based approaches drops significantly if the image is not white balanced. On the other hand, shadows on parts of the face are similar to occlusions and have a large influence on the performance of appearance based face detection systems.

**Occlusions:** Partial occlusions of faces can be caused by objects within the environment (e.g. poles, persons), objects worn by the person (glasses, scarf, mask), other body parts of the person (hands) and shadows.

The research within the last years has concentrated on improving the performance of face detection under different conditions [Yang et al., 2002], including varying illuminations, poses, expressions. Occlusions are one of the aspects that has been neglected in most of the developments so far. Nevertheless, they present a major challenge for most of the face detection approaches, since they are difficult to model.

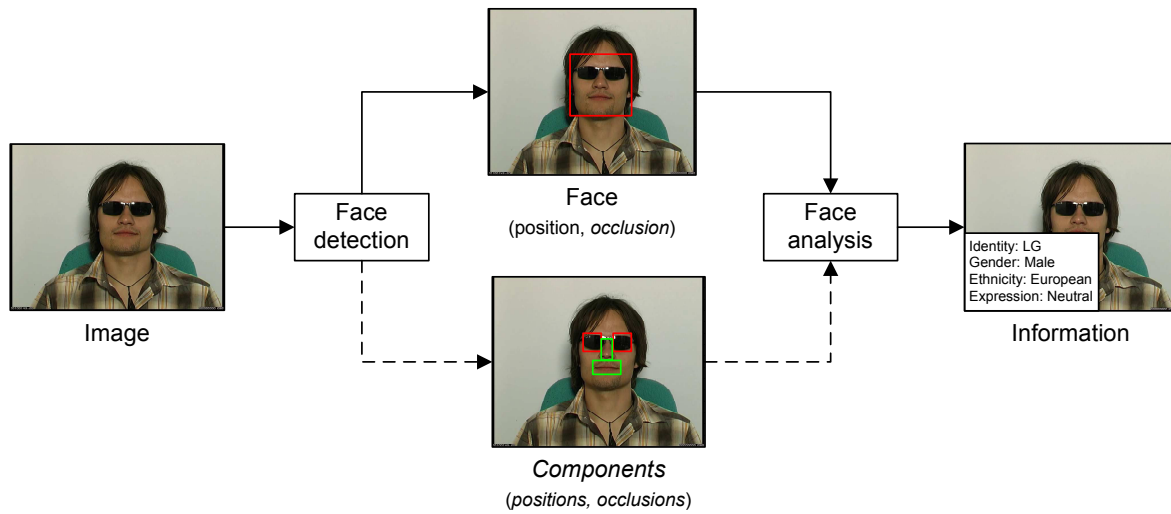
### 6.1.3 Objective

The major objective of this work is to develop a face detection approach that can detect faces robustly even in the presence of partial occlusions. Moreover, to build an *occlusion aware* face analysis system, as shown in figure 6.2, it has to provide additional information regarding the presence and location of occlusions. By using a component based instead of a holistic approach both goals can be achieved in very flexible way. The developed approach will be compared to the well known holistic approach by Viola and Jones [2001a] and the limits of both approaches regarding the different challenges will be explored.

This work has been developed together with Ullrich Mönich<sup>1</sup> (TUB) and is partially described master thesis [Mönich, 2005]. Parts of this work have been published in ICOB 2005 [Goldmann et al., 2005], EI 2006 [Goldmann et al., 2006b], MMSP 2008 [Goldmann et al., 2008c] and TIFS 2007 [Goldmann et al., 2007a].

---

<sup>1</sup>ullrich.moenich@mk.tu-berlin.de



**Figure 6.2:** Overview of the occlusion aware face analysis system. The face detection module provides additional occlusion information (presence, location) beside the typical face information (location, extent) which is used by subsequent face analysis modules to improve the overall performance.

## 6.2 Holistic approach

The well known holistic face detection approach proposed by Viola and Jones [2001a] and its extension by Lienhart et al. [2002] serves as reference for the evaluation.

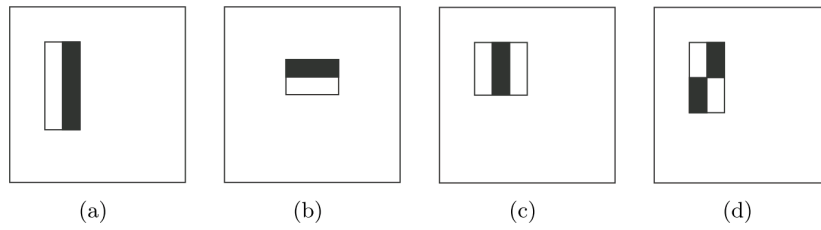
It is based on techniques from statistical pattern recognition and consists of two parts. Within the description step Haar features are extracted to characterize the texture of faces. For the classification a classifier cascade is trained using AdaBoost, which is then used to classify sub windows as faces or non faces. The following sections provide a brief summary of this object detection approach in general. More details can be found in [Viola and Jones, 2001a, 2004; Lienhart et al., 2002].

### 6.2.1 Description

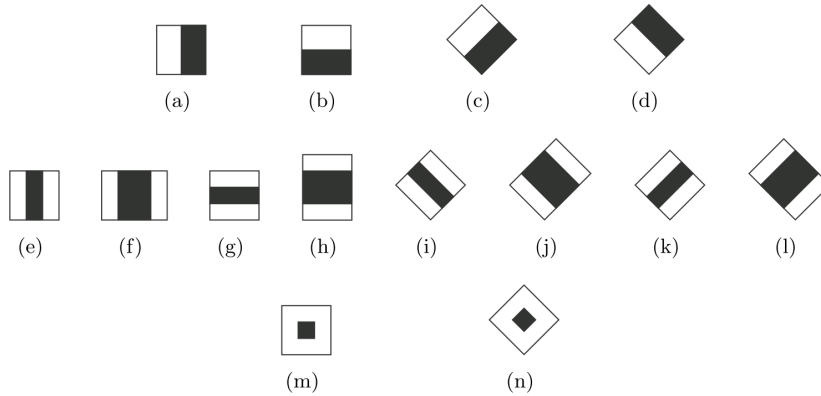
Instead of raw image pixels, low level features are used to describe the image content in order to decrease the intra class variability (between faces) and simultaneously increase the inter class variability (between faces and non-faces). Low level features allow to incorporate domain knowledge which is difficult to learn using a finite number of training samples [Viola and Jones, 2001a]. Furthermore, feature based systems may operate faster than pixel based systems.

#### Haar features

The used feature set consists of several Haar-like features inspired by the Haar basis functions proposed by Papageorgiou [1997]. These features mimic characteristics of the human visual system (HVS) such as edge, line and center surround responses [Lienhart et al., 2002].



**Figure 6.3:** Standard set of Haar features [Viola and Jones, 2001a]: (a-b) single rectangle features, (c) three rectangle features, and (d) four rectangle features.



**Figure 6.4:** Extended set of Haar features [Lienhart et al., 2002]: (a-d) edge features, (e-l) line features, and (m-n) center surround features.

In their original work Viola and Jones [2001a] have proposed a standard set of features (shown in figure 6.3), including two rectangle features, three rectangle features and four rectangle features in both vertical and horizontal direction.

Lienhart et al. [2002] have proposed an extended set of Haar features (shown in figure 6.4) by adding center surround features and diagonal directions. They further show that this extended feature set achieves a better performance than the standard one.

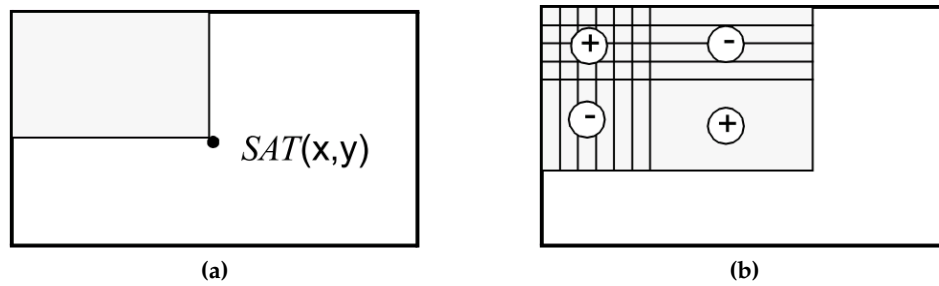
Each of these Haar features  $f_j$  can be computed as the weighted sum of the pixel sums of the individual rectangles  $r_i$  defined as

$$f_j = \sum_{i=1}^N w_i \text{sum}(r_i) \quad (6.1)$$

with the weights  $w_i$  and the number of rectangles  $N = 2$ . The weights  $w_1, w_2$  of the different rectangles (black and white) have opposite signs and are used to compensate for the different size of the rectangles which can be written as

$$-w_1 \cdot \text{area}(r_1) = w_2 \cdot \text{area}(r_2) \quad (6.2)$$

The feature prototypes shown in figure 6.3 and 6.4 are scaled independently in vertical and horizontal direction as well as translated across the image region, which leads to a large



**Figure 6.5:** Illustration of the major idea behind integral images: (a) definition of an integral image and (b) computation of the sum of a rectangle.

and overcomplete<sup>2</sup> set of features.

### Integral image

Both upright and rotated rectangular features can be computed much faster based on so called integral images [Viola and Jones, 2001a] or summed area tables [Lienhart et al., 2002].

For the *upright* case the summed area table  $SAT(x, y)$  is defined as the sum of the pixels  $I(x, y)$  of an upright rectangle from the top left corner  $(0, 0)$  to the bottom right corner  $(x, y)$

$$SAT(x, y) = \sum_{x' \leq x, y' \leq y} I(x', y') \quad (6.3)$$

which can be recursively computed in a single pass over all the pixels. From this the pixel sum of an upright rectangle  $r_i = (x, y, w, h, 0)$  can be determined by four table lookups as

$$\begin{aligned} \text{sum}(r_i) = & SAT(x - 1, y - 1) + SAT(x + w - 1, y + h - 1) \\ & - SAT(x - 1, y + h - 1) - SAT(x + w - 1, y - 1) \end{aligned} \quad (6.4)$$

The general idea of the integral image and the computation of the pixel sum of a rectangle are illustrated in figure 6.5.

For the *rotated* case a rotated summed area table  $RSAT(x, y)$  is computed and used to determine the pixel sum of rotated rectangles in much the same way [Lienhart et al., 2002]. While the recursive computation of the RSAT requires two passes over all the pixels, determining the pixel sum requires four table lookups as well.

Furthermore, the integral image supports fast contrast and brightness normalization within the rectangular regions of the form

$$I'(x, y) = \frac{I(x, y) - \mu}{c\sigma} \quad (6.5)$$

with the mean  $\mu$ , standard deviation  $\sigma$  and the factor  $c = 2$ . While  $\mu$  can easily be derived

<sup>2</sup>A complete basis has no linear dependence between its elements and the same number of pixels as the corresponding image region. The full set of Haar features is many times overcomplete.

from the means of  $SAT$  or  $RSAT$ ,  $\sigma$  requires the sum of squared pixels, which can be derived by computing a second set of  $SAT$  or  $RSAT$  from  $I^2(x, y)$ . Then calculating  $\sigma$  for a given window requires only 4 additional table lookups [Lienhart et al., 2002].

### 6.2.2 Classification

Based on the given feature set and a training set of positive and negative samples the detection is treated as a supervised learning problem. Since the number of features is quite large and only a subset of them may be required for detecting a certain object (e.g. faces) a variant of *AdaBoost* [Freund and Schapire, 1995] is used for selecting appropriate features and training a classifier cascade.

#### AdaBoost

AdaBoost, as a special variant of boosting, belongs to the group of ensemble learning methods [Dietterich, 2001]. The major idea is to build a strong classifier as a combination of several weak classifiers by iteratively adding weak classifiers and reweighting the training samples based on the classification performance.

A weak classifier may be a very simple classifier since it is not expected to perform very well, just a little bit better than random guess. Here a weak classifier  $h_j(x)$  consists of a feature  $f_j$ , an optimal threshold  $\phi_j$  that leads to the lowest error rate and a parity  $p_j$  that indicates the direction of the inequality sign

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) < p_j \phi_j \\ 0 & \text{otherwise} \end{cases} \quad (6.6)$$

Given a set of selected weak classifiers  $h_t(x)$  a strong classifier takes the form of a perceptron, a weighted combination of weak classifiers followed by a thresholding operation

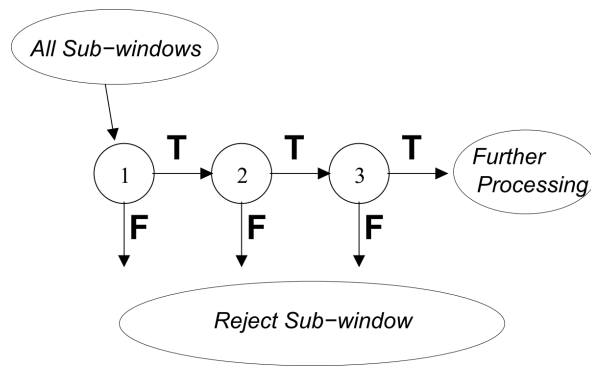
$$h(x) = \begin{cases} 1 & \text{if } \sum_t \alpha_t h_t(x) \geq \frac{1}{2} \sum_t \alpha_t \\ 0 & \text{otherwise} \end{cases} \quad (6.7)$$

with the coefficients  $\alpha_t$  corresponding to the error rates of the weak classifiers  $h_t(x)$ .

The strong classifier is built by iteratively adding the best weak classifier  $h_t(x)$  of each iteration  $t$ . Therefore the weighted error rates  $\epsilon_j$  for the remaining weak classifiers  $h_j(x)$  are computed as

$$\epsilon_j = \sum_i w_i |h_j(x_i) - y_i| \quad (6.8)$$

with the sample  $x_i$ , its label  $y_i \in \{0, 1\}$  and weight  $w_i$ . The weak classifier with the lowest error rate  $\epsilon_t$  is selected as  $h_t(x)$  and the corresponding coefficient is computed as  $\alpha_t = \log(1 - \epsilon_t) / \epsilon_t$ .



**Figure 6.6:** Classifier cascade as a degenerate tree of several classifiers. Within a single stage relevant subwindows are given to the next stage while irrelevant subwindows are immediately discarded.

In order to guide the learning towards difficult samples the weights of correctly classified samples are decreased while those of incorrectly classified samples are retained [Viola and Jones, 2001a]. This is achieved by the following update function

$$w_{i,t+1} = w_{i,t} \beta^{1-e_i} \quad (6.9)$$

with the update factor  $\beta = \epsilon_t / (1 - \epsilon_t)$  and the overall classification error  $e_i = |h(x_i) - y(i)|$ .

### Classifier cascade

In order to achieve a certain performance a monolithic classifier still requires a considerable number of weak classifiers that need to be applied to each of the scan windows. The basic idea of a classifier cascade is to combine several less complex classifiers sequentially (shown in figure 6.6) to achieve the same performance while radically reducing the complexity. A classifier cascade can be seen as degenerated decision tree where each stage is trained to detect almost all objects of interest (low false negative rate) while rejecting a certain fraction of non objects (high false positive rate). Each stage classifies a scan window as either relevant or irrelevant. A positive result from a classifier triggers the evaluation of the next classifier, while a negative result leads to the immediate rejection of the corresponding scan window. Subsequent classifiers are trained using those examples which pass through all the previous stages. As a result, later classifiers face more difficult tasks than former ones, which requires more complex classifiers to achieve comparable performance. But since the number of scan windows decreases for later stages, the overall complexity is much smaller than for a monolithic classifier.

Building the classifier cascade is driven by both detection performance and complexity [Viola and Jones, 2001b]. This leads to an optimization problem that simultaneously minimizes the number of stages as well as the number of weak classifiers and the threshold of each stage for a given target of true positive rate  $TPR$  and false positive rate  $FPR$ . For a

classifier cascade they can be computed as

$$TPR = \prod_{k=1}^K TPR_k \quad (6.10)$$

$$FPR = \prod_{k=1}^K FPR_k \quad (6.11)$$

with the individual rates  $TPR_k$  and  $FPR_k$  of stage  $k$ . Based on given overall target rates and the number of stages, the individual target rates of each stage can be determined. For example, to achieve a overall  $TPR = 0.9$  and  $FPR = 10^{-6}$  with a 10 stage classifier, each stage needs to have a  $TPR_k = 0.99$  (since  $0.99^{10} \approx 0.9$ ) and a  $FPR_k = 0.3$  (since  $0.3^{10} \approx 6 \times 10^{-6}$ ).

The AdaBoost method presented in the previous section is not designed to achieve a high TPR at the expense of larger FPR. Thus the threshold of each strong classifier is decreased to reach the individual goals of each stage and finally the overall goal.

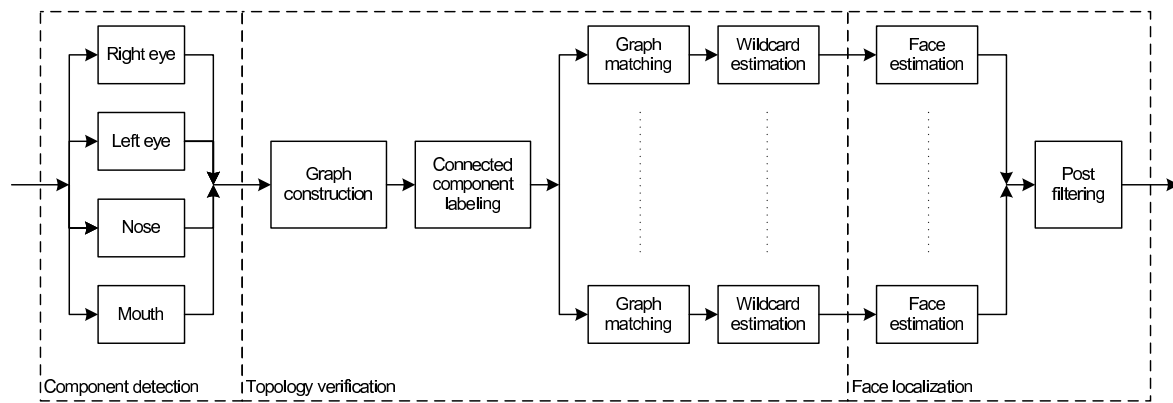
### 6.3 Component based approach

Motivated by recent developments and the human visual perception a component based face detection method has been developed. Thereby, facial components are detected individually and their spatial arrangement (topology) is verified to accept or reject face candidates.

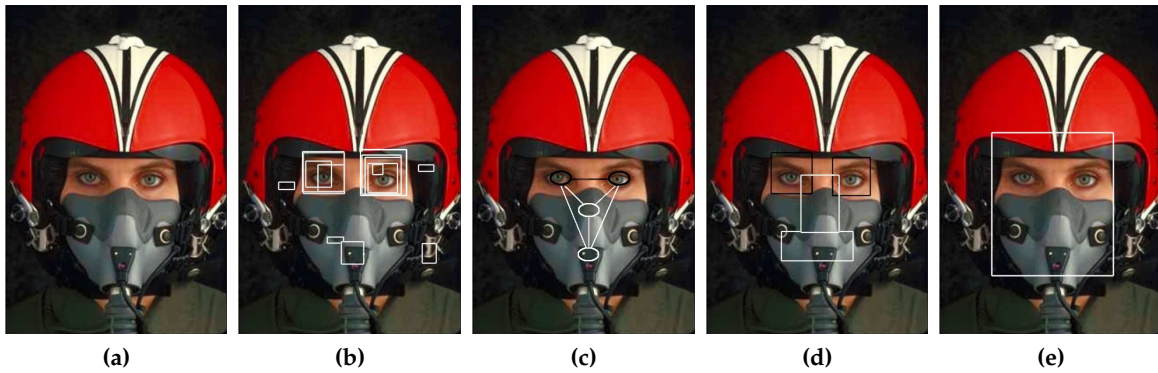
Figure 6.7 provides an overview of the proposed system and its individual steps. In contrast to other systems, techniques from statistical [Jain et al., 2000] and structural pattern recognition [Bunke et al., 2002] are applied. While the component detection is based on techniques from the former domain (see section 3.3), the topology verification relies on concepts from the latter domain (see section 3.5).

Figure 6.8 illustrates the individual steps of the face detection module by showing intermediate results. Figure 6.8(a) shows the input image with partial occlusions. Figure 6.8(b) shows the different components detected individually within the component detection stage. As it can be seen, the nose and the mouth are not detected due to occlusions. Figure 6.8(c) shows the result of the graph matching stage, where the overlayed face graph consists of detected (black) and wildcard (white) components. Based on this the size and location of the wild card components is estimated as it can be seen in figure 6.8(d). Finally, the face region is estimated based on the facial components as shown in figure 6.8(e). Although only two facial components (left and right eye) can be detected by the component detection stage, the face is properly detected.

The following sections describe the individual steps in more detail.



**Figure 6.7:** Overview of the component based face detection method. Within the component detection steps facial components are detected independently from each other. The topology verification step selects combinations of these components that resemble a typical face. Finally the face localization step estimates the missing components and the face region.



**Figure 6.8:** Illustration of the component based face detection method: (a) partially occluded face, (b) detected components, (c) graph matching with overlaid face graph, (d) estimation of wildcard components (e) face localization.

### 6.3.1 Component detection

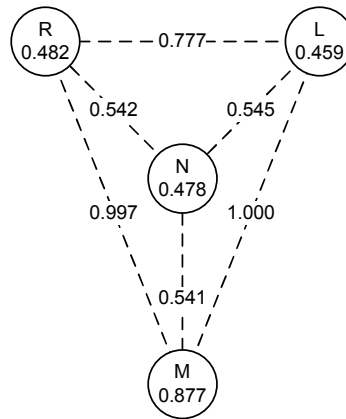
The goal of the component detection is to detect and localize the major facial components, including left eye, right eye, nose and mouth. An individual detector is built for each of the components. All these detectors adopt the same object detection approach described in section 6.2 and differ only in the used training data which is specific for each component.

The individual components are rectangular image regions defined in relation to the anthropometric face regions as already described in section 4.4.

### 6.3.2 Topology verification

The major innovation of the component based face detection system lies in the way how the topology of the different facial components is considered for the face detection task. As it can be seen in figure 6.8(b) the component detection may return too many (e.g. left eye) or too few components (e.g. mouth). While most of the existing approaches [Mohan et al., 2001;





**Figure 6.9:** Facial reference graph with size and distance ratios. Unexpected deviations between the left and the right side are caused by irregularities in the training images.

Huang et al., 2003] use statistical pattern recognition techniques, for modeling the relations between different components, structural pattern recognition techniques are used here.

### Face graph

The topology verification relies on inexact graph matching techniques (see section 3.5), where a facial reference graph is compared with facial candidate graphs in order to accept or reject them as possible faces. In general a graph  $G = (V, E)$  consists of a finite, nonempty set of vertices (nodes)  $V$  and a finite set of edges  $E$ . For modeling faces, the individual components are considered as nodes  $v \in V$  with additional information about the component type  $\kappa_v$  and its size  $\sigma_v$ . Since the aspect ratio of each component type is fixed, only the horizontal dimension is taken. Each edge  $e \in E$  represents the Euclidean distance  $\delta_e = \delta_{ij}$  between the centers of different components  $i$  and  $j$ . Instead of using sizes and differences directly, ratios are used to obtain scale invariance. The graph is naturally invariant to rotation and translation.

### Reference graph

The topology of a typical face is described by a reference graph  $G^R = (V^R, E^R)$  which is created from a set of sample faces by computing the mean of the individual node sizes and edge distances. Each sample graph is constructed by considering the size and the position of manually annotated facial components. The resulting normalized reference graph is shown in figure 6.9. Given the facial reference graph the goal of the topology verification step is to find all subgraphs within the large graph constructed from the detected components that are similar to a typical face.

### Graph construction

The first step after the component detection is to construct a graph from the detected components  $G^D$ . Each facial component becomes a node  $v \in V^D$  and all nodes of different types  $\kappa_v$  are connected with an edge  $e \in E^D$ .

For each node pair  $(i, j) \in V^D \times V^D$  two measures can be defined that represent their deviation to the corresponding nodes  $(r(i), r(j)) \in V^R \times V^R$  in the reference graph  $G^R$ . The size deviation is defined as

$$\Sigma(i, j) = \begin{cases} \frac{\sigma_i/\sigma_j}{\sigma_{r(i)}/\sigma_{r(j)}} - 1 & \text{if } \frac{\sigma_i/\sigma_j}{\sigma_{r(i)}/\sigma_{r(j)}} \geq 1 \\ \frac{\sigma_{r(i)}/\sigma_{r(j)}}{\sigma_i/\sigma_j} - 1 & \text{if } \frac{\sigma_i/\sigma_j}{\sigma_{r(i)}/\sigma_{r(j)}} < 1 \end{cases} \quad (6.12)$$

while the distance deviation is given by

$$\Delta(i, j) = \begin{cases} \frac{\delta_{ij}/\sigma_i}{\delta_{r(i)r(j)}/\sigma_{r(i)}} - 1 & \text{if } \frac{\delta_{ij}/\sigma_i}{\delta_{r(i)r(j)}/\sigma_{r(i)}} \geq 1 \\ \frac{\delta_{r(i)r(j)}/\sigma_{r(i)}}{\delta_{ij}/\sigma_i} - 1 & \text{if } \frac{\delta_{ij}/\sigma_i}{\delta_{r(i)r(j)}/\sigma_{r(i)}} < 1 \end{cases}. \quad (6.13)$$

Edges that violate the typical face topology are discarded based on the following criterion

$$(\Sigma(i, j) > \Sigma_{\max}) \vee (\Delta(i, j) > \Delta_{\max} \wedge \Delta(j, i) > \Delta_{\max}). \quad (6.14)$$

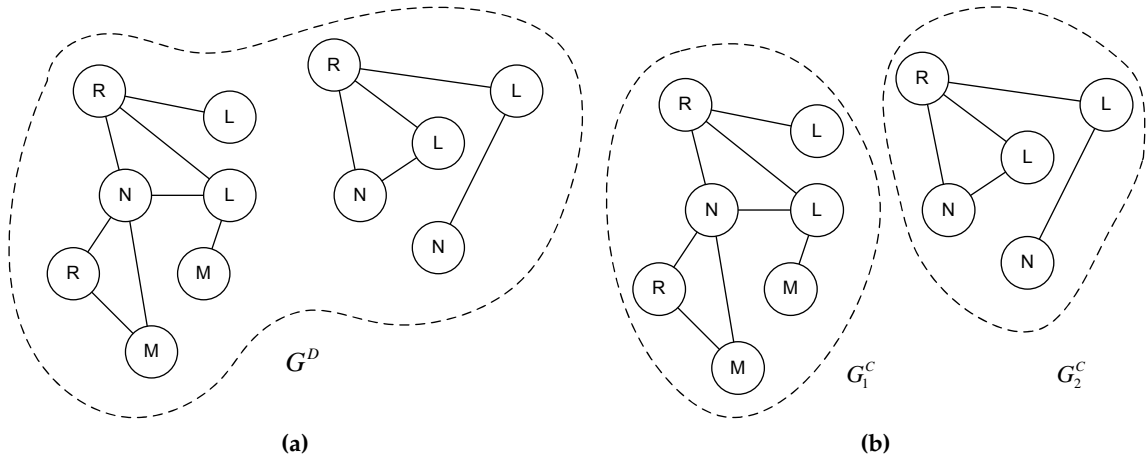
The maximum size deviation  $\Sigma_{\max}$  and the maximum distance deviation  $\Delta_{\max}$  are selected in a way that only edges between components belonging to a single face remain. This is supported by the fact that faces rarely overlap each other and that distances between components of different faces usually violate the face topology.

### Connected component labeling

The graph  $G^D$  is still quite large and consists typically of multiple connected components that correspond to individual faces. The connected component labeling step decomposes the graph  $G^D$  into multiple graphs  $G_i^C = (V_i^C, E_i^C)$ , that are treated as single face candidates. Since the following steps are applied independently for each  $G_i^C$  the index  $i$  is dropped in the following sections. The partitioning helps to increase the speed considerably. Figure 6.10 illustrates that step by showing the graphs before and after the connected component labeling. For real images the graphs are usually much larger with more nodes and edges.

### Graph matching

Each graph  $G^C$  might consist of a variable number of facial components with different types, locations, and sizes. The goal of the graph matching step is to find the best matching sub-graph  $\hat{G}^S$  with respect to the reference graph  $G^R$ . Naturally combinations with a low similarity to the facial reference graph should be discarded.



**Figure 6.10:** Connected component labeling to split the overall face graph into individual face candidates: (a) overall face graph  $G^D$  and (b) two connected components  $G_1^C$  and  $G_2^C$ .

In order to cope with missing and impossible components, a wildcard component  $v \in V^W$  for each type, with  $V^W$  being the set of all wildcard components, is introduced without any size or distance information.

From the resulting graph all possible subgraphs  $G_i^S = (V_i^S, E_i^S)$  with four different and at least two detected components are chosen. For each of them the matching cost

$$c(G_i^S) = \sum_{v \in (V_i^S \cap V^C)} Y(v) + \sum_{(u,v) \in (V_i^S \cap V^C)^2} B(u,v) + \sum_{v \in (V_i^S \cap V^W)} W(v) \quad (6.15)$$

is calculated with respect to the reference graph  $G^R$ . The unary function  $Y(v)$  is used to measure the dissimilarity based on one detected component. Here, the detection reliability based on some skin color criteria may be used. The binary function  $B(u,v)$  measures dissimilarities based on two detected components, such as size and distance ratios. The function  $W(v)$  measures the costs for missing components (wildcards).

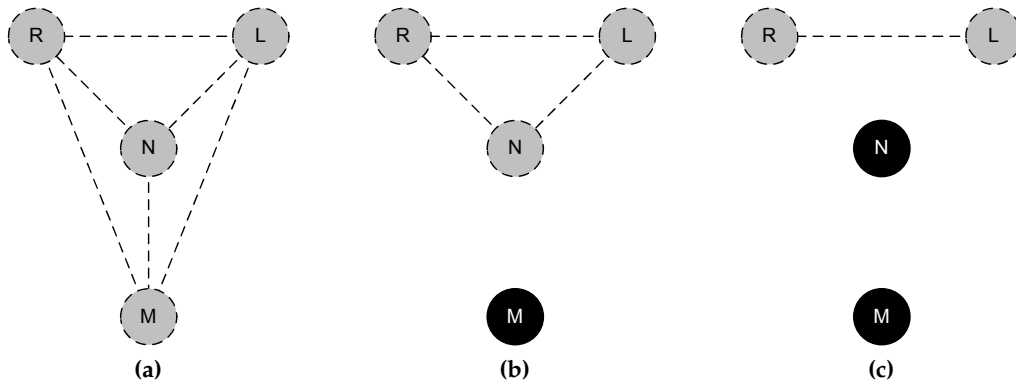
In the actual system, the costs are calculated based the following equations

$$B(u,v) = \begin{cases} P_{ME} & \text{if } \{u,v\} \notin E_i^S \\ \alpha \Sigma(u,v) + \beta \Delta(u,v) & \text{else} \end{cases} \quad (6.16)$$

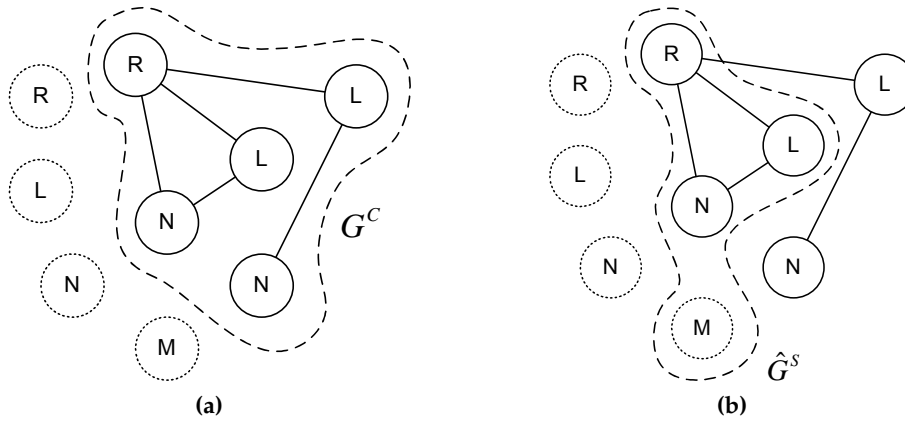
$$W(v) = P_W \quad (6.17)$$

$$Y(v) = 0 \quad (6.18)$$

with  $P_{ME}, P_W$  being constant costs for missing edges and wildcard components, respectively. The constants  $\alpha, \beta$  are used to weight the size and distance ratios. Figure 6.11 depicts the different cost terms of the overall cost function for all possible numbers of wildcard components.



**Figure 6.11:** Cost function with unary (gray), binary (dashed) and wildcard (black) cost terms for different cases: (a) 4 detected components, (b) 3 detected and 1 wildcard component, (c) 2 detected and 2 wildcard components.



**Figure 6.12:** Graph matching with detected (solid) and wildcard components (dashed): (a) before the graph matching with connected component  $G^C$ , (b) after graph matching with best matching subgraph  $\hat{G}^S$ .

Out of all possible subgraphs  $G_i^S$  the one with the smallest cost  $c(G_i^S)$  is chosen and called the best matching subgraph  $\hat{G}^S$ :

$$\hat{G}^S = \arg \min_{G_i^S} c(G_i^S). \quad (6.19)$$

It is considered to be a face if its cost is below the predefined threshold  $c_{\max} = 21.2$  which was found empirically:

$$c(\hat{G}^S) \leq c_{\max}. \quad (6.20)$$

Figure 6.12 illustrates the result of the graph matching step based on the connected component  $G_2^C$  from figure 6.10. It shows the connected component before the graph matching (figure 6.12(a)) and the best matching subgraph (figure 6.12(b)).

### Wildcard estimation

The best matching subgraph returned from the graph matching step may include several wildcard components without any size or location information. In order to estimate the face region reliably the information of all facial components is needed. The missing information of the wildcard components is estimated based on the detected components, the reference graph, and the orientation information of the face.

The orientation information is necessary to resolve ambiguities caused by the mirroring invariance of the graph model that leads to two possible locations for each component. The orientation  $o$  can be computed based on the position of three components  $p_1 = (x_1, y_1)^T$ ,  $p_2 = (x_2, y_2)^T$  and  $p_3 = (x_3, y_3)^T$  using the determinant

$$o = \det \begin{pmatrix} x_1 - x_3 & x_2 - x_3 \\ y_1 - y_3 & y_2 - y_3 \end{pmatrix} \quad (6.21)$$

The rotation angle  $\phi$  is computed using the law of cosine and the distances within the reference graph  $d_{ij}$  by

$$\phi = \arccos \frac{d_{12}^2 + d_{13}^2 - d_{23}^2}{2d_{12}d_{13}} \quad (6.22)$$

The rotation angle  $\phi$  can be either positive or negative. The correct angle is chosen based on the orientation  $o$  as

$$\phi = \begin{cases} \phi & \text{if } o > 0 \\ -\phi & \text{if } o < 0 \end{cases} \quad (6.23)$$

Given the rotation angle  $\phi$ , the position of two detected components  $\vec{p}_1$ ,  $\vec{p}_2$ , and the distances within the reference graph  $d_{12}$ ,  $d_{13}$  the position of the wildcard component  $\vec{p}_3$  can be computed as

$$p_3 = \frac{d_{13}}{d_{12}} A(\vec{p}_2 - \vec{p}_1) + \vec{p}_1 \quad (6.24)$$

with the rotation matrix

$$A = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}. \quad (6.25)$$

### 6.3.3 Face localization

#### Face estimation

After verifying face candidates based on the topology of their components and estimating missing (occluded) components, the final step estimates the region of the faces from the corresponding components. The face region is described by a rectangle  $\vec{r} = (x, y, w, h)$  where  $(x, y)$  corresponds to its center and  $(w, h)$  to its size respectively. The face region is then

computed as

$$x = x_n \quad (6.26)$$

$$y = y_n \quad (6.27)$$

$$w = 2(x_r + x_l) \quad (6.28)$$

$$h = \frac{2(y_r + y_l)}{2 - y_m} \quad (6.29)$$

with the indices  $r, l, n, m$  corresponding to the right eye, left eye, nose and mouth.

### Postfiltering

While the previous steps have been applied independently for each of the connected components  $G_i^C$  the postfiltering considers the whole set of detected faces. Due to the connected component labeling stage, it is very unlikely to have overlapping face regions within this set. Nevertheless, if several face regions overlap the face region with the lowest matching cost is retained and all other face regions are discarded. This strategy is more suitable than just taking the median box of the overlapping faces.

### 6.3.4 Occlusion localization

Due to the component based face detection approach the extension to classify a face and its components as being occluded or not is straight forward.

#### Component classification

The best matching subgraph may consist of detected and missing (wildcard) components. Missing components are usually caused by a large difference between the corresponding region and the statistical model, which lets the component detection discard this region. This appearance change can be caused by occlusions or shadows which can be seen as a special type of occlusion. Given that, a component  $o_c$  is classified as occluded if it corresponds to a wildcard component and classified as non occluded if it corresponds to a detected component which can be written as

$$o_c = \begin{cases} 0 & \text{if detected} \\ 1 & \text{if wildcard} \end{cases} \quad (6.30)$$

#### Face classification

The classification of the face as occluded or not is directly based on the component classification using a simple rule. If at least one component is classified as occluded the face is considered as (partially) occluded.

$$o_f = \bigcup_c o_c \quad (6.31)$$

Parameter	Value	Parameter	Value
Positive samples	2029/4058	Size threshold $\Sigma_{\max}$	1.45
Negative samples	3000	Distance threshold $\Delta_{\max}$	1.25
Symmetry	no/yes	Missing edge cost $P_{ME}$	10
Minimum TPR	0.999	Wildcard cost $P_W$	10
Maximum FPR	0.5	Size weight $\alpha$	1
Feature set	extended	Edge weight $\beta$	1
Boosting method	Gentle AdaBoost	Cost threshold $c_{\max}$	21.2
Weight trimming	0.95		

(a)

(b)

**Table 6.2:** Optimized parameter set for the component based face detection approach: (a) parameters for the training of the component detectors (b) parameter set for the topology verification.

Depending on the application the rule can be adapted, e.g. for lip reading a face may be considered as occluded only if the mouth region is occluded, since eye occlusions would not effect the performance.

## 6.4 Experiments

The goal of the different experiments is to assess the performance of the developed component based face detection approach (section 6.3) and compare it to the state of the art holistic approach (section 6.2). Furthermore, the limits of both approaches with respect to the different challenges (size, view, occlusion) will be explored. Finally, the occlusion classification of faces and their individual components is assessed.

These experiments are based on a set of empirically optimized parameters for the component based face detection approach, which are shown in table 6.2. Furthermore, two versions of the holistic face detector were used. The first one corresponds to the default face detector (stump based, 24x24 pixels, discrete Adaboost) which is provided within the OpenCV library<sup>3</sup>. Unfortunately, no information regarding the training settings is provided. The second holistic face detector has been trained with same images and settings as the component detectors. That these parameters and the used training images are not as optimal for the face detection, will be shown in the experimental results.

Table 6.3 summarizes the obtained component and holistic face detectors which have been used within this work. All are based on the same statistical learning approach that combines Haar features with a Adaboost trained classifier cascade.

<sup>3</sup><http://sourceforge.net/projects/opencvlibrary/>

Title	Size	Adaboost	Stages
Right eye	16x16	Gentle	25
Left eye	16x16	Gentle	24
Nose	12x18	Gentle	28
Mouth	26x11	Gentle	26
OpenCV	24x24	Discrete	25
Holistic	24x24	Gentle	22

**Table 6.3:** Overview of the component detectors (upper part) and holistic face detectors (lower part) which are all based on the same object detection approach using Haar features and a Adaboost trained classifier cascade.

### 6.4.1 Dataset

Similar to other detection tasks, developing and evaluating a face detection method requires a large amount of data. During the training of the detector positive and negative training samples are necessary to learn a reliable face model. For the testing a representative dataset is required to obtain general performance results.

Although, a lot of corpora have been developed for face analysis, it is still difficult to obtain a comprehensive set for developing and evaluating face detection approaches. Several surveys provide comparisons of available corpora based on certain criteria, including media type (image, video), number of images, color (color, grayscale), number of subjects, image size, face size, views (frontal, profile), expressions, illuminations, background. Another very important criteria for automatic evaluation is the availability of ground truth annotations in the form of boxes or points describing present faces. Since the annotation of this ground truth is a very time consuming process only some corpora provide this information.

Table 6.4 provides an overview of available corpora which are suitable for developing and evaluating face detection approaches together with their characteristics. The corpora used within the following experiments are shown in the upper part of the table. As already mentioned several corpora have been used for the experiments. For training the face detection approaches a combination of the 1512 images from the BioID Face Database and 508 images from the AR Face Database have been used as positive samples, while 19973 images from the Data Becker 222222 Database have been used as negative samples. For testing the face detection approaches 143 images from the Corel Gallery Magic 65000, AR Face Database and Caltech 101 Object Categories have been considered. Although this test set is not very large it contains a large variety of variations for a comprehensive evaluation. For evaluating the component/face detection/classification a combination of the well known AR Face Database (described in section A.2.3) and the novel VISNET II Face Database (described in section A.2.3) has been used.



Database	Number	Color	Size	Annotation
AR Face Database	3315	RGB	768x576	yes
BioID Face Database	1521	GS	384x286	yes
Caltech-101 Object Categories	450	RGB	896x592	no
Data Becker 222222	20000	RGB	900x900	no
Corel Gallery Magic 65000	65000	RGB	400x400	no
UMIST Face Database	564	GS	220x220	no
ATT Face Database	400	GS	92x112	no
CMU Frontal Face Database	180	GS	Different	yes
CBCL Face Database	31022	GS	19x19	no
Yale Face Database	165	GS	243x340	no

**Table 6.4:** Overview of available databases for face detection grouped into considered (upper part) and discarded (lower part) ones.

### 6.4.2 Evaluation

#### Detection

Both the face and the component detection can be treated as a detection problem (see section B.2 for more details). In order to evaluate both tasks in a similar way, the methodology of the PASCAL Visual Object Classes Challenge (VOC) [Everingham et al., 2005] has been adopted.

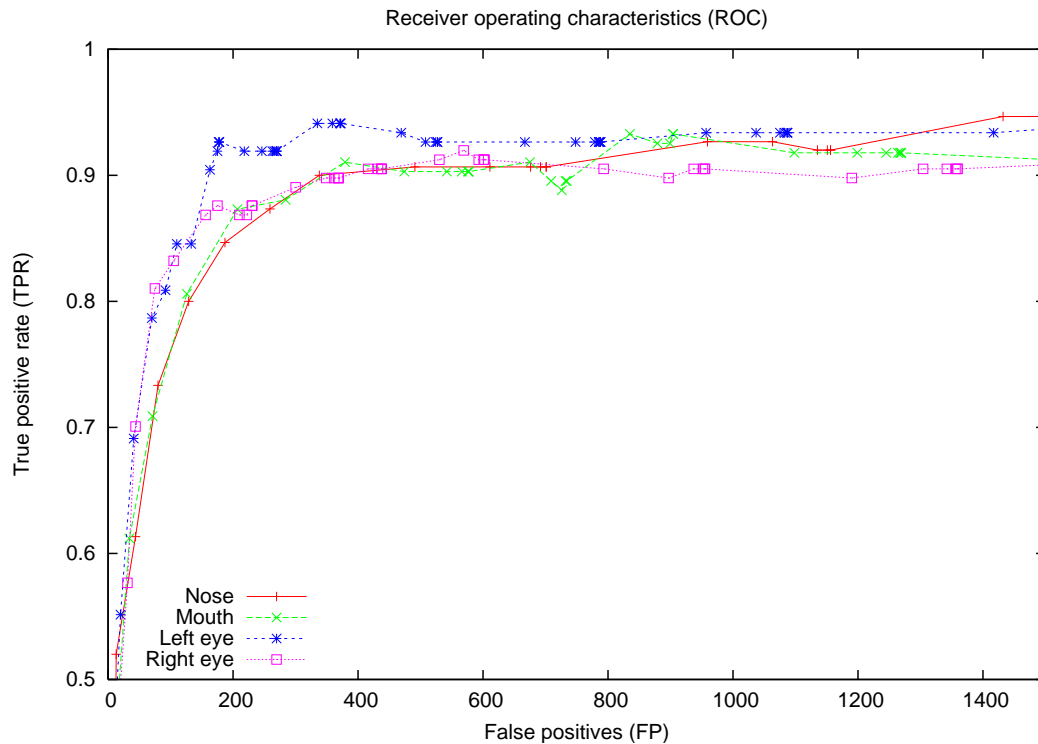
It compares the location of annotated and predicted objects based on their bounding boxes. To be considered as a detection the ratio  $a_o$  between the overlapping area of a predicted box  $b_{pr}$  and a ground truth box  $b_{gt}$  and the overall area of both boxes defined as

$$a_o = \frac{\text{area}(b_{pr} \cap b_{gt})}{\text{area}(b_{pr} \cup b_{gt})} \quad (6.32)$$

has to exceed 50%.

Considering only one-to-one assignments a match matrix between all ground truth and predicted objects is built. Based on this match matrix a confusion matrix with the number of true positives (TP), false positives (FP) and false negatives (FN) are derived. Based on these numbers several measures can be computed including the true positive rate (TPR), false positive rate (FPR), recall (R), precision (P).

In any detection task one can usually achieve a tradeoff between TP and FP by adjusting a detection threshold. Decreasing the detection threshold usually leads to an increase of both TP and FP and vice versa. Based on that, different curves can be derived that show the tradeoff between the different measures. Receiver operating characteristic (ROC) curves (see section B.2.1 for more details), are considered here.



**Figure 6.13:** Comparison of the individual component detectors using ROC curves. Both eye detectors achieve a higher performance than the nose and mouth detectors due to their more unique appearance. The considerable difference between the two eyes may result from differences in the training data.

## Classification

The classification of the face and its components as being occluded or not can be treated as a recognition problem (see section B.4 for more details). Using ground truth and predicted occlusions a confusion matrix can be built, from which the recognition rate (RR) can be computed given the correct classifications and incorrect classifications.

### 6.4.3 Results

#### Component detection

This section summarizes the results of the facial component detection which is the first step within the proposed face detection system.

Figure 6.13 shows the resulting ROC curve for the facial component detection. As it can be seen from the figure, the detection performance of the different facial components varies considerably. Since the training settings and the amount of training images are similar for all components, the difference might be caused by the different levels of uniqueness of the individual component types. Another interesting result is the unexpected difference between the left and the right eye. With the assumption that both eyes have the same level of

uniqueness this might be due to variations in the training images, which lead to a difference in the number of cascade stages between the two eyes as it can be seen in table 6.3.

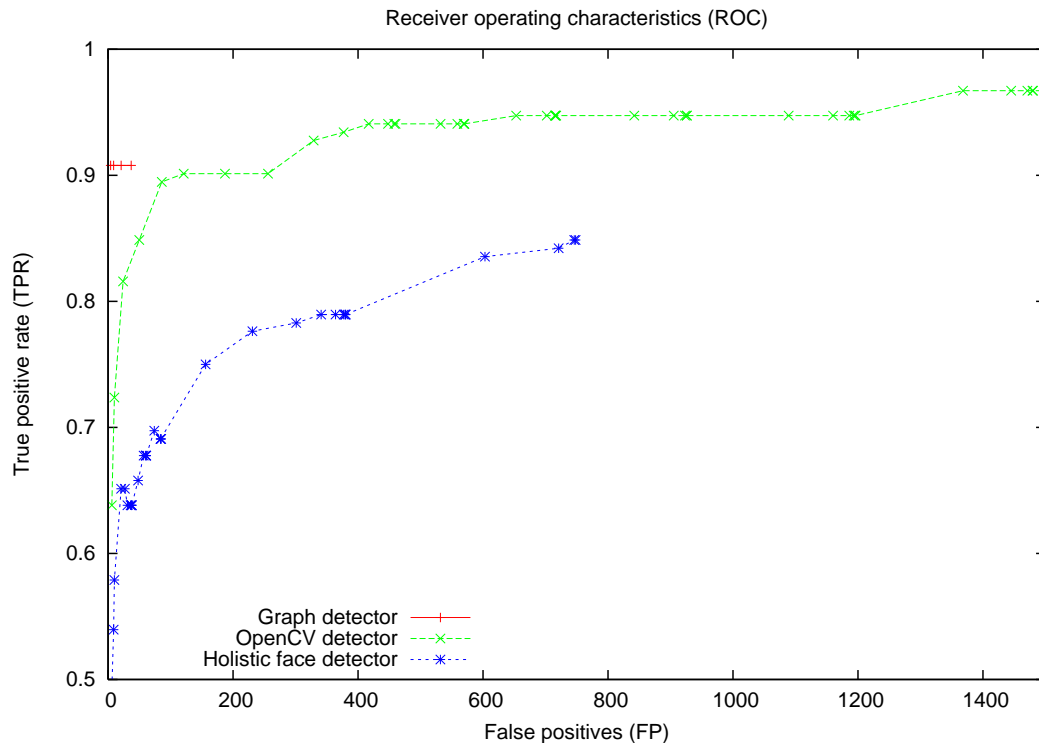
### Face detection

After assessing the performance of the component detectors, this section summarizes the overall performance of the proposed face detection system. Therefore, the proposed component based approach (called graph detector) is compared with two holistic systems. As reference the system proposed by Lienhart et al. [2003] (called OpenCV detector) is used, which utilizes a similar approach to our facial component detectors but applies it directly to the whole face. Since the performance depends a lot on the used training data, another version of this detector (called holistic face detector) is considered for the comparison that is trained with the same data as the component based approach.

The idea of the conducted experiments was to measure the overall performance of the proposed component based detector for partial occlusions, different out of plane rotations and face sizes. The used database does not contain enough samples for analyzing different types of occlusions individually.

Figure 6.14 shows the resulting ROC curves of the experiments with the three detectors. Due to the specific nature of the component based detector it is impossible to generate a large number of false positives, which results in a ROC curve limited to lower FP. Since the component detection step already has a very low number of false positives and the connected component detection discards a lot of impossible combinations it is rather difficult to obtain values for higher FP. Furthermore, the higher FP range is not very interesting since it is not suitable for real applications. Nevertheless, it is obvious that the component based approach is far superior to both holistic approaches. It reaches a  $TPR = 90.8\%$  for  $FP = 4$ . For a similar  $FP$  the reference system reaches only a  $TPR = 63.8\%$ . A  $TPR = 90\%$  is not reached until  $FP = 300$ . The performance improvement is achieved due to the higher robustness of the component based approach regarding occlusions (e.g. wearing a helmet, scarf, sunglasses) and out of plane rotations (tilt, pan). Another interesting result is the big difference between the two holistic detectors. Since both are based on the same approach, the difference must be caused by the training settings and the training data.

By comparing the results of the proposed detector and the reference detector (figure 6.14) with the component detectors (figure 6.13), another interesting fact is discovered. Although the individual component detectors show a lower performance than the holistic reference detector, the overall component based detector shows a much higher performance than the latter one. This improvement justifies the graph based approach for utilizing the topology in a flexible way.

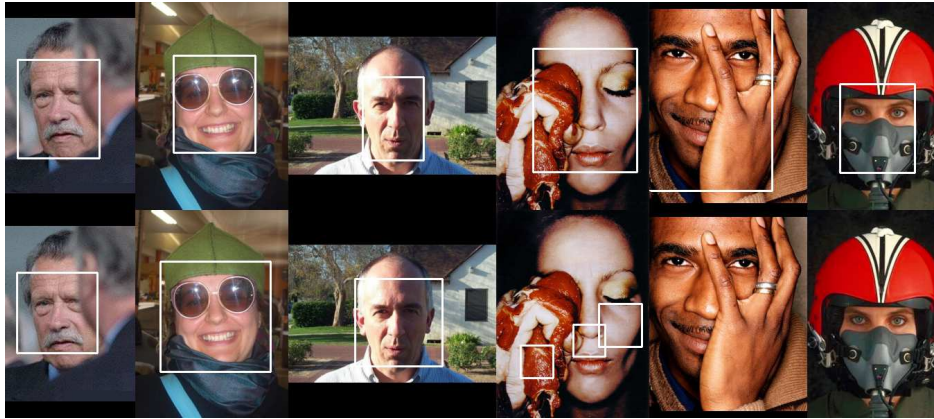


**Figure 6.14:** Comparison of the different face detection approaches based on ROC curves. The proposed component based face detector achieves a much higher performance in the practically relevant left half of the ROC curve. The large difference between the OpenCV and the holistic detector shows the influence of a representative training set and parameter tuning on the performance.

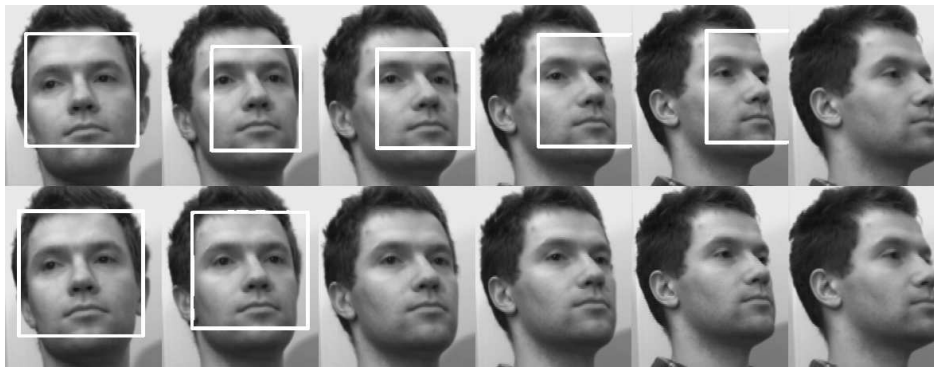
## Challenges

Additional experiments were carried out in order to analyze the performance of the proposed face detection approach more deeply. For these experiments the graph detector is compared against the OpenCV detector as the reference.

In the first experiment, the behaviour of both detectors regarding partial occlusions was analyzed. Partial occlusions can be caused by shadows, hands in front of the face, glasses, hands, and even other persons. Figure 6.15 presents some typical examples of these different types of partial occlusions. In general the graph detector (upper row) is able to handle partial occlusions more reliably than the OpenCV detector (lower row). It seems that the OpenCV detector depends largely on the type of occlusion and the texture difference of the occluding object and a typical face. Furthermore, the face localization of the graph detector in the presence of occlusions is more precise than that of the OpenCV detector, which returns often slightly shifted regions. The performance of holistic face detectors in the presence of partial occlusions depends largely on the training stage. If the training set includes partially occluded faces the detector will more likely detect faces with partial occlusions. On the other hand, this might also increase the number of false positives. Another way is to train multiple detectors for each type of partial occlusion as proposed by Lin et al. [2004]. Nevertheless,



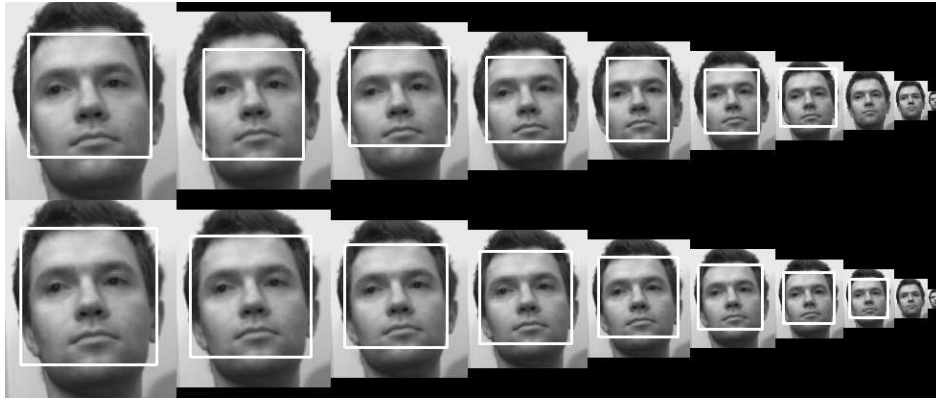
**Figure 6.15:** Comparison of the face detection approaches for different occlusions based on visual samples. Upper row: component based approach, lower row: holistic approach. It can be seen, that the component based approach is both more reliable and more precise in the presence of partial occlusions.



**Figure 6.16:** Comparison of the face detection approaches for different views based on visual samples. Upper row: component based approach, lower row: holistic approach. It can be seen, that the component based face detector is able to handle larger out-of-plane rotations than the holistic face detector.

the component based approach provides a more flexible and natural way to handle partial occlusions.

The goal of the next experiment was to analyze the performance of both detectors for out of plane rotations (pans). The UMIST face database was used for this experiment. It provides images of several persons with different panning angles between 0 and 90 degrees. Figure 6.16 presents representative samples for the graph detector (upper row) and the OpenCV detector (lower row) with different angles (approx. 0, 10, 20, 30, 40, 50 degree). It is important to remember that both detectors have been trained for frontal faces only, which enables a fair comparison. It is obvious that the graph detector can handle much larger out of plane rotations than the OpenCV detector. While the holistic approach works only up to an angle of 15 degrees, the component based approach is able to detect faces up to an angle of 45 degrees. Although the component based detector is still able to detect faces for larger angles, the precision of the face localization drops due to the distortion of the facial components and



**Figure 6.17:** Comparison of the face detection approaches for different sizes based on visual samples. Upper row: component based approach, lower row: holistic approach. While the holistic face detector can detect faces with minimum eye distance of approx. 18 pixels, the component based face detector requires approx. 27 pixels.

the face region. This experiment shows that a component based approach can handle larger out-of-plane rotations than holistic detectors. The reason for this is that the appearance of the whole face changes much more than the appearance of individual facial components. Furthermore, even if a single component may get occluded (e.g. eye occluded by the nose) the component based detector is still able to detect the face. The drop in precision for larger rotation angles suggests its use as prelocalization method and the use of other methods for finer localization.

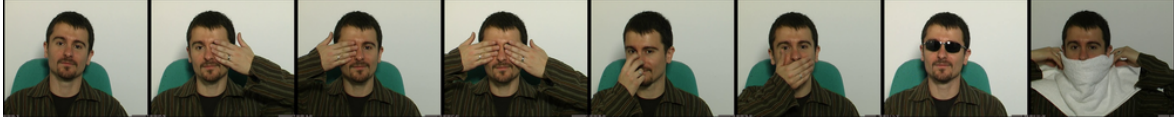
The last experiment was conducted to analyze the performance of both detectors for different resolutions and determine the lower resolution bounds. For this experiment the images of the UMIST Face Database were subsampled to different resolutions. Figure 6.17 shows representative samples for the graph detector (upper row) and the OpenCV detector (lower row) with different resolutions (eye distances of 61, 55, 49, 43, 38, 32, 27, 18, 11, 5 pixels). As expected the holistic detector (eye distance of 18 pixels) can detect faces at lower resolutions than the component based detector (eye distance of 27 pixels). The reason for this is that the facial components are much smaller than the overall face and a higher resolution is required to detect them than for detecting the whole face. In conclusion the component based face detector requires faces with a minimum height of 60 while the holistic face detector requires faces with a minimum height of 40 pixels. This result for the OpenCV face detector is comparable to the one already reported by Cucchiara [2005].

### Component/face classification

This experiment focuses on the performance of the component based face detection approach with respect to occlusions. It evaluates not only the performance of the face and component detection but also the performance of classifying them as occluded or not. The evaluation has been done on a subset of the AR Face Database (see figure 6.18) and the VISNET II Face Database (see figure 6.19).

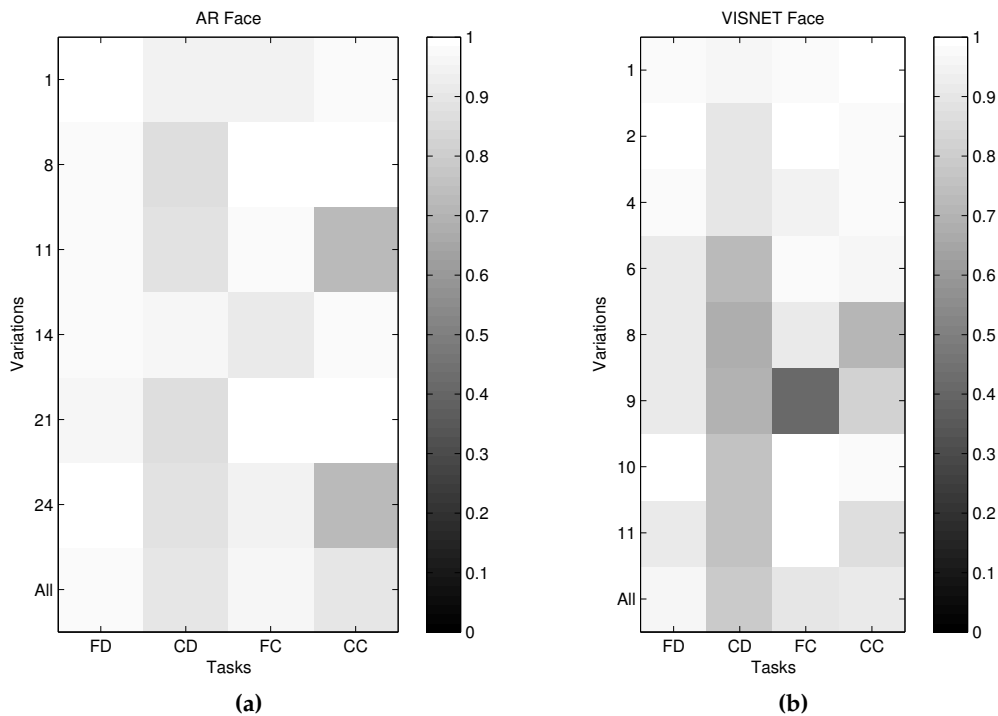


**Figure 6.18:** Subset of the AR Face Database with occlusions. From left to right the variations are 1,8,11,14,21,24.



**Figure 6.19:** Subset of the VISNET II Face Database with real occlusions. From left to right the variations are 1,2,4,6,8,9,10,11.

The face/component detection/classification performance across the different variations of the two databases is shown as a pseudo color plot in figure 6.20. By visual inspection several observations can be made. The performance of the face detection (FD) is quite high for faces without and with occlusions. The performance of the component detection (CD) is not as good due to larger localization error in comparison to the face detection. The performance of the face classification (FC) is very good apart from variation 9 of the VISNET II



**Figure 6.20:** Performance of the face/component detection/classification tasks over the different variations of the (a) AR Face Database and (b) VISNET II Face Database. The face (FD) and component detection (CD) performance is measured as f-measure (F) while the face (FC) and component classification (CC) performance as recognition rate (RR).

Task	AR	VISNET	All
Face detection (FD)	0.978	0.964	0.971
Component detection (CD)	0.896	0.828	0.862
Face classification (FC)	0.963	0.911	0.937
Component classification (CC)	0.897	0.930	0.914

**Table 6.5:** Overall performance of the different component/face detection/classification tasks for the AR Face and the VISNET II Face Database.

Database which corresponds to a hand in front of the mouth. Here the visual similarity between the mouth and the fingers causes the mouth detector to detect a mouth in this region. Subsequently the face is classified as non occluded, since all components can be detected. This illustrates a general problem, if the occlusion is visually similar to the occluded region. For these cases different information (e.g. depth) is required to detect occlusions. The results of the component classification (CC) are quite promising. The performance is lower for some variations (11,24) of the AR Face Database and one variation (8) of the VISNET II Face Database. This is mainly caused by the ambiguity of an occlusion, e.g. beside the mouth a scarf may cover half of the nose, which is not considered within the ground truth.

A summary of the performance of the different tasks is provided in table 6.5 by first averaging the measures over the variations and finally over the databases. The face detection achieves an f-measure of over 97% which demonstrates the robustness of the component based approach regarding occlusions. The localization of the components is not as precise and leads to an f-measure around 86%. The performance of the face and the component classification is very promising with recognition rates of 93% and 91% respectively.

## 6.5 Conclusion

### 6.5.1 Summary

A novel component based face detection approach has been developed that combines techniques from the statistical and the structural pattern recognition domain in an intuitive way to detect faces robustly, even in the presence of partial occlusions. Individual components are detected based on a combination of Haar-like features and an AdaBoost trained classifier cascade. The spatial configuration (topology) of the detected components is verified using inexact graph matching techniques. Therefore, the arrangement of facial components is represented by a graph model that considers size and distance ratios between individual components. In contrast to other component based approaches that utilize statistical pattern recognition techniques for the topology verification, the proposed method can handle components that have not been detected due to occlusions. Given the position of the individual components (either detected or estimated) the face region is derived. Beside the location and extend of detected faces and their components the method provides additional infor-



mation regarding present occlusions by classifying the face and components as occluded or not. This information can be used to improve the performance of following face analysis steps [Goldmann et al., 2008c].

Extensive experiments provide an in depth analysis of the proposed method and a comparison with the well known holistic face detection approach by Viola and Jones [2001a]. It is shown that the component based approach outperforms the holistic approach and achieves a higher robustness especially in the presence of occlusions and out of plane rotations. Due to the higher resolution needed by the component detectors, it requires a slightly higher face resolution than the holistic approach. This result supports the idea of a hierarchical face analysis approach similar to the human visual perception (see chapter 4 for a discussion) for improved performance in a large variety of conditions. Furthermore, it is shown that the component based approach can provide reliable information regarding the presence and location of partial occlusions by classifying the face and its components as occluded or not.

### 6.5.2 Future work

Although the experiments show that the proposed face detection approach outperforms the well known approach by Viola and Jones [2001a] considerably a few directions for future research are discussed here.

The component detection step plays a crucial role for the overall detection performance. While the adopted component detection approach is very fast, it is not very precise due to the Haar-like features that provide only a very rough contrast description of the texture within an image region. In order to increase the performance of the component detection more sophisticated texture features such as edgelets [Wu and Nevatia, 2007a], local binary patterns (LBP) [Ojala et al., 2000] and histograms of oriented gradients (HOG) [Dalal, 2006] may be used.

Since the component detector scans the image at different positions and scales it usually returns multiple overlapping detections for each component. The graph matching step handles this problem implicitly by choosing the best matching subgraph which considers only the best fitting detection of each component. Nevertheless, a large number of subgraphs with only minor cost differences exists due to the other very similar detections. If the image changes slightly due to noise or illumination another subgraph may be selected as the best matching subgraph leading to considerable changes in face location and extend. This can be solved by either fusing multiple detections [Viola and Jones, 2004] before the graph matching or combining similar subgraphs into a median subgraph [Bunke et al., 2002].

Although the approach has been developed for detecting faces it is generic enough to be applied for other objects as well. In order to make it trainable for other object classes an automatic way for selecting the individual components is required. One possible way is to consider salient points or regions that share a common topology across the different training samples. Commonly used point detectors are Harris [Harris and Stephens, 1988], Laplacian

[Lindeberg, 1998], and difference of Gaussians (DoG) [Lowe, 2004]. The local image regions around these points can be described by means of scale invariant feature transform (SIFT) [Lowe, 2004] or shape contexts [Belongie et al., 2001].

## Chapter 7

# Face recognition

### 7.1 Introduction

In recent years face recognition has received substantial attention from researchers in biometrics, pattern recognition and computer vision. This common interest among researchers from diverse fields is motivated by the fact that facial appearance is one of the major clues for humans to identify each other. Besides, there are a large number of applications that require face recognition technology, including automated surveillance, access control, mugshot identification, human computer interaction and multimedia search and retrieval.

The general goal of face recognition is to describe the facial appearance of persons in a robust way to recognize their identities. Beyond this biometric task, facial appearance can also be used to recognize the gender or ethnicity of a person and to estimate her age (see figure 7.1 for some examples).

Like any other face analysis task, face recognition has to cope with the common challenges (pose, illumination, expression, occlusions, aging). The major problems for face recognition are caused by the relation between inter and intra subject variations. On the one hand, the *inter subject variations* can be quite small due to the similarity of appearance across individuals (see figure 7.2). On the other hand, large *intra subject variations* can occur due to different poses, illuminations, expressions and partial occlusions (see figure 7.3).



**Figure 7.1:** Face recognition beyond determining the identity of a person. (a) gender recognition, and (b) ethnicity recognition.



**Figure 7.2:** Small inter subject variations of faces between twins (a) and non twins (b).



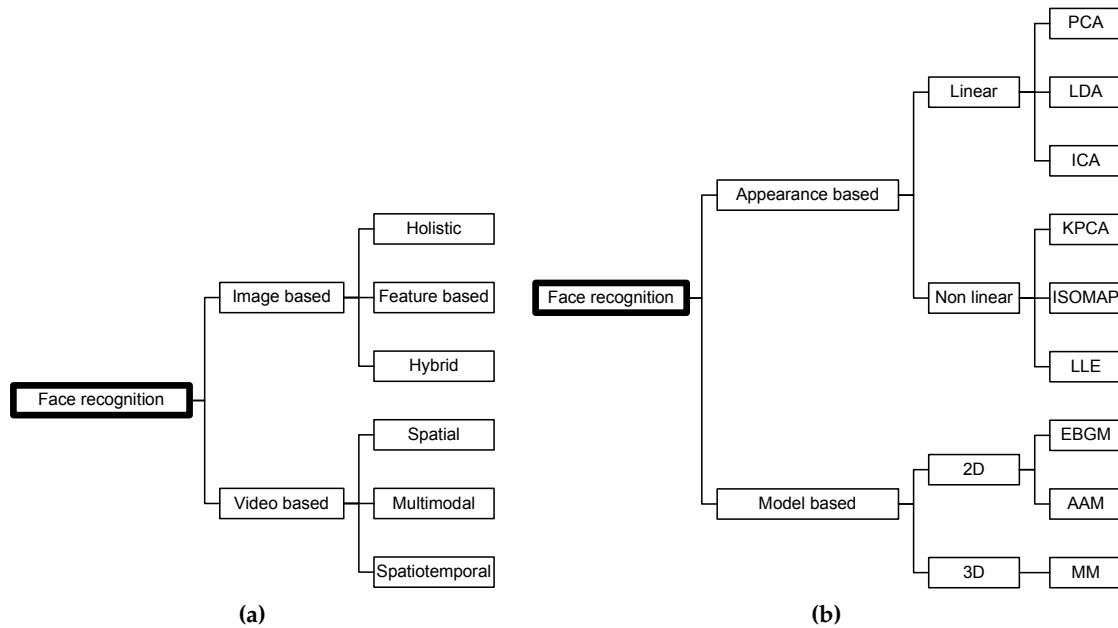
**Figure 7.3:** Large intra subject variations of faces in relation to a neutral face (a) due to illuminations (b), expression (c), and partial occlusions (d).

### 7.1.1 Related work

Since face recognition has been a very active research field for almost two decades, a large number of methods have been developed. Several surveys [Heseltine et al., 2003; Lu, 2003; Phillips et al., 2005; Zhao et al., 2000, 2003; Zhou and Chellappa, 2006] have been published, that review and compare existing approaches according to different criteria.

The most recent survey from Zhao et al. [2003] provides an in depth review of the face recognition domain until 2003. Existing approaches are grouped into image and video based approaches depending on the media type which is considered as shown in figure 7.4(a). *Image based approaches* are further subdivided based on the face representation into holistic, feature based and hybrid approaches. Considering the used information *video based approaches* can be subdivided into spatial, spatiotemporal and multimodal methods.

Lu [2003] focuses on image based approaches and categorizes them into appearance and model based approaches as shown in figure 7.4(b). *Appearance based* approaches are solely based on the image data and do not consider intermediate 3D models. They are further subdivided into linear approaches based on principal component analysis (PCA) [Turk and Pentland, 1991a], linear discriminant analysis (LDA) [Belhumeur et al., 1997], and independent component analysis (ICA) [Bartlett et al., 1998] and nonlinear approaches using kernel



**Figure 7.4:** Taxonomies of face recognition approaches according to (a) Zhao et al. [2003] and (b) Lu [2003].

principal component analysis (KPCA) [Schoelkopf et al., 1998], ISOMAP [Yang, 2002], and local linear embedding (LLE) [Roweis and Saul, 2000]. Model based approaches construct a face model that allows to capture the different variations. Prior knowledge about human faces as heavily utilized to design the model. Interesting approaches include elastic graph bunch matching (EBGM) [Wiskott et al., 1997], active appearance models (AAM) [Cootes et al., 2001] and 3D morphable models (3DMM) [Vetter and Blanz, 2003].

Following the taxonomy of Zhao et al. [2003] the different categories will be reviewed and some representative approaches for each them will be discussed below

### Image based approaches

As the name suggests these approaches are developed for single images but can be also applied to videos by considering frames independently. Many approaches have been proposed during the past decades, which makes a categorization based on used techniques rather difficult. Thus image based approaches can be categorized based on how faces are represented:

**Holistic approaches:** These approaches usually consider the whole face region for the recognition. *Correlation based methods* [Brunelli and Poggio, 1993; Nefian, 1996] are the simplest approach, where the matching takes directly place in the image space of the face as a bidirectional array of intensity values. These methods have some well known disadvantages, including the sensitivity to illumination changes, the large computational and storage costs. In order to decrease the feature dimension, several feature reduc-

tion techniques have been applied to face recognition. The so called eigenface approach by [Turk and Pentland, 1991a,b] applies the *principal component analysis (PCA)* to reduce the feature dimensionality while retaining most of the relevant information. Several extensions have been developed including selfeigenfaces [Torres et al., 2000], two dimensional PCA [Yang et al., 2004] and probabilistic subspaces [Moghaddam et al., 1998]. Bartlett and Sejnowski [1997]; Bartlett et al. [1998] have proposed two architectures based on the *independent component analysis (ICA)* by either finding statistically independent basis images or coefficients (factorial codes). Belhumeur et al. [1997] applied the *linear discriminant analysis (LDA)* and called their approach fisher-faces. After mapping the features into an intermediate subspace using PCA, classical LDA is applied to obtain the final subspace. Since the face manifold can not be entirely described using linear models, *kernel PCA (KPCA)* [Schoelkopf et al., 1999] has been applied for face recognition by Yang [2002]. This approach was extended by Kim et al. [2002] which combined the KPCA based feature extraction with a support vector machine (SVM). Since face recognition is essentially a supervised learning tasks, various techniques from this domain have been considered. Several neural network based (NN) approaches have been developed for face recognition. A neural network is an interconnected group of artificial neurons that uses a computational model to process information. Lin et al. [1997] use a probabilistic decision based neural network (PDBNN), which is an extension of a decision based neural network with a mixture of Gaussians as discriminant function. The system by Lawrence et al. [1997] is based on convolutional neural network (CNN). More recently *support vector machines (SVM)* have been used for face recognition [Guo et al., 2000]. The approach uses eigenfaces for the representation and linear SVM's for the classification. A face recognition approach based on *Bayesian decision theory* has been developed by Shakhnarovich and Moghaddam [2004]. It distinguishes between intra and extra personal variations and computes the a posteriori probability using the Bayes rule.

**Feature based approaches:** These approaches usually extract local features or components (e.g. eyes, nose, mouth) and their location, shape and appearance are fed into a structural classifier. *Hidden Markov models* have been applied for face recognition. Nefian and Hayes [1998] developed a framework for face recognition based on pseudo2D hidden Markov models (P2DHMM). Motivated by the natural order of facial components (hair, forehead, eyes, nose, mouth) the face is divided into vertically overlapping blocks. For each of the blocks lower dimensional features are extracted using Karhunen Loeve transform (KLT) and feed to a left right 1D continuous HMM. Wiskott et al. [1997] have developed a face recognition system based on *elastic graph bunch matching (EBGM)*. It is based on the dynamic link architecture (DLA) [Lades et al., 1993] which models an object as an adapted graph with nodes located at fiducial points. The individual nodes are described using a set of Gabor jets with different orientations and

scales. For each of the subjects an elastic bunch graph is built, that simultaneously describes the appearance and location of these fiducial points.

**Hybrid approaches:** Comparable to the human visual perception, these methods consider both global and local descriptions for the recognition. Pentland et al. [1994] extended the standard eigenface approach [Turk and Pentland, 1991a] to *modular eigenspaces*. The general idea is apply the original approach to several facial features (eyes, nose, mouth) and combine the scores by a cumulative sum. Penev and Atick [1996] proposed to use *local feature analysis (LFA)* for face recognition. LFA is a biologically inspired feature analysis technique that considers topographical information for feature reduction. A sparsification process selects the best topographic set based on the reconstruction error. An *active appearance model (AAM)* [Cootes et al., 1998] combines the statistical description of a shape and an appearance model. Matching the model to an image involves finding the model parameters that minimize the difference between the image and the synthesized model projected onto the image. Edwards et al. [1998] were the first that applied the AAM for face recognition. A single AAM is fitted to the training images of all persons. The resulting model parameters are projected into a discriminant subspace and a prototype for each person built.

### Video based approaches

More recently video based face recognition has emerged as a new research field [Zhao et al., 2003]. In comparison to image based face recognition it offers several advantages. Abundant image data can be used to select “good views” to perform recognition. The temporal continuity established by tracking can be used to compensate facial expression and pose changes. Multimodal informaton (face, speech, captions) can be used to increase the robustness of the recognition task. Nevertheless it brings along some inherent challenges. In general the quality (resolution, sharpness) of a video is much lower than that of images. This usually leads to face regions with smaller size and less details.

Existing approaches can be categorized depending on the type of information that is considered for the recognition:

**Spatial approaches:** Methods belonging to this category apply one of the classical image based approaches individually to each frame and combine the scores of multiple frames to improve the overall results. McKenna and Gong [1998] developed an access control system based on face recognition. Moving objects are detected and tracked based on clustering of spatio temporal zero crossings and Kalman filtering. Faces within these object are detected and tracked using radial basis function (RBF) networks. Face recognition based on principal component analysis (PCA) or linear discriminant analysis (LDA) is applied individually for each frame and the decisions for a tracked object are combined over time using a probabilistic voting scheme.

**Multimodal approaches:** Beside the visual modality these methods consider also other modalities (speech, captions) to obtain more reliable recognition results. A person authentication system for an ATM scenario was developed by Choudhury et al. [1999]. The recognition module consists of a speaker recognition and a face recognition module, which are combined by a fusion module. The face recognition module is based on the probabilistic eigenface approach [Moghaddam et al., 1998]. The speaker recognition module uses Mel frequency cepstrum coefficients (MFCC) and hidden Markov models (HMM). A Bayesian network is applied for combining the output of both recognition modules.

**Spatiotemporal approaches:** These methods exploit spatial (facial appearance) and temporal (facial motion) information simultaneously for face recognition. Zhou et al. [2003] proposed a system that considers these spatio-temporal characteristics for the recognition. While most other approaches deal with the uncertainties in face tracking and recognition separately, their system performs both tasks simultaneously. A time series state space model characterizes both the motion and the identity using a motion vector and an identity label. The joint posteriori probability is estimated at each time instant using a sequential importance sampling and propagated to the next time instant.

### 7.1.2 Challenges

While most of the face recognition approaches discussed above work quite well in well controlled scenarios, face recognition in uncontrolled scenarios is still difficult [Zhao et al., 2003; Zhou and Chellappa, 2006] due to various challenges. They will be discussed below along with methods to cope with them:

**Illuminations:** The illumination problem is caused by different lighting (direction, intensity, color) which changes the facial appearance quite considerably. The difference induced by the illumination is often larger than the difference between individuals, leading to wrong recognition results [Adini et al., 1997]. Existing approaches to handle the illumination problem can be grouped into 4 categories [Zhao et al., 1999]. *Heuristic approaches* such as simple contrast normalization [Moghaddam and Pentland, 1997], histogram equalization [Sung and Poggio, 1998] and symmetric shape from shading (SSFS) [Zhao et al., 1998] rely on heuristic methods to compensate for illumination changes. *Image comparison approaches* [Adini et al., 1997; Jacobs et al., 1998] consider different image representations (edges, derivatives, filters) and distance measures to increase the robustness regarding illuminations. *Class based approaches* try to model appearance changes due to different illuminations within the training stage, which requires a representative set of training samples covering different illumination conditions. Recent approaches include 3D linear subspaces [Shashua, 1997], illumination cone [Belhumeur and Kriegman, 1998] and spherical harmonics [Basri and Jacobs,



2003]. In *model based approaches* a 3D face model is used to synthesize the virtual image from a given image under the desired illumination. 3D morphable models [Blanz et al., 2002] are one of the methods that follow this approach.

**Pose:** It is not surprising that the performance of face recognition systems drops significantly in the presence of large pose variations. While in plane rotations (roll) can be easily handled, out of plane rotations (pan, tilt) are a big challenge due to the 3D structure of the face. Existing approaches can be divided into 3 groups [Zhao et al., 1999]. *Multi view approaches* such as the template based correlation matching [Beymer, 1993] and the multi view illumination cone method [Georghiades et al., 2001] require multiple view images from each person during training and testing. *Hybrid approaches* consider multiple view images during the training but operate on single view images for testing. Well known approaches include multi view eigenfaces [Moghaddam and Pentland, 1994], eigenlight fields [Gross et al., 2002] and partial PCA/LDA [Rama and Tarres, 2006]. Finally, *single view approaches* require only a single image during training and testing and they are usually based on either view invariant features [Manjunath et al., 1992] or frontal view synthesis based on 3D models [Zhao and Chellappa, 2000]. While hybrid approaches have been the most popular up to now, single view approaches have not received much attention.

**Occlusions:** While a lot of methods have been proposed for handling different illuminations and views, only a few methods for handling occlusions have been developed. Nevertheless facial occlusions present a major challenge within uncooperative scenarios such as surveillance or multimedia retrieval. In contrast to illumination and pose it is rather difficult to model occlusions reliably due to their large variety. Existing approaches can be grouped depending on the representation of the face which is used to improve the reliability in the presence of occlusions. One straightforward strategy is to use a *holistic approach* but train the system with images of occluded faces. Instead of using any knowledge regarding present occlusions these approaches rely on a representative training set to handle occlusions reliably. The major problem with this strategy is that occlusions can be very different both in appearance and location, which makes it impossible to train them. Another option is to consider a *component based approach* that divides the face into several parts and considers only the parts which are not occluded. Anyway this requires additional information about the presence and location of occlusions. Modular eigenspaces [Pentland et al., 1994] are one of the potential methods, that can be adapted to consider occlusion information. [Martinez, 2002] have developed a similar approach where the face is divided into several local parts that do not exactly correspond to facial components. Each of these components is projected to an individual eigenspace and modeled by a mixture of Gaussians. Occlusions are detected by analyzing the distance with the corresponding eigenspace. Finally, the last strategy for handling partially occluded faces is to use a *near holistic*

*approach* which can be seen as a tradeoff between holistic and component based approaches. The major idea is to decrease the importance of occluded parts and consider only the rest of the face for the recognition. Belonging to this category, the lophoscopic PCA (LPCA) [Rama and Tarres, 2005] is a natural extension of the eigenface approach, where several models are created through masking occluded parts. Therefore instead of having only a single eigenspace an individual eigenspace for each type of occlusion is built. The local feature analysis (LFA) [Penev and Atick, 1996] can be also considered within this group. It combines the PCA with the analysis of local information around some critical points of the face.

### 7.1.3 Objective

The goal of this work is to develop a generic approach for face description and recognition that can be applied in various application scenarios. Therefore, it needs to cope with the most prominent challenges within these application domains, which are illuminations and partial occlusions. Furthermore, the approach should be suitable for several face recognition tasks, including matching, clustering, and recognition.

This work has been developed together with Toni Rama<sup>1</sup> (UPC) who contributed the lophoscopic PCA part and was involved in the experiments. Parts of this work have been published in FG 2008 [Rama et al., 2008].

## 7.2 Approach

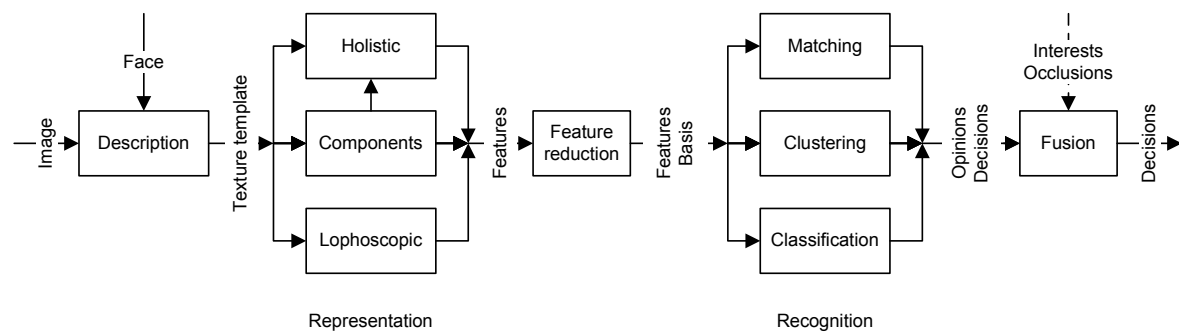
In order to reach the objective an appearance based approach was chosen for the description and recognition of faces. Figure 7.5 provides an overview of the proposed method that consists of 4 major steps. The description step extracts a texture template that is invariant to translation, scaling and in plane rotations as well as variations in illumination strength and direction. Based on this template the next step extracts representations of different complexity and adaptability. Since the dimensionality of the resulting feature vectors is usually quite high, feature reduction techniques are used to reduce the number of dimensions while retaining the most important information. The resulting face descriptions are then used within the recognition step to match faces to each other, cluster faces based on similarity, or predict the identity of a person based on the face. Depending on the representation multiple experts are used to represent a face that can be fused to improve the overall recognition performance. The following sections describe the individual steps in more detail.

### 7.2.1 Description

Appearance based face recognition methods typically rely on *texture templates*, which are rectangular image patches of predefined size. In order to make it robust regarding pose (in

---

<sup>1</sup>tonirama@gps.tsc.upc.edu



**Figure 7.5:** Overview of the face recognition system. Given an image and the location of the face provided by the face detection module, the face recognition starts by describing the face with a texture template from which different representations are extracted. Since the resulting feature dimensionality is usually quite high feature reduction techniques are applied. Depending on the application, different recognition tasks are supported, including matching, clustering and classification. Finally, the individual experts for each of the representations are fused to improve the performance. Within this step additional occlusion information can be considered for selecting a set of reliable experts.

plane rotation) and illumination variations (strength and direction) two steps are required, namely geometric transformation and intensity normalization.

### Geometric transformation

In order to extract a texture template from the face region, it is transformed using a *2D similarity transformation*, that includes scaling, translation and rotation (see section 3.2.2 for more details). To compute the 4 parameters of the corresponding transformation matrix at least 2 point correspondences are needed. Usually only the pupils are considered for the transformation which is problematic if eyes are occluded and the position of them cannot be reliably estimated. In order to handle these cases and make the transformation more reliable, 2 other points (nose tip, mouth center) are additionally considered. If all points are visible a least square solution can be found for the transformation matrix which is more robust to localization errors of the components. In the case of occluded feature points only the visible points are considered for the normalization. Figure 7.6 illustrates the geometric transformation by showing the face before and after the process.

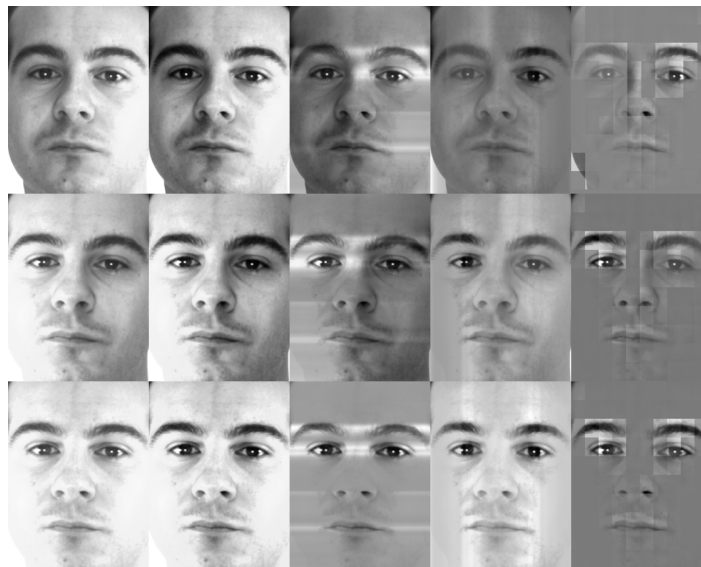
### Intensity normalization

The goal of this step is to make the texture template invariant to illumination changes. This typically includes the strength, the direction and the color of the light source. Since the texture template considers only the intensity, different colored light sources will not have a large influence on it. Thus no color normalization is applied here.

Beside other things, the contrast of an image is largely affected by the strength of the light. In order to make the texture template invariant to this image enhancement techniques (see section 3.2.1) can be applied. Two well known techniques for image normalization are histogram stretching and histogram equalization. While the former technique applies



**Figure 7.6:** Illustration of the texture template extraction by applying a 2D similarity transformation based on four feature points: (a) input image, and (b) output template.

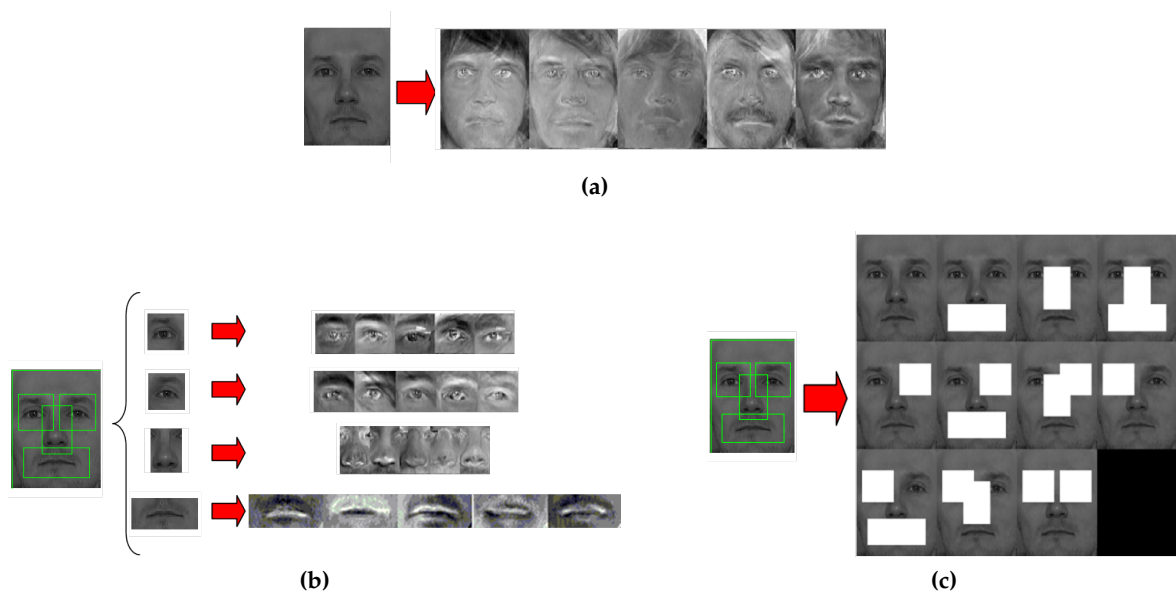


**Figure 7.7:** Illumination compensation methods for different illumination directions. From top to bottom: left side, right side, frontal illumination. From left to right: original image, global illumination compensation and local illumination compensation in horizontal, vertical and block-wise manner.

a linear mapping function to the image pixels to span the full range of intensity values, the second technique applies nonlinear mapping functions to give the intensity histogram a uniform shape. Since the later may suppress important face information, contrast stretching is applied globally to compensate for varying light strength which can be seen in the first and the second column of figure 7.7.

Apart from the intensity of the light, the appearance of a face can vary largely with the direction of the light. This is caused by the 3D structure of the face where depending on the light direction certain parts cast shadows onto other parts. While this effect is minimal for frontal or diffuse light, it may be quite large for light from lateral directions as it can be seen in the first column of figure 7.7.

Since global contrast stretching does not compensate for different illumination direc-



**Figure 7.8:** Illustration of the different face representations: (a) holistic approach with a single expert, (b) component based approach with 4 experts, and (c) lophoscopic approach with 11 experts.

tions, local contrast stretching is applied. This can be done by dividing the whole face region into subregions and applying the contrast stretching independently. The division can be done horizontally, vertically or block-wise. While the two former are optimal for special light directions, block based processing allows to compensate very diverse illumination directions as it can be seen in the last column of figure 7.7.

### 7.2.2 Representation

Given the extracted face description in form of the texture template, several face representations with different focus and complexity can be created. Following the goal of providing a set of flexible and robust representations, the 3 approaches illustrated in figure 7.8 have been considered.

The representations are sets of subregions of the overall face that differ in the number and definition of considered parts (experts). All of them are based on a combination of several face regions defined with respect to the overall face region in table 4.1. Given these definitions, table 7.1 compares the representations based on several criteria including number of experts, overall feature dimensionality, correlation between experts and fusion possibility.

#### Holistic

The holistic approach (figure 7.8(a)) represents the face as a whole with a single expert. Thus it considers all parts of the face as equally important independent of the application or environmental conditions. It is less flexible than the two other representations but also less

Approach	Experts	Dimens.	Corr.	Fusion
Holistic	1	5625	No	No
Components	4	4175	No	Yes
Lophoscopic	11	45175	Yes	Yes

**Table 7.1:** Overview of the different representations with interesting characteristics including number of experts, dimensionality, correlation and fusion possibility.

complex. Theoretically the whole physiognomical face region can be used for the representation. In practice is it usually better to consider a smaller subregion and discard irrelevant parts as it can be seen in figure 7.8(a).

### Components

The component based approach (figure 7.8(b)) represents the face by 4 experts that correspond to individual facial components (left eye, right eye, nose, mouth). Thus it considers only the most important structural parts of a face and discards others. The low spatial overlap of the 4 components leads to 4 uncorrelated experts, that can be combined quite flexibly depending on the application or environmental conditions.

### Lophoscopic

The lophoscopic approach (figure 7.8(c)) can be seen as the inverse of the component based approach, where the face is represented by a set of experts corresponding to the different combinations of masking a maximum of 2 components within the whole face. Each of the resulting 11 experts than discards different parts of the face as unimportant and considers the rest. Since there is overlap between the different experts they are largely correlated. Depending on the application or environmental conditions they can be combined or a certain expert can be selected.

### 7.2.3 Reduction

One of the major problems of any appearance based face recognition approach is that the extracted texture templates usually consist of a large number of pixels. This problem is illustrated in table 7.1 which shows the overall dimensionality of the different representations. The combination of this large number of dimensions and the usually low number of samples, may lead to the curse of dimensionality. Thus it is necessary to reduce the number of dimensions before the recognition step. Several feature reduction methods (see section 3.3.2) have been applied to face recognition, including principal component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA). Within the current system PCA was chosen since it has moderate complexity and does not require any class information which might not be available in certain applications.

Given a 2D texture template there are basically two ways how the PCA can be applied to the data. The standard way is to treat the texture template as single feature vector of size  $s = w \times h$  by scanning it row or column wise. This is also referred to as *1D PCA* [Turk and Pentland, 1991a]. Another way, often called *2D PCA* [Yang et al., 2004] is to treat either the rows or the columns of the texture template as feature vectors and apply the PCA only to one dimension. Although this leads to a more reliable estimate of the principal components due to the reduced feature dimensionality it also discards spatial information in the other dimension. Only the 1D PCA is considered within this work.

### 7.2.4 Recognition

As already discussed in section 4 recognition in this work may refer to several related tasks including matching, clustering and classification. Thereby this work is restricted to the identity of a person instead of gender and ethnicity.

*Face matching* refers to measuring the dissimilarity between two face descriptions. Depending on the representation a description may consist of several experts that are compared to each other and fused together. Although any of the distance measures described in section 3.3.1 may be used for this task the Euclidean distance is the most commonly used

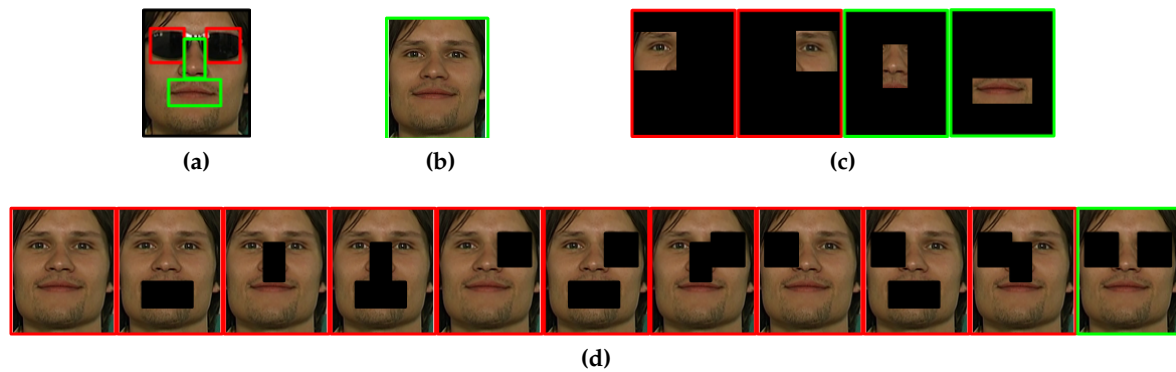
*Face clustering* describes the process of grouping similar face descriptions together. This can be used to automatically determine the number of distinct identities from a set of images or videos. Any of the clustering methods described in section 3.3.4 is suitable for this task and the optimal choice depends largely on the task and the data.

*Face classification* usually refers to predicting the identity of a face based on several trained models and a classification rule. Any of the classification approaches described in section 3.3.5 can be applied.

### 7.2.5 Fusion

Depending on the representation only a single (holistic) or multiple experts (components, lophoscopic) are used to describe a face. The output of multiple experts can be combined by different postmapping fusion methods (see section 3.4.2) to improve the performance of the face recognition task. Both opinion (max, min and product rule) and decision level (majority voting) fusion are considered here.

Without any additional information all experts for the component based approach (4) and the lophoscopic approach (11) are fused equally. Nevertheless, this might not be the best approach in the presence of partial occlusions or for certain application, since some experts may be not as reliable or important as others. Therefore additional knowledge such as occlusion information from the component based face detection approach (see chapter 6) may be used to improve the reliability of the face recognition. It is used to select a reliable subset from the set of available experts. For the *component based representation* which considers 4 experts representing the facial components, the straightforward way is to take only the



**Figure 7.9:** Multiple expert fusion for the different face representations with and without a priori information: (a) input image with occluded (red) and non-occluded (green) components, (b) holistic representation without fusion, (c) component based representation with fused (green) and discarded experts (red), and (c) lophoscopic representation with selected (green) and discarded (red) experts.

experts corresponding to non occluded components. For the *lophoscopic representation* only the expert corresponding to the present occlusion situation is selected. Figure 7.9 illustrates that idea for a sample where both eyes are occluded.

## 7.3 Experiments

The goal of the conducted experiments is to assess the performance of the developed face recognition approach and the different representations in the presence of partial occlusions. Furthermore, different fusion methods for combining multiple experts of a single representation are compared to each other. Finally, the adaptive fusion based on occlusion information is explored.

### 7.3.1 Dataset

Face recognition has received a lot of attention for the last decades. Along with the large number of methods, several databases for the development and evaluation have been created.

A comprehensive review of publicly available databases is given by Gross [2005]. Table 7.2 provides a comparison of considered databases based on several criteria including number of images, subjects, poses, illuminations, expressions, occlusions, and sessions. While most of the available databases contain a large variety of poses, illuminations, and expressions only a very small number of realistic occlusions is available. Thus only two databases (upper part of the table) have been used in the experiments.



Title	Im.	Subj.	Pos.	Ill.	Expr.	Occl.	Sess.
AR Face Database	3536	136	1	3	3	2	2
VISNET II Face Database	4070	37	1	1	3	8	1
Yale Face Database	165	15	1	3	5	1	1
CMU PIE Database	41368	68	13	43	3	0	1
UPC Face Database	1188	44	5	3	3	3	1

**Table 7.2:** Comparison of selected databases suitable for face recognition based on number of images, subjects, poses, illumination, expressions, occlusions, and sessions grouped into used (top) and discarded (bottom) databases.



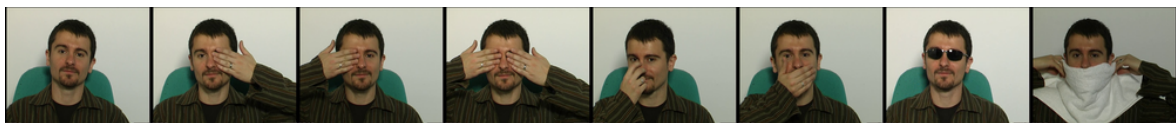
**Figure 7.10:** Subset of the AR Face Database with occlusions. From left to right the variations are 1,8,11,14,21,24.

### AR Face Database

The AR Face Database (described in section A.2.3) contains 3536 images of 136 (76 male, 60 female) subjects with 22 variations (illumination, expression, occlusion, time). For the experiments only a subset of the database has been used. To study solely the influence of occlusions, only the variations with frontal illumination and neutral expression (1,8,11,14,21,24) have been considered as shown in figure 7.10 for a single subject. Furthermore, the subset has been limited to 100 subjects (50 male, 50 female).

### VISNET II Face Database

The VISNET II Face Database (described in section A.2.5) contains 4070 images of 37 subjects (34 male, 3 female) with 11 variations (expressions, occlusions). For the experiments only a subset of the database with the variations (1,2,4,6,8,9,10,11) have been used as shown in figure 7.11 for a single subject.



**Figure 7.11:** Subset of the VISNET II Face Database with real occlusions. From left to right the variations are 1,2,4,6,8,9,10,11.

Version	Approach	Experts	Fusion	AR	VISNET
1	HPCA	Single	None	0,38	0,70
2	CPCA	All	Majority	0,52	0,64
3	CPCA	All	Min	0,32	0,61
4	CPCA	All	Max	0,48	0,46
5	CPCA	All	Sum	0,46	0,66
6	CPCA	All	Product	0,41	0,69
7	CPCA	Multiple	Majority	0,59	0,70
8	CPCA	Multiple	Min	0,44	0,69
9	CPCA	Multiple	Max	0,70	0,74
10	CPCA	Multiple	Sum	0,72	0,79
11	CPCA	Multiple	Product	0,67	0,80
12	LPCA	All	Majority	0,39	0,68
13	LPCA	All	Min	0,38	0,67
14	LPCA	All	Max	0,41	0,72
15	LPCA	All	Sum	0,40	0,69
16	LPCA	All	Product	0,39	0,68
17	LPCA	Single	None	0,47	0,74

**Table 7.3:** Face recognition performance of the versions over the different databases. From top to bottom: HPCA (1), CPCA without (2-6) and with (7-11) occlusion information, LPCA without (12-16) and with (17) occlusion information.

### 7.3.2 Evaluation

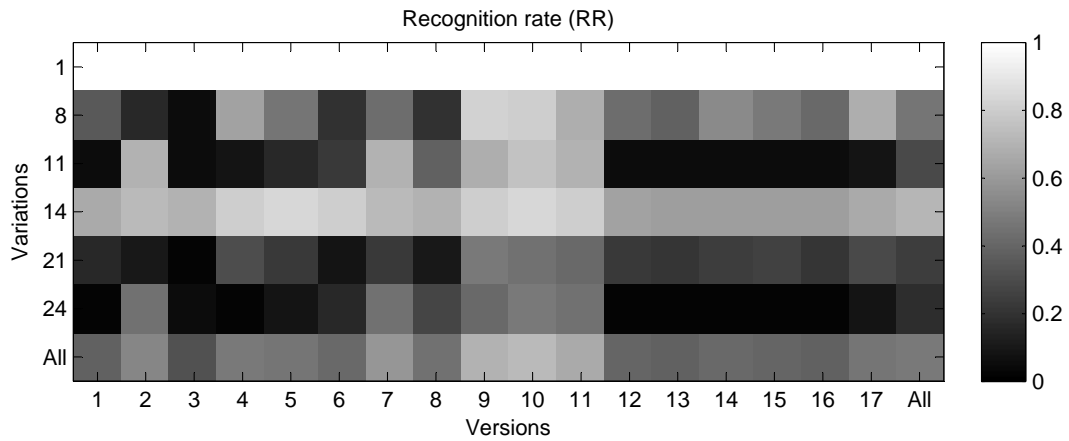
While the different recognition tasks (matching, clustering, classification) are usually evaluated in different ways, their relative performance is usually comparable. The same applies to the different classification scenarios such as verification, identification, and watchlist.

Within these experiments face recognition has been treated as an identification task and evaluated as a recognition problem (see section B.4). Based on the ground truth and the predicted labels for the given dataset a n-ary confusion matrix is built, from which the recognition rate (RR) can be computed.

### 7.3.3 Results

The goal of this set of experiments is to assess the performance of the different representations (holistic, components, lophoscopic) and the different fusion strategies (max, min, sum, product) in the presence of partial occlusions. Furthermore, it is analyzed how the performance improves if additional occlusion information is considered within the fusion step.

Based on this 17 versions of the appearance based face recognition approach have been compared to each other as shown in table 7.3. It summarizes the experiments by providing the overall performance of the versions for the different datasets. The best recognition rates for HPCA, CPCA, LPCA without additional information are 38%, 52%, 41% for the AR Face

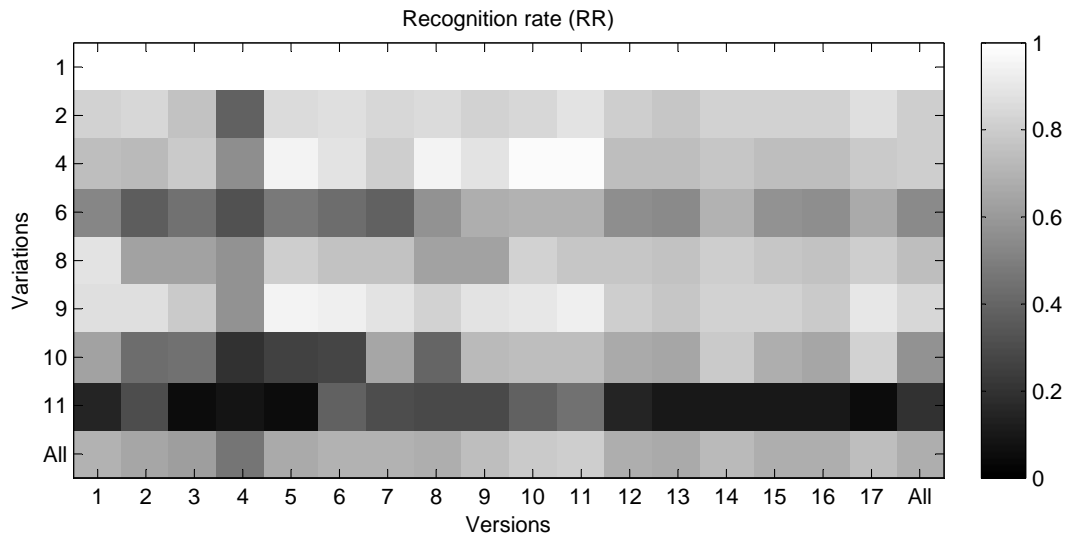


**Figure 7.12:** Face recognition performance of the different versions over the variations of the AR Face Database. The last row and column provide the average over the variations and approaches, respectively.

Database and 70%, 69%, 72% for the VISNET II Face Database respectively. With the use of occlusion information the recognition rates of CPCA, LPCA improve to 72%, 47% for the AR Face Database and 80%, 74% for the VISNET II Face Database. It is important to note that these performance measures include occlusions and non occlusions. If only occlusions are considered the improvement is even higher. In the following a more detailed analysis of this results for the different occlusion types is presented.

The first experiments is based on the subset of the AR Face Database. Each version is trained with a single image per person without any occlusion (variation 1) and tested on all other variations (1,8,11,14,21,24). Figure 7.10 shows the performance of each version over the different variations as a pseudo color plot. Furthermore, the average performance of each version over the different variations (last row) and the average performance of each variation over all the versions (last column) are also provided. By visual inspection several observations can be made. First of all the performance over the different variations is generally lower for occlusions (8,11,21,24) than for non occlusions (1,14). This is caused by the lower amount of discriminative information which makes subjects harder to distinguish. The performance for session 1 (1,8,11) is generally better than for session 2 (14,21,24) since samples from session 1 are used for training and the face appearance changes considerably between the two sessions. Concerning the different versions HPCA is outperformed by both CPCA (2-11) and LPCA (12-17). It is interesting to see that LPCA performs worse for the scarf (11,24) than for the sun glasses (8,21) while CPCA achieves similar performance for both. This is caused by the fact that the scarf covers more than the mouth region of the face. While the performance of the different fusion methods is quite similar for the LPCA (12-16) it varies considerably for the CPCA (2-6). Considering the occlusion information within the fusion step improves the results for both LPCA and CPCA. The individual approaches can be ranked in the following order: CPCA (10), LPCA (17) and HPCA(1).

The second experiment uses the VISNET II Face Database. Each version is trained with



**Figure 7.13:** Face recognition performance of the different versions over the variations of the VISNET II Face Database. The last row and column provide the average over the variations and approaches, respectively.

1 image per person without any occlusion (variation 1) and tested on 4 images for each variation (1, 2, 4, 6, 8, 9, 10, 11). Figure 7.13 shows the performance for each of the versions over the different variations in a similar way as before. In general the performance is better than for the AR Face Database, since the number of persons is considerably lower. Furthermore, it is evident that the performance depends largely on the variation. As expected the best performance is achieved for faces without any occlusions (1). For occlusions of one components (2, 4, 8, 9) the performance is generally higher than for occlusions of two components (6, 10, 11). The largest performance drop is caused by the scarf (11), since it covers a large portion of the face. Again the occlusion information improves the results for both LPCA and CPCA. The overall ranking of the different versions is the same as for the AR Face Database.

## 7.4 Conclusion

### 7.4.1 Summary

An occlusion aware face recognition approach has been developed that considers additional information regarding the presence and location of partial occlusions. It relies on the component based occlusion information from the proposed face detection approach (described in section 6) and selects reliable parts of the face for the recognition. Therefore, an appearance based face recognition approach has been chosen and different face representations (holistic, components, lophoscopic) are considered. While the developed method relies on principal component analysis (PCA) and thus can be seen as an extension to the well known eigen-face approach it is generic and not limited to this feature reduction technique. The PCA was chosen to develop a general purpose face recognition approach, that does not require class

information which may not be available for some tasks such as matching or retrieval. Depending on the representation multiple experts are used to describe a face which are fused together using postmapping fusion. Based on the occlusion information the most reliable experts are selected from the overall set, which is comparable to the human visual perception that focuses on visible parts of the face.

Extensive experiments have been carried out to assess the performance of the developed face recognition system in the presence of partial occlusions. The different representations have been compared to each other and different post mapping fusion methods have been explored with and without the use of additional occlusion information. The experiments show that the best performance can be achieved by the component based followed by the lophoscopic and the holistic representation. Selecting reliable experts based on the occlusion information improves the performance considerably for both the component based and the lophoscopic representation.

#### 7.4.2 Future work

Although a considerable performance improvement can be achieved in the proposed system by considering occlusion information for face recognition, several directions for future work are remaining.

As an appearance based approach the proposed method relies on texture templates which are extracted from the image by applying a 2D transformation to the detected face region. In order to handle varying illumination contrast stretching is applied globally and locally. Depending on the conditions other preprocessing methods may be considered [Heseltine et al., 2002] including color normalization, statistical and filtering methods.

Using principal component analysis (PCA) for feature reduction has been largely motivated by the fact that it is well understood technique for feature reduction and does not require any class information which may not be available for some tasks, e.g. matching and clustering. Nevertheless, other techniques, e.g. linear discriminant analysis (LDA) [Belhumeur et al., 1997] and independent component analysis (ICA) [Baek et al., 2002], have been applied to face recognition and shown superior performance with respect to PCA. Due to the generic nature of the approach, these techniques can be easily incorporated.

Since the conducted experiments use only a single training sample for each person, a minimum distance to means (MDM) classifier is used which will not be the best method if multiple training samples are available. In that case, other classification approaches (described in section 3.3.5) such as Bayesian classifier (BC) or support vector machine (SVM) may lead to improved performance.

Currently fusing multiple experts is achieved by unweighted score combination. Since the reliability of the experts may differ even for non occluded faces weighted score combination may be considered to improve the results. The corresponding weights can be derived from the recognition performance of the individual experts. The same applies to the selec-

tion of experts based on the occlusion information which in the current method can be seen as an associative switch [Xu et al., 1994]. Depending on the amount of overlap the weight of each expert may be adapted.

## Chapter 8

# Multimodal person search

### 8.1 Introduction

#### 8.1.1 Motivation

With the increasing amount of available multimedia data, efficient systems for searching and retrieving audiovisual (AV) documents are required. Traditional systems based on keywords are quite inefficient due to the time consuming annotation as well as linguistic and semantic ambiguities. Therefore, content based multimedia systems have been proposed that search and retrieve AV documents based on audio and visual features extracted from the content itself.

While content based multimedia retrieval has been very active research topic, less work has been done in the field of person search and retrieval, where the goal is to find audiovisual (AV) documents or parts of it where a specific person is present in the audio or video stream. This is especially interesting since persons are one of the most relevant objects within multimedia data. The general idea is the following: given a large set of audiovisual (AV) documents (e.g. from YouTube<sup>1</sup>) containing individuals giving talks, the goal is to find and retrieve the clips of a specific person based on a sample provided to the search engine. Typical application scenarios of such a system are illustrated in figure 8.1 including official video podcasts, personal video blogs and broadcast news. For most of these scenarios it can be assumed that the voice present in the audio stream and the face present in the video stream belong to the same person.

#### 8.1.2 Related work

As already mentioned little work has been done in the specific field of multimodal person search and retrieval. Nevertheless, related work can be found in two major areas: *content based multimedia retrieval* (see section 2.4) and *multimodal biometrics* (see section 2.3). The former deals with the search of multimedia documents without the emphasis on a certain

---

<sup>1</sup><http://www.youtube.com/>



**Figure 8.1:** Application scenarios for multimodal person search and retrieval: (a) official video podcast (b) personal video blog (c) broadcast news.

object class [Lew et al., 2006]. The latter focuses on the identification of persons based on different biometric traits such as face, gait, voice and fingerprints [Bowyer et al., 2006]. Some approaches, e.g. Image Search (Google)<sup>2</sup> and Google Portrait (IDIAP)<sup>3</sup> find specific persons within images through a combination of keyword based search and face detection. Another approach taken by Riya<sup>4</sup> combines user tagging and visual analysis of images to support the search and retrieval of individuals.

### 8.1.3 Objective

In contrast to these approaches, the goal of this work is to develop a system that combines audiovisual analysis of humans with content based multimedia retrieval techniques for efficient search and retrieval of audiovisual content based on the present humans. Another objective is to explore the performance and limits of such a system with respect to the different modalities and retrieval paradigms.

This work has been developed together with Amjad Samour<sup>5</sup> (TUB) who contributed the audio analysis part, was involved in the development of the overall system and the conducted experiments. A preliminary version of this work has won the best poster award at KSPJ 2007 [Samour et al., 2007a] and was published at SAMT 2007 [Goldmann et al., 2007b].

## 8.2 System overview

Figure 8.2 provides an overview of the proposed system for multimodal person search and retrieval. It consists of an online and an offline parts. For a given database the *offline* part splits the multimodal data into an audio and a video stream, runs the audio analysis and the video analysis independently and stores the extracted information in a database. The

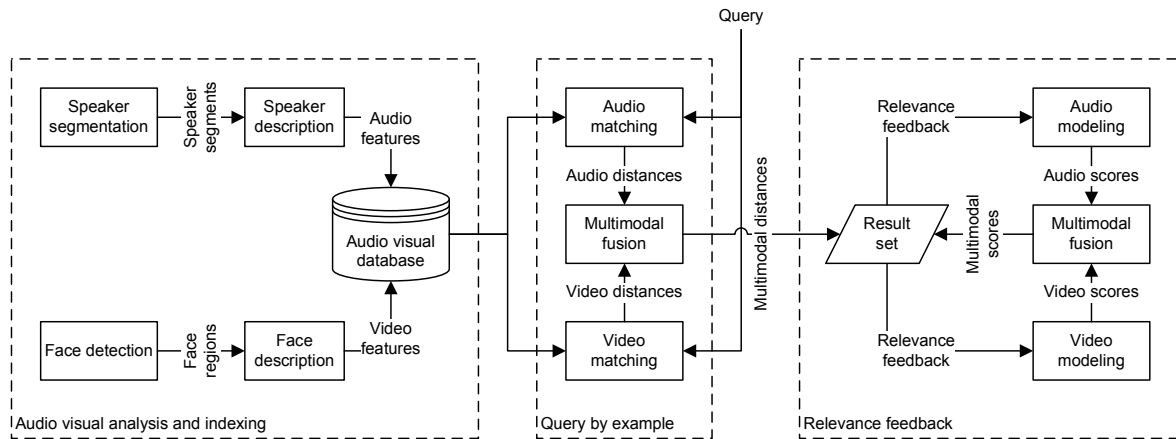
<sup>2</sup><http://images.google.com/>

<sup>3</sup><http://www.idiap.ch/googleportrait>

<sup>4</sup><http://www.riya.com/>

<sup>5</sup>[samour@nue.tu-berlin.de](mailto:samour@nue.tu-berlin.de)





**Figure 8.2:** Overview of the system for multimodal person search and retrieval. Within the offline phase an audiovisual description of the humans present within the videos is extracted. The on-line phase starts with an audiovisual example which is compared to the audiovisual models in the databases. Relevance feedback is used to iteratively refine the search result.

*online* part itself consists of two steps, a query by example step to start the search process by providing a sample and a relevance feedback loop to refine the search based on the feedback provided by the user. Again the audio and video information is treated individually during the matching and modeling, but finally combined within the multimodal fusion step.

The current system provides a simple web based user interface, shown in figure 8.3, to select the query sample, display the current result set, and ask the user to provide feedback. To support the analysis of the different modalities (audio, video, multimodal) the corresponding result sets of the current iteration are provided along the individual columns. For the evaluation the relevant and irrelevant items for this query are highlighted in green and red, respectively.


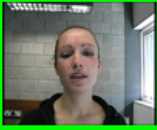
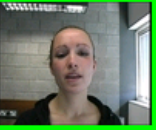

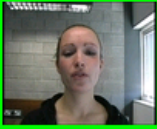
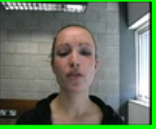


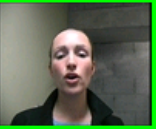





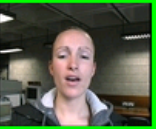
### 8.2.1 Audio analysis

The goal of the audio analysis part is to retrieve audio segments based on the voice characteristics of a person independent of the spoken content. It consists of an optional speaker segmentation step which segments the speech stream into individual speaker segments and a speaker description step that extract suitable audio features to describe each speakers voice.

#### Speaker segmentation

The goal of the speaker segmentation step is to divide the audio stream into temporal segments corresponding to individual speakers. Therefore change points between different speakers are detected with a metric based segmentation approach similar to the one proposed by Delacourt and Welekens [2000].

The audio stream is divided into frames of 40 ms duration for which well known Mel

Rank	Audio	Score	Video	Score	Multimodal	Score
1		-1238.280		0.000		0.000
2		-646.054		5.328		0.000
3		-484.800		5.527		0.000
4		-426.157		5.671		0.000
5		-417.134		6.312		0.031

**Figure 8.3:** Web based user interface (WUI) of the multimodal person search application. The ranking of the individual modalities are shown along the columns, with the relevant items (green) and the irrelevant items (red).

frequency cepstrum coefficients (MFCC) [Stevens, 1957] are computed. Given this sequence of audio features the Bayesian information criterion (BIC) is used to determine speaker changes. This is achieved by moving a sliding window over the stream and considering the corresponding features vectors as a single (whole window) or two individual Gaussian processes (two half windows). Then the decision if the window contains a change point or not can be interpreted as a model selection problem based on the BIC value. The temporal segment between two change points is then considered to belong to the same speaker.

### Speaker description

The goal of the speaker description step is to extract a robust description of the speakers voice characteristics independent of the spoken content and environmental conditions. Again, well known MFCCs have been adopted since they provide a compact representation of the spectral characteristics of an audio signal that resembles the human auditory system.

Given a speaker segment MFCCs are extracted in the same way as in the speaker segmentation by dividing the audio stream into frames of 40 ms length, applying a Hamming window and computing the power spectrum. The power spectrum is transformed to the Mel-scale by applying a set of triangular filters. Finally, the discrete cosine transform (DCT) is applied to compute the cepstral coefficients from the Mel-spectrum leading to a feature

vector of size 13 for each window. In order to reduce the temporal characteristics of the spoken content within a segment and to create a robust model of the spectral characteristics of the speaker's voice, each speaker segment is described by the arithmetic mean computed over all the windows.

### 8.2.2 Video analysis

The goal of the video analysis part is to detect and describe frontal faces in the visual stream in a large variety of environments. It consists of a face detection step which detects and localizes visible faces and a face description step that extracts a robust description of the face which is used for the matching later.

#### Face detection

The face detection is based on the component based approach described in more detail in section 6. It has been shown that this approach can not only detect partially occluded faces, but also detect the location of these occlusions. This additional information may be used to select non-occluded samples of the persons face to increase the robustness of the retrieval process.

#### Face description

The face description is based on the holistic approach described in more detail in section 7 which is an extension of the well known *eigenface* method proposed by Turk and Pentland [1991a]. The module developed within chapter 7 has been slightly adapted by considering a smaller template size ( $30 \times 40$  instead of  $75 \times 100$ ). This was motivated by the lower face resolution in the VALID Database (see section A) in comparison to the previously used databases.

### 8.2.3 Query by example (QBE)

The idea behind the query by example (QBE) paradigm (see section 2.4) for retrieval is that a user is asked to provide a sample that represents his search intention. This sample is analyzed in the same way as all the samples in the database and compared to them based on some criteria.

In the current system each sample is represented by two feature vectors. For each modality (audio, video) the distances between the corresponding sample and the documents in the database are computed. Several metrics have been considered for the matching (see section 3.3.1). Initial experiments showed that the *euclidean distance* provides the best results for the this task. The individual distance sets are combined in the multimodal fusion step.

### 8.2.4 Relevance feedback (RF)

The idea behind relevance feedback (RF) (see section 2.4) is to bridge the semantic gap and improve the retrieval results by integrating the user into the retrieval process to obtain a better estimate of the search intention. RF approaches can be categorized according to several criteria [Crucianu et al., 2004]. Out of the possible *time periods* only the current session (current and the previous rounds) are considered. From the possible *sources* only the information of the current user is considered. The relevance feedback process itself typically consists of a learning and a selection step. Within the *learning* step a model of the user's search intention is built based on the provided feedback. From the machine learning approaches described in section 3.3 the single Gaussian model and the support vector machine have been selected, since they naturally support different types (positive and positive/negative) of feedback. After the learning step, a *selection* step determines the items which are returned in the result set. Selecting the most positive items returns the best matches regarding the user's search intention which may not be the best strategy for intermediate iterations. On the other hand, selecting the most informative items tries to reduce the ambiguities by obtaining feedback for critical items. In order to allow the user to stop the iterative process at any time when he is satisfied with the retrieval results, only the former selection criteria is used.

**Positive feedback** The first relevance feedback strategy considers only positive feedback, i.e. relevant items to estimate the user's search intention [Su et al., 2003] (see section 3.3.3 for more details on density estimation). Therefore the relevant items are assumed to follow a *single Gaussian* (SG) distribution

$$p(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1}(\vec{x} - \vec{\mu})\right) \quad (8.1)$$

with the feature dimensionality  $D$ , the mean vector  $\vec{\mu}$ , and the covariance matrix  $\Sigma$ . These parameters are estimated from the positive samples provided by the user. The matching is performed by predicting the likelihood  $p(\vec{x})$  for each sample in the database given the previously estimated model  $(\vec{\mu}, \Sigma)$ .

**Positive and negative feedback** The second relevance feedback strategy considers both positive and negative feedback, i.e. relevant and non-relevant items to create a model for discriminating between them. Therefore *support vector machines* (SVM) proposed by Vapnik [2000] (see section 3.3.5 for more details) are adopted which try to find the optimal separating hyperplane that maximizes the margin between the two classes. From the different kernel methods  $k(x, z)$  that can be used to map the features into a higher dimensional space for linear separability, the radial basis function (RBF) is used, since it has been shown to provide the best performance under various conditions [Guo and Li, 2003]. For matching the samples of the database to the trained SVM, the distance from the decision boundary

[Guo and Li, 2003] is used which is defined as

$$d(\vec{x}) = \frac{\alpha k(\vec{x}, \vec{z}) + \beta}{\alpha \vec{z}} \quad (8.2)$$

with the kernel function  $k(\vec{x}, \vec{z})$  the support vectors  $\vec{z}$ , the scaling parameter  $\alpha$ , and the bias  $\beta$ .

### 8.2.5 Multimodal fusion

Until now audio and video modality have been treated independently. The goal of the multimodal fusion step is to improve the performance by combining this complementary information (see section 3.4 for a summary of the information fusion). For increased flexibility only post mapping fusion is considered here.

As already mentioned in section 3.4, score combination methods require the scores of the different modalities to have a common range. Therefore, the location and scale of their distributions is modified to map them into equal ranges. The *z-score normalization* and its adaptation the *3-sigma normalization* have been proven to be quite reliable and are used here.

The normalized scores (probabilities, distances) of the audio and video modality are combined by score level fusion methods, since they have been proven to provide a better flexibility than the other approaches. Several score combination rules are considered within this work including the *product*, *sum*, *min*, *max* rule.

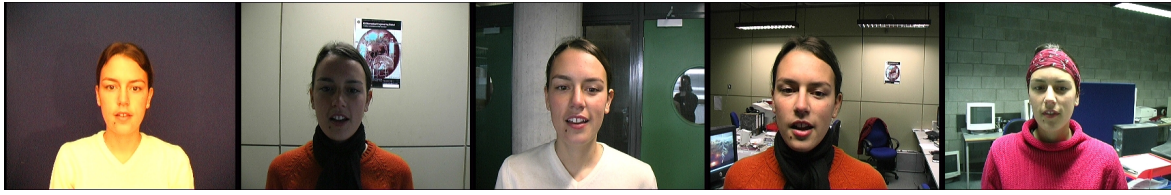
## 8.3 Experiments

Several experiments have been conducted to assess the performance and explore the limits of the developed multimodal person search system. More specifically the following aspects have been analysed

- Comparison of unimodal (audio, video) and multimodal system
- Comparison of different query paradigms including query by example (QBE) and relevance feedback (RF)
- Influence of different result set sizes onto the retrieval performance and speed
- Required number of iterations until convergence of the retrieval results

### 8.3.1 Dataset

While a large number of publicly available datasets exist for general image and video retrieval, there is no dataset available for multimodal person search and retrieval. Therefore,



**Figure 8.4:** Sample of the VALID Database showing an individual in 5 different environments.

the VALID Database<sup>6</sup> (see section A.4.1 for more details) has been adopted for the experiments. Although it is primarily made for developing and evaluating biometric systems, it is very similar to the chosen application scenario.

The VALID Database consists of 1060 multimedia clips of individuals in head and shoulder view saying a short sentence or counting numbers. Each of the 106 persons (27 female, 79 male) is captured in 5 environments (1 studio, 4 office) leading to 10 samples each. Both the acoustical (noise, reverberation) and visual characteristics (illumination, background) of the environments are quite diverse, making the data even more realistic for the given application scenario. Figure 8.4 shows the different environments for a single individual.

### 8.3.2 Evaluation

The problem of multimodal person search has been evaluated as a retrieval problem (see section B.3) based on manually annotated ground truth which splits the documents into relevant and non relevant sets. Given a ranked list of the documents in the database provided by the system, well known retrieval evaluation measures are computed. In order to provide reliable performance measures each sample (1060 files in total) has been considered as a query and the results have been averaged.

It is well known, that the policy of the user providing relevance feedback can have a strong impact on the evaluation results [Crucianu et al., 2004]. In order to provide reproducible results an automatic evaluation process without any real user has been used, that automatically selects all relevant and all non-relevant items within a given result set. Since this is less realistic than just marking some samples and even make mistakes, the reported results can be seen as an upper bound performance, which may not be achieved in reality. Nevertheless, since the number of relevant items is quite small (only 10 samples) the real performance will be quite close to this limit.

Generally retrieval measures can be divided into precision/recall and rank measures (see section B.3 for more details). As the most suitable measures from each category the *average precision (AP)* and *normalized average rank (NAR)* have been used throughout the experiments. The former measure describes the average ratio between relevant and retrieved items at the position of relevant items. The latter corresponds to the average rank of all the relevant items normalized over the number of items in the database.

<sup>6</sup><http://ee.ucd.ie/validdb/>

Modality	Approach	AP	NAR
Audio	QBE	0.196	0.280
Audio	SG	0.425	0.226
Audio	SVM	0.803	0.089
Video	QBE	0.317	0.126
Video	SG	0.633	0.062
Video	SVM	0.802	0.036
Multimodal	QBE	0.324	0.168
Multimodal	SG	0.753	0.087
Multimodal	SVM	0.991	0.001

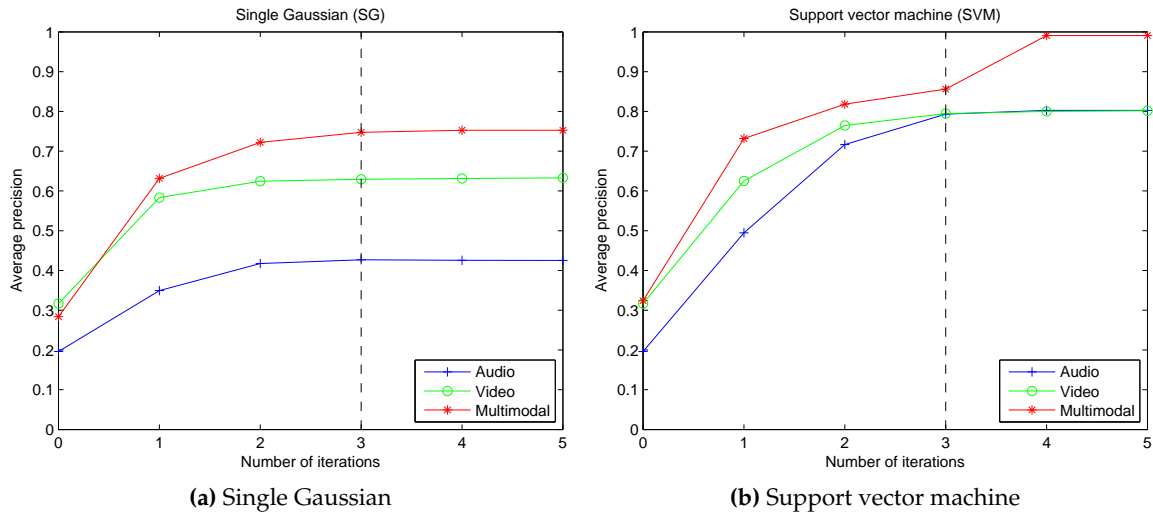
**Table 8.1:** Performance of the different modalities (audio, video, multimodal) and retrieval approaches (QBE, SG, SVM) based on two measures after convergence.

### 8.3.3 Results

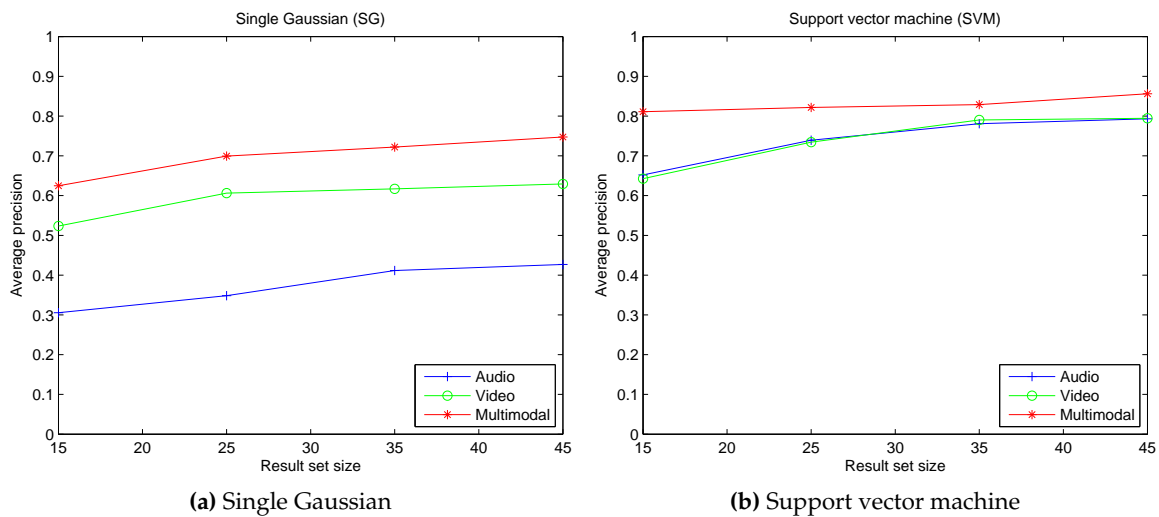
Table 8.1 provides a comparison of the different modalities (audio, video, multimodal) and retrieval approaches (QBE, SG, SVM) by showing the average precision and normalized average rank after convergence (5 iterations) of the relevance feedback. Comparing the different relevance feedback approaches with the query by example paradigm a large performance improvement can be achieved across all modalities. For the audio modality the average precision improves by 0.229 and 0.607 for the SG and the SVM based RF, respectively. For the video modality an improvement of 0.316 for the SG and 0.485 for the SVM are achieved. The gain is even larger for the multimodal system with 0.429 for the SG and 0.667 for the SVM. Comparing the different modalities to each other shows that the performance of all retrieval approaches can be improved by combining the audio and video information. While the improvement is only marginal for the QBE (0.007) it is much larger for the SG (0.120) and the SVM (0.189). The results are achieved with the optimal score fusion rule for each of the approaches. While the sum rule is the best fusion method for the QBE and the SVM, the product rule is the most suitable for the SG.

Figure 8.5 provides a more detailed view of the iterative retrieval process by plotting the average precision vs. the number of iterations. Iteration 0 actually corresponds to the initial query by example and iteration 5 to the final retrieval result after convergence reported in table 8.1. The results are shown for the maximal result set size of 45 items. It can be seen that the performance for both the SG and the SVM based approach converges after 3–4 iterations. While the video modality achieves a higher performance than the audio modality for the SG approach, both modalities are comparable for the SVM based approach. For both approaches the performance increases when audio and video information are fused to exploit the complementarity.

Figure 8.6 focuses on another aspect of the retrieval process by plotting the average precision vs. the result set size after 3 iterations. As expected the performance increases for



**Figure 8.5:** Performance of the different modalities and retrieval approaches over the number of iterations for a result set size of 45 items.



**Figure 8.6:** Performance of the different modalities and retrieval paradigms over the result set size for 3 iterations.

larger result set sizes, since more samples are provided as feedback which leads to a better model of the users search intention. For the SG approach the performance varies about 0.10 over the different result set sizes across modalities. For the SVM approach the variation of the unimodal systems (0.15) is larger than for the multimodal system (0.05). Nevertheless, it is interesting to see, that a larger result set size does not improve the performance pretty much, which allows to reduce the users feedback without large performance drops.

Finally, a short analysis of the users efforts in terms of iterations, result set size and required browsing time is provided. With the assumption that it takes a user 4 s to play a single sample and judge it either as relevant or irrelevant, a single iteration for a result set



size of 25 samples takes about 100 s. Considering 3 iterations as the average, a complete search and retrieval session takes about  $300\text{ s} = 5\text{ min}$ . In comparison to manually searching through all the 1060 items in the database which takes about  $4240\text{ s} = 70\text{ min}$  the effort is reduced by factor 14. While the retrieval performance may decrease for larger databases, the complexity reduction will be even larger.

## 8.4 Conclusion

### 8.4.1 Summary

An original system for multimodal person search and retrieval has been developed, that allows to efficiently search persons within audiovisual clips based on face and voice characteristics. It supports different query paradigms including query by example and relevance feedback. The relevance feedback can be either positive (one class) based on a single Gaussian or positive/negative (two class) based on a support vector machine. Furthermore, the audio and video modality are combined to reduce ambiguities and improve the overall performance.

The experiments show that through the combination of relevance feedback and multimodal fusion a very high retrieval performance can be achieved. Regarding the different relevance feedback approaches the two class RF based on a SVM constantly outperforms the one class RF based on the SG. Furthermore, by analyzing the performance over the number of iterations and result set sizes it is shown that for the given database the user effort can be reduced by factor 14 from the manual to the interactive search.

### 8.4.2 Future work

Possible future work may go into several directions.

In order to measure the performance on more realistic data, a database of real video blogs, talks and news may be created. Within this direction it is interesting to explore how the performance of the system is influenced by different number of samples and individuals.

So far the system considers only a single face sample (image) of the whole video stream for face recognition, which may be extended too multiple samples to improve the reliability of the visual part.

In the current system the audio and video modality are fused with equal weights, which might not be the optimal way. Depending on the performance of each modality under certain environmental conditions the weights may be adapted based on a priori information.

In order to evaluate the system with a large number of real users, an accessible web based user interface would be suitable solution. Furthermore, this would allow to assess the difference between the upper bound and the real performance of such a system.



## Chapter 9

# Visual person search

### 9.1 Introduction

#### 9.1.1 Motivation

With the increasing amount of available multimedia data efficient systems for searching and retrieving audiovisual (AV) documents are required. Traditional systems based on keywords are quite inefficient due to the time consuming annotation as well as linguistic and semantic ambiguities. Therefore, content based multimedia retrieval systems have been proposed that search and retrieve AV documents based on audio and visual features extracted from the content itself.

Every retrieval process usually starts with query where the user describes his search intention in a way the search system can interpret. For image and video retrieval a large variety of query paradigms have been developed including

**Query by keyword:** This traditional paradigm is directly adopted from text retrieval. Therefore each image needs to be tagged by a comprehensive set of keywords that can be automatically extracted from related textual sources such as surrounding text on web pages or closed captions or manually annotated by users in form of games such as the ESP Game<sup>1</sup> or Peekaboom<sup>2</sup>. The search itself then matches the query keywords to the associated keywords of the images and retrieves those that match most of them.

**Query by example:** Most content based image retrieval systems use this paradigm which requires a sample that appropriately describes the users search intention. The major problem is that for some applications such a sample may not be available which requires the user to browse the database to find a suitable sample.

**Query by drawing:** The use of human sketches for specifying a query has been proposed for content based image retrieval [Wang et al., 1997; Egenhofer, 1996]. Considering

---

<sup>1</sup><http://www.espgame.org/>

<sup>2</sup><http://www.peekaboom.org/>

the difficulty of exact drawing and the need for some artistic skills, this method is only applicable for a limited set of content such as single objects or shapes. For general image and video retrieval sketches are too time-consuming and the retrieval results are usually not exact enough.

**Query by concept:** Retrieving images or videos based on concepts [Wu et al., 2004] can be seen as a supervised learning problem, where classifiers (see section 3.3.5) are trained to detect a predefined set of concepts within the data and label it accordingly. During the search the user may choose a combination of these predefined categories and images or videos that fulfill this concept will be retrieved.

**Query by visual thesaurus:** Recently a new query paradigm for content based image retrieval has been proposed [Boujemaa et al., 2003] that provides the user with a summary of the database which is created by grouping items based on their similarity using unsupervised learning techniques (see section 3.3.4). Given this visual thesaurus, a user can quickly retrieve images based on a logical combination of these visual words.

The major idea of this work is to explore the query by visual thesaurus paradigm for visual person search.

### 9.1.2 Related work

Similar to multimodal person search (described in chapter 8) only little work has been in the specific field of visual person search. Nevertheless, related work can be found in two major areas: *content based image retrieval* and *surveillance*. The former deals with the search of images without the emphasis on a certain object class [Lew et al., 2006]. The latter focuses on the description of persons and their behavior within an environment. In the following one representative work for each of the categories are summarized.

Boujemaa et al. [2003] proposed the query by visual thesaurus paradigm for general image retrieval to overcome the limitations of the classical query by example approach. Therefore, an image is segmented into a set of homogeneous regions which are described with their average color. Given a database all the present images are grouped into several categories based on their visual similarity using competitive agglomeration clustering. For each of these categories a prototype is selected that represents the category within the visual thesaurus. The query is expressed as a set of positive and negative region categories which can be described as “find images that contains regions like these and not like those”. Given the images corresponding to the individual region categories the set of retrieved images is determined by logical rules. This original idea forms the basis for the developed system.

Quite recently, Hansen et al. [2007] developed a system for the automatic annotation of persons within a surveillance scenario. Beside the height and the gaze direction they also considered the appearance of a person. Therefore the human body is represented by 3 parts (hair, upper body, and lower body) and each of those parts is described with one

of 11 predefined color terms [Berlin and Kay, 1969]. The color terms are assigned using a rule based partitioning of the HSV color space. Given the annotation it is possible to search for people based on the query by concept paradigm. The major issue of describing the appearance with a set of predefined terms is that they may not be representative for the database and the distribution among the classes may be very irregular. In the worst case all persons may be assigned to a single color only.

### 9.1.3 Objectives

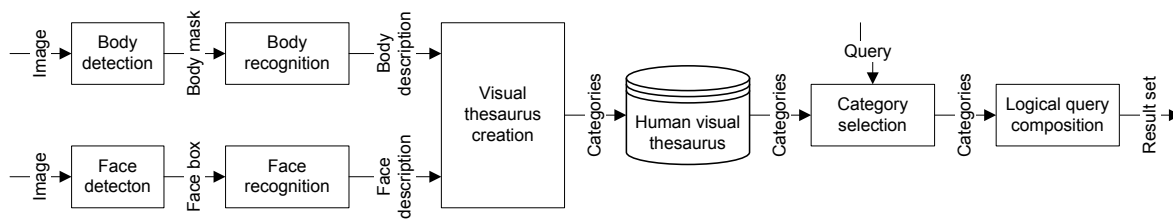
The objective of this work is to develop an efficient system for appearance based search of humans in images and videos. In order to support a wide range of applications and search intentions it considers several channels of information, including the face and the body of a person described both in a holistic and a component based way. For the search and retrieval process the query by visual thesaurus paradigm is adopted to provide an intuitive and efficient way for the user to describe his search intention. The so called *human visual thesaurus* is inspired by the way witnesses describe the appearance of involved humans by splitting it into several parts and choosing among a set of available descriptive terms (see section 4.2 for more details on this).

Although this work builds on the idea of Boujemaa et al. [2003] it differs in several ways. Instead of creating a single visual thesaurus for all low level image regions, an individual visual thesaurus is built for each of the body parts. Considering multiple thesauri requires also an adaptation of the logical query composition. For the body description we adopt the average color descriptor but consider 4 different color spaces to find the optimal one for a perceptual grouping. The face description is based on higher dimensional descriptors which makes the clustering more difficult. For the visual thesaurus creation 3 clustering methods with different characteristics are considered. Finally, we provide a comprehensive evaluation of the quality of the human visual thesaurus based on internal and external clustering evaluation criteria.

An article describing this work will be submitted to ACIVS 2009.

## 9.2 System overview

Figure 9.1 provides an overview of the proposed system for visual person search based on a human visual thesaurus. It consists of an offline part and an online part. During the *offline* phase the images within a database are analyzed, the bodies and faces of humans are detected and their appearance is described using several channels. Based on the extracted descriptions a human visual thesaurus is built by grouping human body parts based on their visual similarity. Within the *online* phase the user describes his search intention by choosing among the categories of the individual body parts. The selected categories are used to retrieve an individual set of images with humans that correspond to the chosen



**Figure 9.1:** Overview of the system for visual person search based on a human visual thesaurus. During the offline phase a human visual thesaurus is built by grouping the individual body parts based on their visual similarity. Within the online phase the user combines categories for each body part into a logical query and the system retrieves the corresponding images.

categories for each of the parts. The logical query composition step combines these image sets into the final result set based on logical operations.

### 9.2.1 Body analysis

The goal of the body analysis part is to detect and describe the appearance of the bodies of humans present within the images or videos.

#### Body detection

The body detection module detects and segments humans that are present within the image and delivers a binary segmentation mask for each of them. Since the considered database was captured in front of a green screen, chroma keying techniques [Jack, 1996] were used for the segmentation. In more realistic scenarios motion segmentation [Hu et al., 2004] or model based object detection methods such as the ones proposed by Dalal et al. [2006] and Wu et al. [2008] could be used.

#### Body description

The body description module (see section 5 for more details) splits the human body into several parts and describes them using visual low level features such as color and texture. All the body parts (whole, head, upper and lower) are considered for the visual search. Inspired by the original work of Boujemaa et al. [2003] the average color is used for the description of the individual body regions. Furthermore, this is motivated by the promising performance of it reported in section 5.3 for body recognition and its low dimensionality which supports the clustering. Nevertheless, to achieve an optimal visual grouping of the regions with respect to their perceptual visual similarity, different color spaces [Poynton, 2008] are explored including *RGB*, *YUV*, *HSV* and *CIE Lab*.

### 9.2.2 Face analysis

The goal of the body analysis part is to detect and describe the appearance of frontal faces of humans present within the images or videos.

### Face detection

The face detection module (see chapter 6 for more details) detects frontal faces of humans that are present within the image and provides their location and extend in form of the bounding box to the subsequent face description module. For the current scenario the holistic approach described in section 6.2 is considered since there are no facial occlusions present and the resolution of the used dataset is rather low.

### Face description

The face description module (see chapter 7 for more details) normalizes the face region and describes it in a holistic way with a  $30 \times 40$  pixel texture template. To avoid the curse of dimensionality which may lead to bad clustering results the dimensionality of the feature vector is reduced by applying principal component analysis (PCA). The number of dimensions is automatically determined based on the cumulative sum of explained variance.

### 9.2.3 Visual thesaurus creation

The goal of this step is to group the individual body and face parts based on the similarity of their corresponding visual low level features. In contrast to the approach by Boujemaa et al. [2003] this is done for each part individually which leads to a set of visual thesauri that together form the human visual thesaurus.

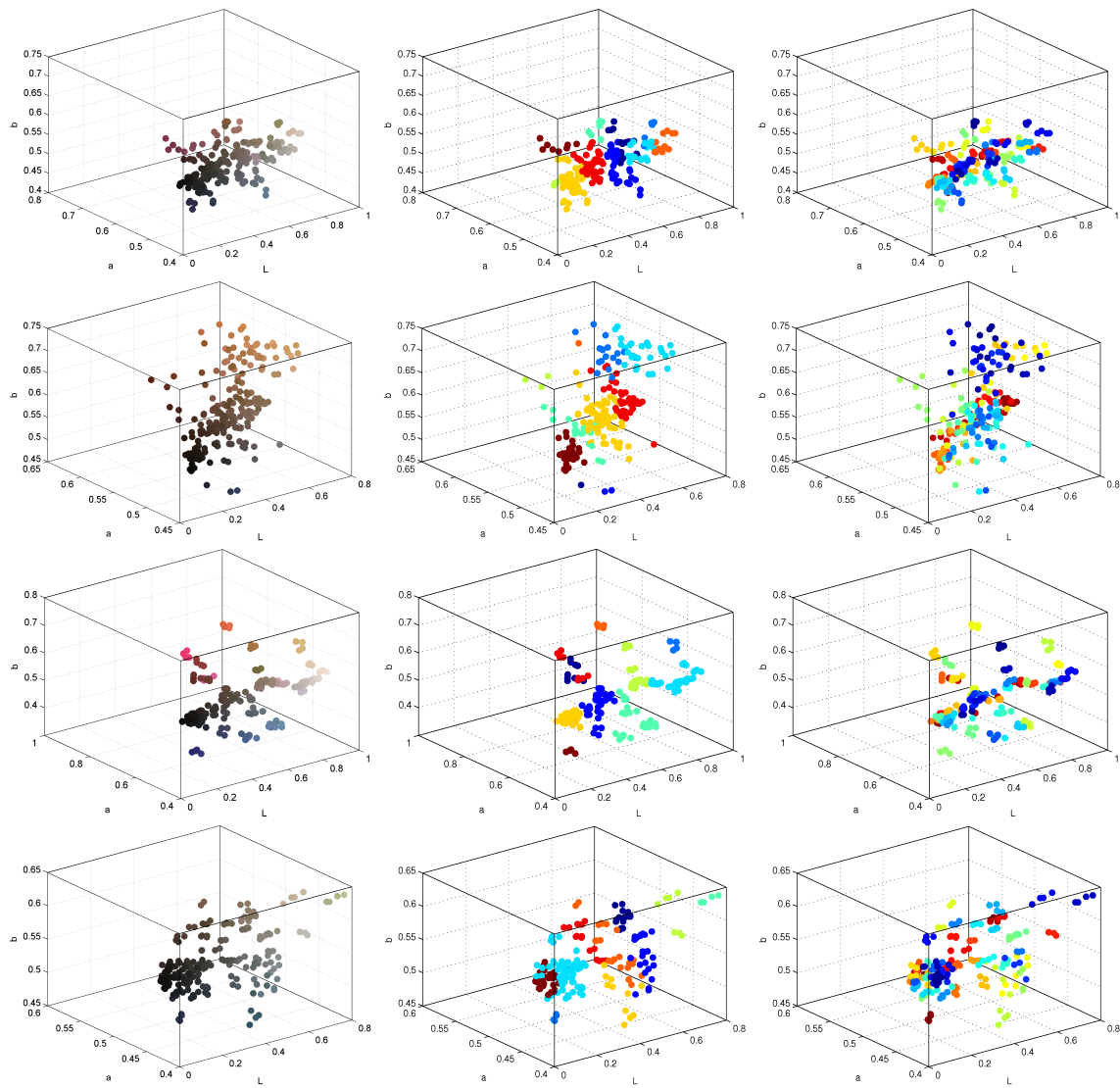
### Preprocessing

Since unsupervised learning (clustering) relies only on the features itself it is very important to normalize the individual dimensions into a common range and remove present outliers.

The normalization is especially important for some color spaces since individual color channels may have different ranges. Several methods have been developed for data normalization (see section 3.4.3 for an overview). Since the range of the individual features is known a priori *min-max normalization* with predefined limits is applied.

### Clustering

The grouping of the body parts according to their visual similarity is achieved through unsupervised learning or clustering (see section 3.3.4 for an overview). Within the current system only non hierarchical thesauri are used, but this may be extended to hierarchical ones to support several levels of granularity. For the partitioning several clustering methods with different characteristics have been considered including *kmeans* (KM), *fuzzy c-means* (FCM) and *agglomerative clustering* (AGG). Since the latter is a hierarchical clustering method, the final partition is obtained by pruning the dendrogram based on a predefined number of clusters.



**Figure 9.2:** Color distribution and predicted clusters and ground truth classes within the Lab color space. The rows correspond to the individual body parts (whole, head, upper, lower). The columns from left to right provide the color distribution, the 10 predicted clusters using agglomerative clustering, and the 52 ground truth classes for the costumes.

Figure 9.2 shows a clustering example for the individual body parts using the average color feature in the Lab color space when agglomerative clustering is applied. It shows the color distribution of the individual body parts within the color space, the 10 predicted clusters after clustering and the 52 ground truth classes. Depending on the distribution of the different classes (costumes) among the feature space there is a considerable overlap between them, which causes the clustering to group them together and leads usually to a lower number of predicted classes than ground truth clusters. Nevertheless, this is actually the correct behavior since the goal is to distinguish only between visually distinct costumes. By comparing the color distributions of the individual body parts with each other, it is interesting to see that both the shape and tendency to form natural clusters varies considerably. Espe-



cially the later fact may influence the performance of the clustering process. The Hopkins statistics (HS) [Banerjee and Dave, 2004] can be used to measure the clustering tendency. For the Lab color space the clustering tendencies are  $HS = \{0.88, 0.83, 0.91, 0.85\}$  for the whole body, head, upper and lower body, respectively. Given that all of these values are close to 1 applying clustering to this data will lead to well defined groupings. Furthermore, it confirms the visual impression that the upper body has the highest and the head has the lowest clustering tendency among the different body parts.

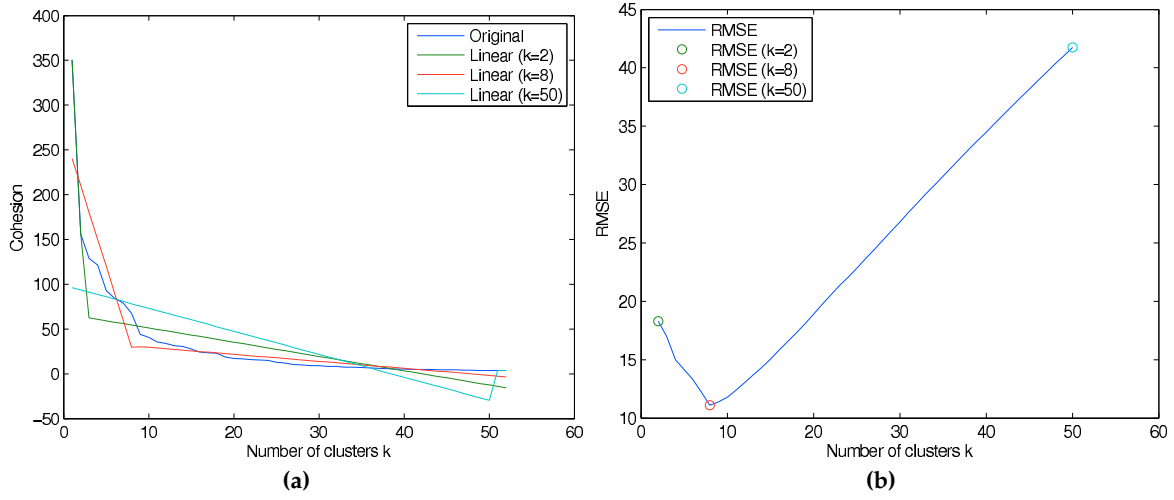
### Determine the optimal number of clusters

One of the biggest issues with clustering is to automatically determine the number of clusters that describes the data in the best way. Since clustering can be seen as an unsupervised learning problem, there is usually no human guidance on how many clusters are appropriate for a certain application. While some algorithms, such as agglomerative clustering provide ways to automatically determine the optimal number of clusters during the clustering process for other approaches such as k-means it needs to be predefined a priori. A general way to determine the optimal number of clusters is to run the clustering for a varying number of clusters within a suitable range and evaluate the results according to some evaluation criteria. Finding the optimal number of clusters can then be translated into finding the optimal point within an evaluation measure vs. number of cluster curve.

An evaluation measure that is widely used for this purpose is the cohesion which is defined in section B.6. Plotted against the number of clusters the curve usually exhibits a monotonic decreasing shape with a distinct knee as it is shown in the example in figure 9.3(a). Several methods have been developed to find the knee of this curve including the largest ratio difference between two point, the first point with the second derivative above some threshold, the point with the largest second derivative and the point on the curve that is furthest from the line fitted to the entire curve. Here the so called L-method proposed by Salvador and Chan [2004] has been adopted. In contrast to most of the other methods, it considers all points of the curve at the same time which makes it less sensitive to outliers and local trends that may not be statistically significant.

The location of the knee is determined by exploiting a typical property of these evaluation curves which states that the parts on the left and the right from the knee are approximately linear (see figure 9.3(a) for an example). If two lines are fitted well to the right and the left side of the curve, the crossing point between these two lines will be near the knee. Therefore, for each possible number of clusters  $c \in [2; b - 2]$ , with  $b$  being the maximum number of clusters, the points are partitioned into a left set  $L_c$  and a right set  $R_c$ . One line is fitted to each of the datasets and the root mean square error (RMSE) between the corresponding line and the original curve is computed. Both are combined into the overall RMSE defined as:

$$RMSE(c) = \frac{c-1}{v-1} RMSE(L_c) + \frac{b-c}{b-1} RMSE(R_c) \quad (9.1)$$



**Figure 9.3:** Determine the optimal number of clusters by finding the knee in cohesion vs. number of cluster curves.

Since the  $RMSE(c)$  provides a measure on how well the two lines at a certain  $c$  fit the original data, the optimal number of clusters  $\hat{c}$  can be derived by finding the  $c$  with the minimum  $RMSE(c)$  given as

$$\hat{c} = \underset{c}{\operatorname{argmin}} RMSE(c) \quad (9.2)$$

This method is illustrated for an example in figure 9.3, with the cohesion vs. number of cluster curve on the left and the corresponding RMSE vs. number of cluster curve on the right. The two-line approximation of the curve and the corresponding RMSE values are shown for  $c \in \{2, 8, 50\}$ . In this case the optimal  $\hat{c} = 8$  which is close the position of the knee.

### Visual thesaurus

Given a set of clusters  $C^p = \{c_i^p : i \in [1, k^p]\}$  for an individual body part  $p$  the visual thesaurus is formed by grouping the humans according to their cluster memberships and selecting one of them as the *prototype* for each category. This is achieved by averaging all the feature vectors that belong to a single cluster and select the human whose feature vector is the closest to this average feature vector in terms of a suitable distance metric (see section 3.3.1 for more details on matching). For all the considered parts the *Euclidean distance* was adopted. The prototypes are used to represent the category within the visual thesaurus as it can be seen in figure 9.4 which shows the visual thesauri for the individual body parts.

#### 9.2.4 Query by visual thesaurus

The query processing differs a lot between the query by example and the query by visual thesaurus paradigm. In the *query by example* paradigm (used for the multimodal person

search described within chapter 8) the user selects an image which is matched against all the images present in the database and the images are ranking according to their similarity. The ranked list is returned as the retrieval result and can be limited to a predefined number of items to get a fixed size result set. Thus the overall goal of this paradigm is to rank relevant images as high as possible. The *query by visual thesaurus* paradigm follows a different idea. The user describes his search intention by selecting categories from the visual thesaurus that are close to the his mental image. Through the clustering process each of the selected categories is directly linked to a set of relevant images, which can be simply retrieved by a hash lookup. Since there is no matching between the query and the database contents the retrieved result set is not ranked as for the query by example paradigm. Depending on the number of category members the result set may contain different number of items.

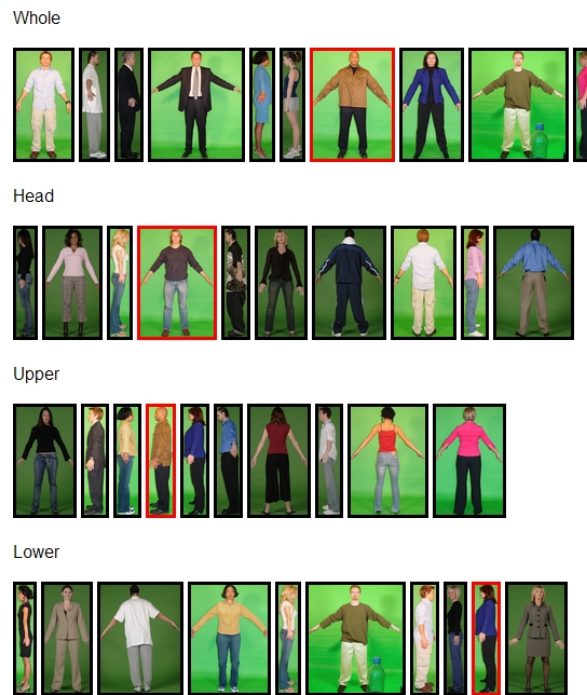
In relation to the traditional query by example paradigm it offers the following advantages

- Instead of a real sample only a mental image is required, which is translated into a query by choosing appropriate categories from the thesaurus.
- The complexity is shifted from the online phase to the offline phase since the query processing requires only a hash lookup and logical operations.
- If the precision of the retrieval is not satisfactory the user can make the query more specific by adding more categories through AND fusion.
- If the recall of the retrieval is too low the user can make the query more generic by adding more categories through OR fusion.
- It natively supports multimodal queries (search for several visually distinct humans) within a single query.

It also has some disadvantages which are mainly caused by the unsupervised learning approach

- There is no guarantee that the clustering leads to an optimal set of categories for the retrieval task.
- Image are assigned hard to a single category, which may lead to a loss in recall if only a single region is selected.
- Depending on the number of clusters there is generally a tradeoff between recall and precision.

The basic idea of the query by human visual thesaurus is illustrated by the example in figure 9.4 which shows the web based query interface of the search engine with the different body thesauri (whole, head, upper, lower) with 10 categories each. Since the face thesaurus can only be used for the retrieval of humans in frontal view it is excluded from this specific



**Figure 9.4:** Query interface with individual thesauri for the different body parts and some selected categories (red).

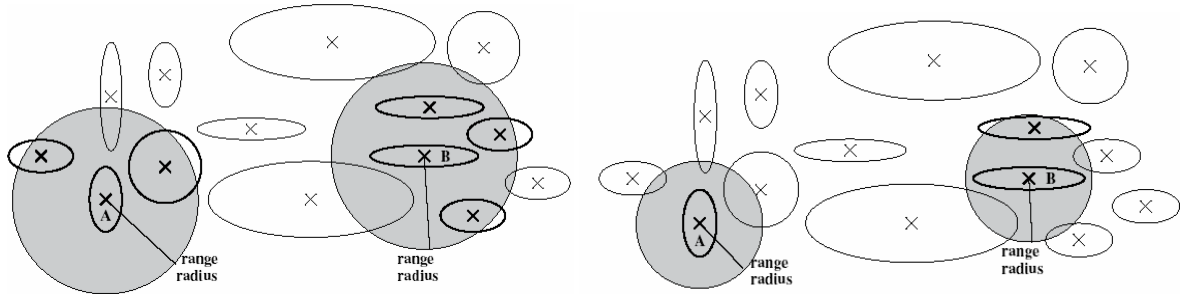
example. The individual result sets for each of the parts and the overall result set after the combination are shown in figure 9.6.

### Range queries

As already mentioned above, there is a tradeoff between recall and precision which is largely determined by the number of categories the data is grouped into during the clustering process. A large number of categories usually results in small sets of very similar items per category which gives a high precision. On the other hand, a small number of categories usually leads to bigger sets of items with lower similarity which gives a high recall. The third issue that has to be considered with respect to the number of categories is that too many categories will decrease the efficiency of the query process, since the user needs to select a large number of categories. Suitable values for the number of categories for each part are between 10 and 20 depending on the dataset.

Boujemaa et al. [2003] have proposed a simple way to improve the overall retrieval performance of the system through so called *range queries*. The basic idea is to define similarity on two levels. On feature level categories contain only very similar objects and provide a fine granularity to achieve a high precision. On cluster level for each of the selected categories a set of neighbor categories is included in the query to improve the recall without sacrificing the precision.

The neighbor categories of a selected category are determined by computing the dis-



**Figure 9.5:** Illustration of the neighbor selection for the range queries [Boujemaa et al., 2003]. Depending on the range radius a certain number of neighbor categories are considered as positive categories beside the selected categories.

tances between the prototypes of the selected category and all other categories of this thesaurus and choosing the ones within a range radius which can be adjusted during the online phase. In order to use a single radius for different features the distances are normalized by dividing through the maximum distance between the cluster prototypes in the thesaurus. This maps the distances into the range  $[0; 1]$ . Thanks to this range queries the search is less dependent on the initial database partition, since close categories are considered together.

Figure 9.5 illustrates the range query scheme for two examples with different range radii. Given the prototypes of the selected category  $A$  and  $B$  all the categories whose prototypes are within the range radius around the selected prototype are considered as neighbors. For the larger range radius this yields  $2 + 3 = 5$  neighbor categories, while for the smaller range radius it gives only  $0 + 1 = 1$  neighbor categories.

### Logical query composition

The concept of logical query composition for the human visual thesaurus differs from the one proposed by Boujemaa et al. [2003]. While in the original work only one thesaurus is constructed and the user is allowed to select positive and negative query categories, within the proposed system multiple thesauri are used from which the user selects only positive categories. All non-selected categories are considered as negative categories and excluded from further processing. Given a set of selected categories and a predefined range radius, the neighbor categories are determined as described above. Both together form the set of positive categories which is considered for extracting the result set of relevant images.

For performing the logical query composition an inverse index is used to retrieve the set of relevant images  $I_k^p$  for each of the positive categories  $c_k^p$  of a single thesaurus  $p$ . These images are combined into the overall set of relevant images for thesaurus  $p$  using OR fusion which can be written as

$$I^p = \bigcup_k I_k^p \quad (9.3)$$

The relevant image sets  $I^p$  of the individual thesauri  $p$  can be combined into an overall

set of relevant images  $I$  using AND or OR fusion which can be written as

$$I = \bigcup_p I^p \quad (9.4)$$

$$I = \bigcap_p I^p \quad (9.5)$$

The user can freely choose between these two operations depending on his interest. For example, if a user wants to find persons wearing a red shirt and blue trousers he may select the appropriate categories from the upper body and lower body thesauri and choose AND fusion to search for this specific combination. If he is interested in finding humans that wear either a red shirt or blue trousers or both he may choose OR fusion which gives more generic results. In the case that the user does not select any category for a thesaurus  $p$  the image set  $I^p$  will be empty and in combination via AND fusion with the other thesauri will cause an empty set  $I$ . In this case  $I_p$  is not considered for the fusion with the other image sets.

The logical query composition for the selected categories of the query from figure 9.4 is illustrated in figure 9.6. It starts by showing the result sets  $I^p$  of the individual thesauri  $p$  which are combined into the final image set  $I$ . Since AND fusion is considered for this example only the images that are present in all the individual result sets are included into the final result set. This figure also provides some visual impression of the clustering results since only a single category has been selected for each of the body parts. The selected category for the whole body contains all 4 samples of this very unique costume. The selected category for the head contains mainly middle brown heads with either homogeneous middle brown skin or a combination of bright skin and dark hair. This already illustrates the major issue of the head description. While the selected category for the upper body contains only a small number of brown shirts, the selected category for the lower body contains a large number of dark grey or black trousers, which illustrates the fact that there is less variety for the lower than for the upper body.

### 9.3 Experiments

Several experiments have been conducted to assess the performance and explore the limits of the developed system for visual person search based on the query by visual thesaurus paradigm. Since most of the processing is done during the offline phase the evaluation focuses on the quality of the built visual thesauri. With respect to the clustering the following aspects have been considered

- 3 cluster methods (KM, FCM, AGG)
- Different number of clusters (both manually set and automatically determined)

Furthermore, for each of the different channels (body, face) other facets were analyzed. While for the body thesaurus this included

Whole



Head



Upper



Lower



Combination (and)



**Figure 9.6:** Result sets for the individual thesauri and after the logical query combination. Since the AND rule was chosen for the inter thesaurus combination only the images (humans) present within all the individual result sets are included into the final result set.

- 4 representations (whole, head, upper, lower)
- Only 1 feature (average color)
- 4 color spaces (RGB, YUV, HSV, Lab)

for the face thesaurus the following parameters were used

- Only holistic representation

- Texture template of 30x40 pixels size
- Only grayscale color

### 9.3.1 Dataset

All the experiments are based on the Free Character Database (see section A.2.2 for more details), which is the only available database that provides images with a sufficient resolution to explore body and face analysis together. It contains 216 images of 22 persons with 54 costumes in 4 different views (front, back, left, right). While only the 54 frontal images will be used for the face analysis, all 216 images will be used for the body analysis.

### 9.3.2 Evaluation

Since the creation of a visual thesaurus is based on unsupervised learning, it has been evaluated as a clustering problem (see section B.6 for the details). Both internal measures (silhouette coefficient) based on the features and predicted clusters as well as external measures (purity, coverage, quality) based on ground truth classes and predicted clusters are considered here. The external measures are computed using the manually annotated costume and identity labels for the body and face, respectively.

### 9.3.3 Results

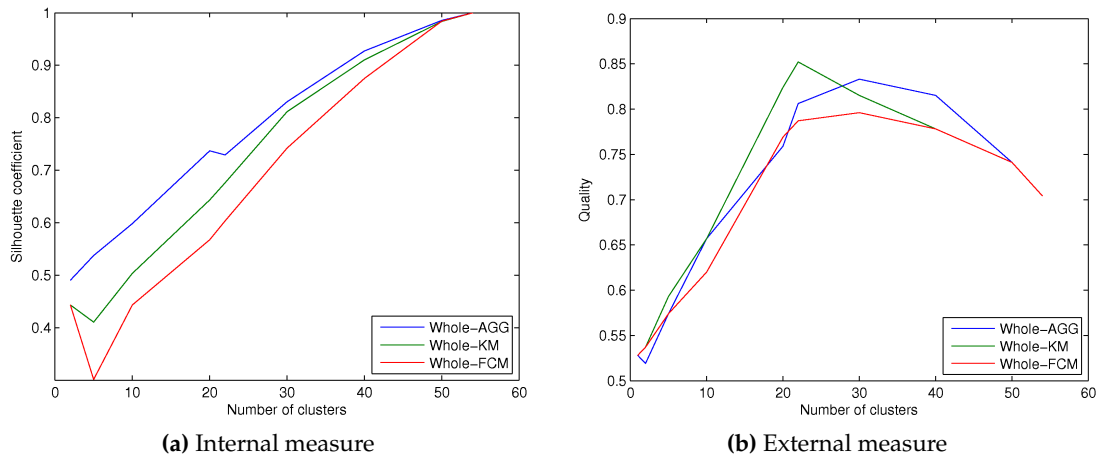
#### Face thesaurus

Figure 9.7 provides a summary of the face thesaurus experiments that considered only the holistic face representation, different clustering methods and number of clusters. The performance is evaluated in terms of silhouette coefficient and quality vs. the number of clusters. Considering the internal quality using the silhouette coefficient the different clustering methods can be ranked in the following order: agglomerative clustering, k-means and fuzzy c-means. What is especially interesting, is that while both partitioning approaches (KM and FCM) exhibit a significant drop at  $k = 6$  the silhouette coefficient of the agglomerative clustering is monotonically increasing as expected. Using the external quality measure the ranking is slightly different: k-means, agglomerative clustering, fuzzy c-means. With respect to the highest quality the approaches not only differ in the value but also in the number of clusters for which this is achieved. While the KM achieves its highest quality of 0.85 for 22 clusters which is equal to the number of classes, both FCM and AGG reach their highest quality of 0.79 and 0.83 for 30 clusters.

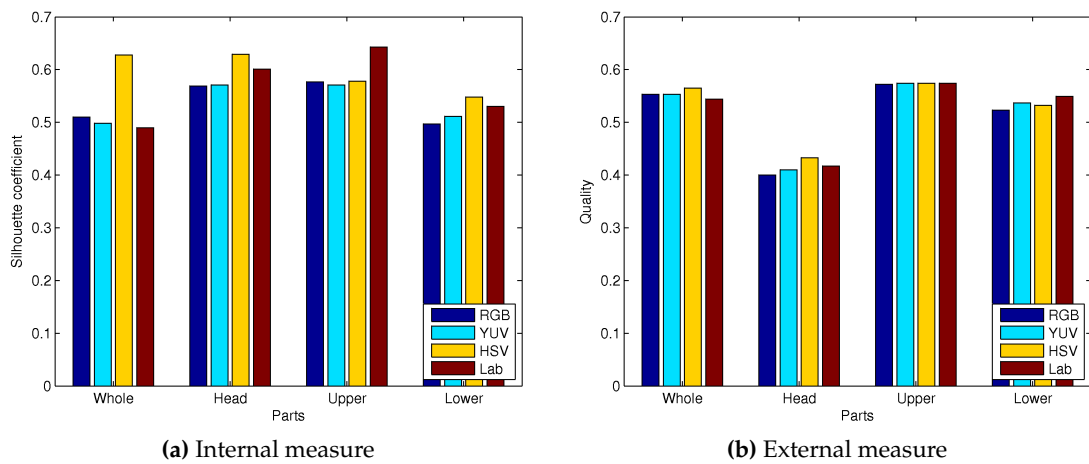
#### Body thesaurus

The goal of the first body thesaurus experiment was to compare the different color spaces with each other and determine the optimal one for the clustering. Therefore all the color





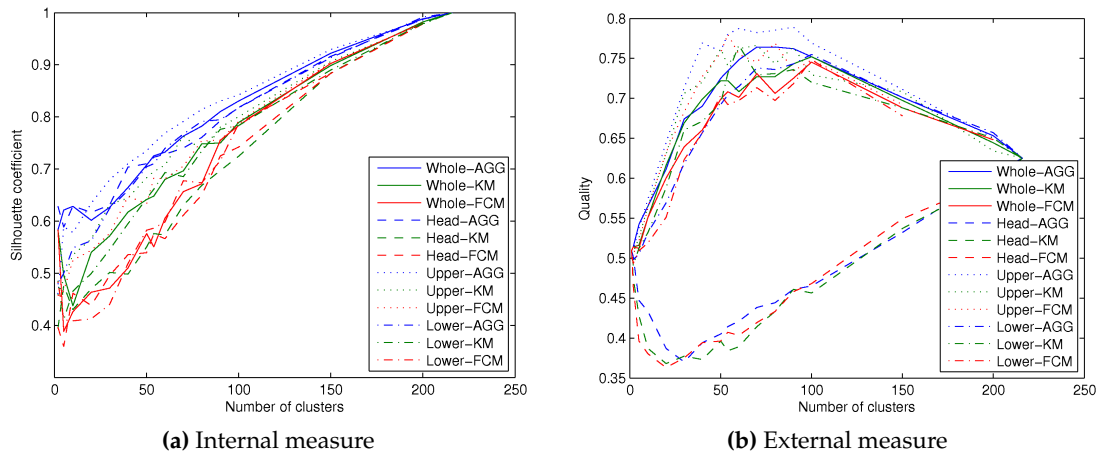
**Figure 9.7:** Internal and external quality assessment of the visual face thesaurus for different clustering methods. The best performance with respect to the internal criterion is achieved by the agglomerative clustering. On the other hand, the best performance regarding the external criterion is achieved by k-means clustering.



**Figure 9.8:** Internal and external quality assessment of the visual body thesaurus for different color spaces. It can be seen, that HSV and Lab color spaces achieve a better grouping quality than the RGB and YUV color spaces.

spaces were clustered using the different clustering methods for a predefined  $k = 10$ . Since the results were consistent over the different clustering methods figure 9.8 provides only the performance of the agglomerative clustering. It compares the quality of the color spaces across the different body parts using internal and external quality criteria. It can be seen that the HSV and Lab color space consistently outperform the RGB and YUV color space due to the more perceptual color description. Considering all the parts and both criteria the HSV color space is slightly better than the Lab color space which makes it the optimal choice for the following experiment.

The goal of the second experiment was to explore the performance of the different clus-



**Figure 9.9:** Internal and external quality assessment of the visual body thesaurus for different clustering methods. According to the results the clustering methods can be ranked: agglomerative clustering, k-means, fuzzy c-means. The ranking of the individual body part is the following: upper body, whole body, lower body, and head.

tering methods across different number of clusters and determine the optimal method which achieves the best quality for the individual body thesauri. Based on the results from the previous experiment only the HSV color space is considered here. Figure 9.9 provides a comparison of the clustering methods across the different body parts in the same way as figure 9.7 provides it for the face thesaurus. The clustering methods (AGG, KM, FCM) are grouped by color (blue, green, red) and the body parts (whole, head, upper, lower) are grouped by line style (solid, dashed, dotted, dash-dotted). Considering both the internal and the external criteria the clustering methods can be ranked in the following way: agglomerative (blue), k-means (green) and fuzzy c-means (red). In the same way the following ranking of the body parts can be derived: upper body (dotted), whole body (solid), lower body (dash-dotted) and head (dashed). With respect to the different criteria the observations from the face thesaurus are confirmed. The internal quality is increasing with an increasing number of clusters, while the external quality is increasing until a prominent peak and decreases after that. The only exception is the head for which the inverse happens. This is caused by the confusion between skin and hair colors, which leads to a large variation across different views. The optimal quality of the upper body (0.78), the whole body (0.76) and the lower body (0.76) is achieved with a  $k = 60 - 70$ . The optimal quality for the head (0.60) is much lower and only achieved for the maximum  $k = 216$ .

### Search and retrieval process

The experiments on the visual thesaurus creation were complemented by a set of retrieval runs which were evaluated subjectively. The general conclusion is that the human visual thesaurus provides an efficient way to search for persons based on a mental image only.

With respect to the individual body parts, the upper and the lower body provide the most intuitive results since the average color provides a quite reliable description of the mostly homogeneous regions. For the whole body the results are a little less intuitive since the average color is not a very reliable description if the appearance of the individual body parts differs too much. The major problem with the head is that there is currently no distinction between the skin and the hair which causes ambiguities between different views. The use of the face thesaurus is not as intuitive as the one for the body thesaurus which may be caused by the holistic representation and the larger feature dimensionality. Nevertheless, it still provides a suitable way to search for faces based on prominent differences such as complexion, beards, glasses etc.

## 9.4 Conclusion

### 9.4.1 Summary

Within this work an efficient system for visual person search has been developed. It is based on the recently proposed query by visual thesaurus paradigm [Boujemaa et al., 2003] that goes beyond the simple query by example approach. The original work has been extended towards a human visual thesaurus that consists of an individual thesaurus for each of the parts a human description is composed of. In this way it provides an efficient summary of the persons present within the database that is comparable to the way humans describe each other. Given such a human visual thesaurus the search does not require a real image for the query but only a mental image (memory) of how the person of interest looks like. The search intention is then described by selecting appropriate categories from the individual part thesauri and composing them into a logical query. Especially the latter step allows for a large range of queries reaching from very specific ones such as “find people with red shirts and blue trousers” to more generic ones such as “find people with dark shirts or dark trousers”.

The system is based on the hierarchical human analysis framework described in chapter 4 and the individual modules described in the chapters 5, 6, and 7. For the body description both the holistic (whole) and the component based representations (head, upper, lower) are used and the individual parts are described using the average color descriptor within different color space (RGB, YUV, HSV, Lab). For the face description only the holistic representation is used and described using a PCA reduced texture template. Different clustering methods (kmeans, fuzzy cmeans, agglomerative) have been explored to find the optimal one for the visual thesaurus creation.

Extensive experiments have been carried out to assess the quality of the visual thesaurus which is crucial for an efficient retrieval using this query paradigm. Both internal and external clustering measures have been considered. Regarding the body thesaurus it has been shown that the HSV and Lab color spaces achieve a better grouping than the RGB and YUV

color spaces. The most suitable clustering method is the agglomerative clustering with average linkage for all the body parts. The quality ranking of the individual body parts (upper, whole, lower, head) confirms the results from the body recognition experiments described in chapter 5. With respect to the face thesaurus agglomerative clustering provides again the best clustering results closely followed by the k-means clustering. Besides evaluating the clustering quality the actual search and retrieval process was evaluated subjectively. The results show that a human visual thesaurus provides an efficient and intuitive way to search for people based on a mental image only.

#### 9.4.2 Future work

Although the developed human visual thesaurus provides already an efficient way for searching persons within images and videos without the need for an initial sample, it may be extended into several directions.

One way to improve the quality of the visual thesaurus is to postprocess the clusters produced by the clustering algorithm [Tan et al., 2005]. Loose clusters with a large number of elements may be split into several clusters to improve the precision of the retrieval. Close clusters with a small number of elements may be merged to improve the recall of the retrieval.

In order to make the approach more suitable for large datasets and provide a coarse-to-fine categorization of the database content, a hierarchical visual thesaurus may be created. Therefore, the agglomerative clustering method, which has been used for partitioning here, could be applied. Furthermore, other hierarchical clustering methods, such as divisive clustering and hierarchical kmeans, can also be considered.

The query interface can be further improved by a distance based arrangement of the individual categories in order to make the selection of multiple categories more intuitive. Therefore, distance-preserving projection methods can be used to map the high-dimensional features to a low-dimensional representation.

Finally, the system should be evaluated on larger databases and other application scenarios where the appearance of the clothes provides an important cue for the search of people such as sports and surveillance.

## Chapter 10

# Personalized human computer interaction

### 10.1 Introduction

#### 10.1.1 Motivation

With the rapid development in hardware and software for audiovisual capture and analysis, electronic equipment gains more and more intelligence and is transformed into “smart” devices that interact with humans in a more active way (see chapter 1 for more details on this). An interesting application is the replacement of the classical Automated Teller Machine (ATM) by a more intelligent cash machine [Ignasiak et al., 2007] that provides additional functionality: advanced costumer authentication, voice and gesture based interaction, environment monitoring and unusual behavior detection.

#### 10.1.2 Related work

Ignasiak et al. [2007] proposed an architecture for such an intelligent cash machine that consists of the following modules:

**Environment analysis module:** This module provides the sensoric capabilities for the cash machine through the audiovisual analysis of humans near the cash machine. It detects and tracks people within the audiovisual data, identifies them and recognizes their intentions by interpreting gestures, expressions and speech commands.

**Conversational agent module:** This module is responsible for the interaction between the cash machine and the user, based on the information regarding the user and the environment provided by the environment analysis module. It controls the behavior of an audiovisual avatar that represents the agent itself in the human computer interaction.

**Speech synthesis module:** Controlled by the conversational agent module it provides the voice for the avatar.

**Visualization module:** Controlled by the conversational agent module it provides the visualization of the avatar.

**Cash machine module:** This module resembles the mechanical part of the intelligent cash machine and is comparable to the classical ATM. In contrast to the classical approach it is controlled by the more intelligent conversational agent.

**Storage and retrieval module:** It is a very important module from the surveillance point of view. It not only stores the audiovisual content and extracted meta-data from the environment analysis module but also adds security measures to restrict the access to this data.

### 10.1.3 Objective

The objective of this work is to develop the visual part of the environment analysis module, that provides some of the required functionality by detecting and tracking people, their faces and hands in order to identify them and recognize their gestures. Furthermore, the focus is laid on a tight integration of the different channels (body, face and hands) and an exchange of information between the modules to increase the performance of the overall system. More specifically we will explore how the identity of a person can be used to improve the reliability of the hand detection and gesture recognition.

This work has been developed together with Daniel Rodriguez<sup>1</sup> (UniS), Surachai Ongkittikul<sup>2</sup> (UniS) and Mustafa Karaman<sup>3</sup> (TUB) within the scope of the European Network of Excellence (NoE) VISNET II<sup>4</sup>. The contributions to the overall system are the face detection and tracking module, the face recognition module, and the skin detection step used within the hand detection and tracking module. The work has been published in ELMAR 2008 [Rodriguez et al., 2008].

## 10.2 System overview

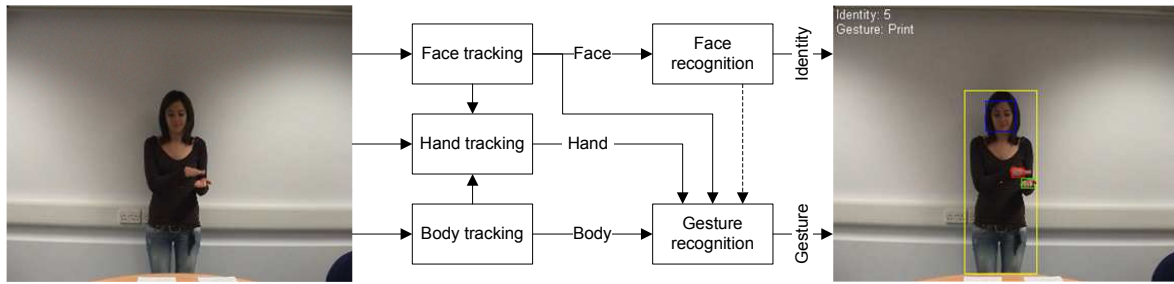
Figure 10.1 provides an overview of the developed system, its individual modules and the interaction between them. A static camera mounted at the cash machine captures a video of the user in medium to wide angle view as it can be seen on the left. The *body tracking* module detects and tracks the bodies of people and provides their position and extent to the hand tracking and the gesture recognition. The *face tracking* module detects and tracks the faces of the people and provides their locations to the gesture and face recognition. The *face recognition* stage takes the detected faces and recognizes people based on their facial appearance. The *hand tracking* detects and tracks the hands of a person and provides their

<sup>1</sup>daniel.rodriguez@surrey.ac.uk

<sup>2</sup>surachai.ongkittikul@surrey.ac.uk

<sup>3</sup>karaman@surrey.ac.uk

<sup>4</sup><http://www.visnet-noe.org/>



**Figure 10.1:** Overview of the system for personalized human computer interaction.



**Figure 10.2:** Intermediate results of the individual tracking modules.

position to the gesture recognition. The *gesture recognition* uses the location and extent of the person and its body parts to recognize the signs of the user. The results of the whole visual analysis are shown on the right including the location of the body, face and hands as well as the identity and gesture currently performed by the person. Figure 10.2 illustrates the output of the individual tracking modules by showing some intermediate samples.

The following sections describe the individual modules in more detail. For modules that utilize an approach described previously within this thesis only a short summary with a link to the related section is provided. On the other hand, modules that have not been described before will be discussed in more detail such as the skin modeling and the gesture recognition.

### 10.2.1 Body detection and tracking

In the given scenario the users are observed by a static camera and the number of users within the field of view is quite small (between 1 and 3).

**Body detection** Since the video data is captured by a static camera, person detection is based on background subtraction and object classification [Hu et al., 2004]. For background subtraction the approach proposed by Karaman et al. [2006] was adopted. The method models the background using a Gaussian Color Model (GCM) and classifies pixels as foreground/background based on the distance to this model. A post-processing stage restricts

foreground regions to regions of interest which are obtained by temporal differencing. Finally objects are detected by grouping adjacent pixels using connected component labeling. In order to distinguish people from other objects shape and size heuristics are applied.

**Body tracking** A region based tracking approach [Gupte et al., 2000] is used to track the detected people between frames. People in adjacent frames are matched based on their position. This leads to a distance matrix that compares all currently detected with all previously tracked bodies. Hungarian matching [Munkres, 1957] is used to extract one-to-one assignments between the detected and tracked bodies. Certain heuristics are considered to handle appearing and disappearing objects.

### 10.2.2 Face detection and tracking

In the given scenario, the user usually faces the cash machine during the interaction, so the goal of this step is to detect and track frontal faces.

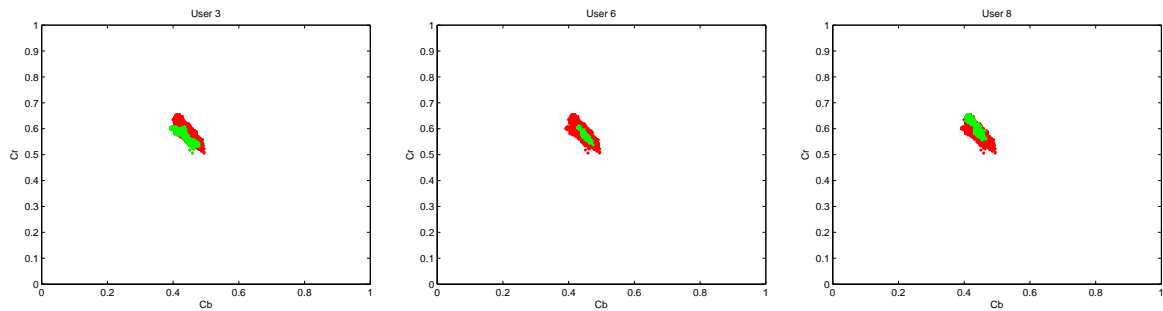
**Face detection** Since the cash machine is a cooperative scenario where the person wants to interact with the machine, occlusions are not very likely. Therefore, the holistic face detection approach by Viola and Jones [2004] is adopted (see section 6.2 for more details). It is based on the combination of Haar features and a classifier cascade trained using Adaboost. The Haar features are binary texture features, that can be efficiently computed using an integral image. The classifier cascade consists of multiple strong classifiers, which are applied sequentially to the remaining face candidates after each stage. Each of the strong classifiers combines multiple weak classifiers, that consist of a single Haar feature and a weight, by summation and thresholding. Both the strong classifiers and the classifier cascade are trained using Adaboost.

**Face tracking** In order to track faces between frames, the region based tracking approach from the body tracking is adopted. In combination with the body tracking certain heuristics are used to handle disappearing and reappearing faces.

### 10.2.3 Face recognition

In order to support user specific interactions, people in front of the camera need to be identified. Considering the application scenario, face recognition is the most suitable approach. In the present system the holistic approach (described in section 7.2) was adopted. A face region provided by the *face tracking* module is normalized to a predefined size and orientation. Uneven illumination is compensated using local contrast adjustment. Since the number of features for the extracted templates is quite large, principal component analysis (PCA) is applied to reduce the dimensionality. The features are projected into the reduced face space and a minimum distance classifier is used for the classification. To increase the robustness of





**Figure 10.3:** Distribution of skin color within the YCbCr space for all users (red) and three individual users (green). The individual skin color distributions have quite different locations and a smaller variation than the overall skin color distribution.

the face recognition, the information of multiple samples from the tracked face are combined using majority voting (described in section 3.4.2).

#### 10.2.4 Hand detection and tracking

Besides body and face, the gesture recognition module largely relies on the motion of the hands, in relation to each other and the other body parts.

**Skin modelling and detection** Since hands are highly articulated objects it is quite difficult to detect them based on shape or appearance. Thus, hand detection is based on an initial skin detection step similar to the one used by Habili et al. [2004].

Although skin color naturally forms a tight cluster within the different color spaces, robust skin detection is quite difficult due to the following challenge: While skin color varies considerably between different ethnicities and illuminations, clothes and the background may have skin-like colors which makes them difficult to distinguish from real skin. This situation is illustrated in figure 10.4 which shows samples from the VISNET II Cash Machine Database (described in section A.3.1) with persons of different skin colors and skin/non-skin colored clothes.

The confusion between skin and non-skin may be reduced by using specific skin color models for each user, as it is shown in figure 10.3 where the individual skin distribution (green) of the users (3,6,8) of the VISNET II Face Database is overlayed onto the overall skin distribution (red) of all users. This example shows that the individual skin color distributions have a different locations and smaller variation than the overall skin color distribution.

Based on this observation it seems reasonable to assume that the skin detection can be made more insensitive to skin colored clothes by using user specific skin models instead of a generic model since the overlap between skin and non skin pixels can be decreased. The major question is how to built and consider user specific models for skin detection. Gener-

ally all the developed methods model the skin color as a multivariate Gaussian distribution (described in section 3.3.3) in the YCbCr color space. Therefore the pixels are transformed from RGB color space to the YCbCr color space and the Y component is discarded to make the skin model insensitive to illumination changes. A pixel  $\vec{x}$  is classified as skin/non-skin based on the probability  $p_M(\vec{x})$  given the skin model  $M = (\vec{\mu}, \Sigma)$  and a thresholding operation. The threshold  $\hat{p}$  is determined from the training data in the way that 5% of the pixels are considered as outliers to cope with non-skin pixels inside the detected face region.

Given that, 3 different methods with different specialization have been developed. The *generic* approach uses a common skin model  $M$  and threshold  $\hat{p}$  for all users, while the *specific* approach uses an individual skin model  $M_u$  and threshold  $\hat{p}_u$  for each user  $u$ . The *hybrid* approach is a combination of both, by using a common skin model  $M$  and individual thresholds  $\hat{p}_u$ . The hybrid and the specific approach can be either based on offline trained skin models for each user and their online recognized identity provided by the *face recognition* or on online trained skin color models based on the face region provided by the *face detection*.

Given the binary skin mask, the individual hands are detected by discarding skin pixels outside the person mask, provided by the *person detection* module, grouping the remaining pixels into regions using connected component labeling and classifying the skin blobs in relation to the face position provided by the face detection module.

**Hand tracking** For the simultaneous tracking of two hands *particle filtering* [Arulampalam et al., 2002] is employed. This method also known as conditional density estimation (condensation) [Isard and Blake, 1998] is a popular approach that recursively constructs the posterior probability distribution function of certain features. Based on the initial position of both hands provided by the *hand detection* module each hand is tracked using a individual particle filter. To resolve the ambiguity between both hands and the face when they are close or occluding each other kmeans clustering (described in section 3.3.4) is employed. More details regarding the hand tracking can be found in a recently published article by Ongkittikul et al. [2008].

### 10.2.5 Gesture recognition

For the gesture recognition the approach of Bowden et al. [2004] was adopted, which recognizes hand gestures and breaks them into a limited set of building blocks (visemes), as used in automated sign recognition approaches [Vogler and Metaxas, 2001]. In contrast to other approaches, it considers not only the shape of the hands but also their movement, location with respect to each other and to other body parts (face, torso).

The approach is based on a two-stage classification process. The first stage is concerned with data reduction where each frame is reduced to a single binary feature vector that fully describes all defined gesticulation units present in the frame. The second stage classifies a

Id	HA	TAB	SIG	SIG-B	DEZ
0	No hands	No hand	Still	Single hand	–
1	Right high	Face	Up	Move apart	B
2	Left high	Chest	Down	Move together	$\bar{B}$
3	Side by side	Neutral space	Left	Move united	$\hat{B}$
4	In contact	Stomach	Right		A
5	Crossed	Other			$\dot{A}$
6					5
7					O
8					G
9					V
10					W

**Table 10.1:** Possible labels of the different stage 1 classifiers. The hand-shape names in DEZ are defined as in the British Sign Language [Brien, 1992].

sequence of observed feature vectors into one of the trained gestures.

**Stage 1 classifier** The stage 1 classifiers used to generate the feature vectors describing all the defined gesticulating units appearing within a frame is the same as in [Kadir et al., 2004]: *HA* (classifies hand location with respect to each other), *TAB* (hand location with respect to key body locations), *SIG* (hand movement) and *DEZ* (hand-shape). However, in this work the classifiers are applied to both right (R) and left (L) hand instead of the dominant hand only, and the movement of both hands with relation to each other is described by a new *SIG-B* classifier. This gives a binary feature vector of  $6 + 2 \times (6 + 5 + 11) + 4 = 54$  dimensions with the following format: HA/TAB-R/TAB-L/SIG-R/SIG-L/SIG-B/DEZ-R/DEZ-L. The possible labels for each of these classifiers can be seen in Table 10.1.

**Stage 2 classifier** Having reduced each frame in an observed video sequence into a binary feature vector, the stage 2 classifier selects a label from a set of pre-trained gesture models that best describes the sequence of binary feature vectors. Therefore, each of the possible binary feature vectors is mapped to a unique symbol that corresponds to a state in a first order Markov chain.

Since the dimensionality of the binary feature vector is very high, building a full ergodic Markov chain is too costly. However, many of the possible features are mutually exclusive and even fewer transitions between the states are possible due to the physical limitations of the body. Thus, only a partial ergodic Markov is built based on the states and transitions that occur in the training data. For each gesture  $g$  the training leads to an individual sparse state transition matrix  $P_g(s_t|s_{t-1})$ .

For the recognition, an observation sequence  $s$  of length  $T$  is matched against the models of the different gestures  $w$  by computing the probability  $P(w|s) = \prod_{t=1}^T P_w(s_t|s_{t-1})$ . The gesture  $w$  with the maximum probability  $P(w|s)$  is assigned to the observation sequence



**Figure 10.4:** Samples for each of the users considered within the database.

according to the maximum a posteriori MAP rule. In contrast to the work of Kadir et al. [2004] nominal probabilities are used if an exact match for each vector transition can not be found.

The identity of a user determined by the *face recognition* module enables the use of individual gesture classifiers  $P_w^u$  trained for each user  $u$ . In this manner, inter signer variations caused by different gesticulation patterns and hand shapes can be eliminated. This offers the advantage that classification performance is increased since only intra signer variations need to be modeled which is particularly important for hand shapes. A single generic  $P_w$  and multiple user specific classifiers  $P_w^u$  have been considered within the experiments.

### 10.3 Experiments

This section describes a set of experiments that have been conducted to assess the performance of selected individual steps and the overall system.

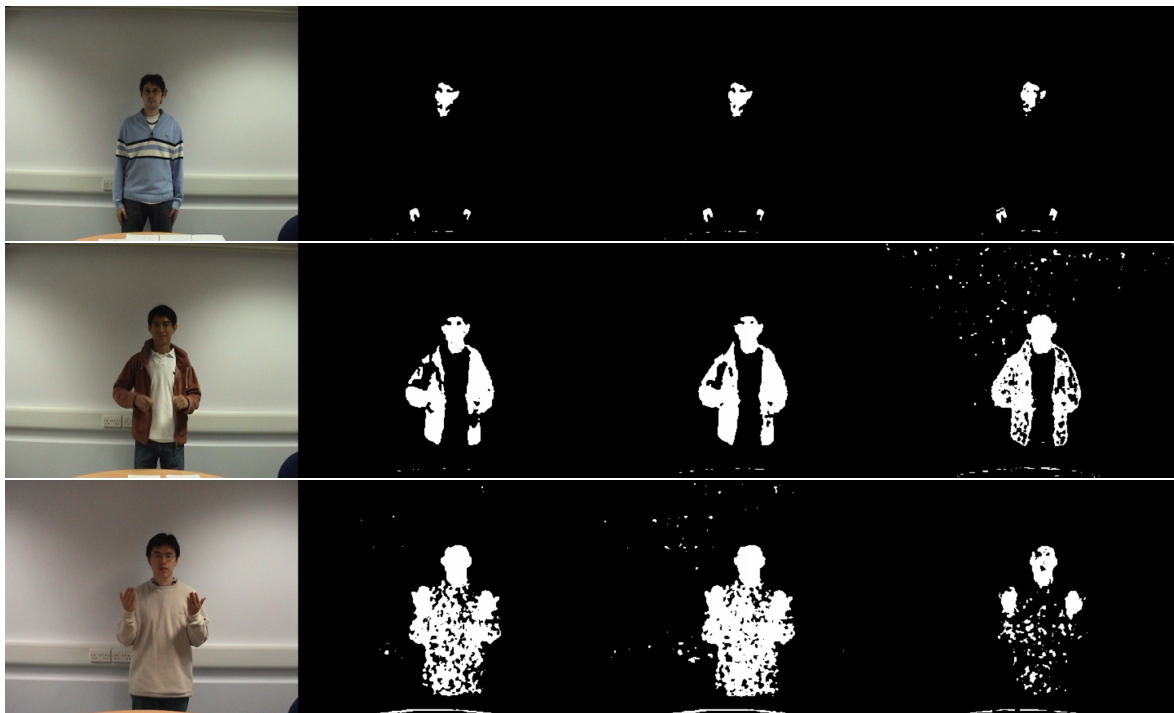
#### 10.3.1 Database

All experiments are based on the newly created VISNET II Cash Machine Database (see section A.3.1 for more details) that contains typical movements and gestures expected for an intelligent multimedia user interface such as the cash machine application. A sample for each of the 10 users in the database can be seen in figure 10.4.

#### 10.3.2 Skin detection

Skin detection can be evaluated as a binary segmentation problem (see section B.5). Since manual segmentation of an image is a very time consuming process, the different skin detection approaches were compared only subjectively.

Figure 10.5 shows representative samples for the users 1,3,10 from the database. The first column shows the original frame and remaining columns show the results of the different skin detection approaches (general, hybrid, specific). While the approaches perform



**Figure 10.5:** Comparison of different skin detection approaches (general, hybrid, specific) based on visual examples.

equally well for the easy example in the first row, and equally bad for the very difficult example in the second row, the specific approach achieves a much better performance for the moderately difficult example in the third row. In the first example the specific approach is slightly more sensitive than the generic and the hybrid approach, which leads to a slight under-segmentation of the skin regions. In the third example the over-segmentation of the skin region can be greatly reduced since the skin color of the user differs from its skin colored clothes. Nevertheless, the specific approach does not provide a solution for the second example, where the skin color of the user is too similar to its skin colored clothes.

### 10.3.3 Face recognition

The face recognition has been evaluated on a subset of the whole database, by considering only 10 videos (one for each person) from the scenario subset. Initially identification was chosen as the operation mode, but it can be easily extended to verification, which is the more typical for the cash machine application scenario. Using a holdout validation procedure recognition rates (see section B.4) of 95% and 100% have been obtained for frame level and track level identification, respectively.

Approach	1	2	3	4	5	6	7	8	9	10	Average
General	0.79	0.64	–	–	0.71	–	–	1.00	0.91	–	0.41
Specific	0.82	0.72	0.36	0.85	0.57	–	0.82	1.00	0.93	0.79	0.69

**Table 10.2:** Comparison of different gesture recognition approaches (general, user specific) based on the recognition rate.

### 10.3.4 Gesture Recognition

The gesture recognition has been evaluated on the *individual gesture* subset of the database and has been treated as a recognition problem (see section B.4). From the 3 samples per user and command, 2 samples were used for training and 1 sample was used for testing of the gesture recognition module.

Table 10.2 compares the results for the generic approach that uses both generic skin detection and gesture recognition models and a system that is based on specific skin detection and gesture recognition models. The recognition rates are provided for each of the users individually and averaged across them. Due to the unreliable performance of the skin detection which lead to failure of the hand tracking, the gesture recognition could not be applied for some of the users (marked with –). While the hand tracking based on the generic skin model worked reliably for only 5 users (1,2,5,8,9), with the specific skin color model the hand tracking worked for all users apart from user 6, because of the shirt with the short sleeves. Furthermore, the performance of the gesture recognition could be improved through specific gesture models for all users apart from user 5. Considering all the users the average recognition rate is improved by 28% from the generic to the specific approach.

## 10.4 Conclusion

### 10.4.1 Summary

This chapter presents a system for personalized human computer interaction (HCI) that can be used as the visual part of the environmental analysis module within an intelligent cash machine. The system detects and tracks persons, their hands and faces within an environment and recognizes their identities and gestures. The focus of this work was on the tight integration of the individual modules to improve the overall performance of the system. This is achieved considering the identity of the user to create user specific skin and gesture models. The user specific skin model leads to a more robust hand tracking in the presence of skin colored clothes. Together with the user specific gesture model which decreases the inter signer variations, a promising performance gain can be achieved for the gesture recognition step.

### 10.4.2 Future work

Although the current system delivers already a promising performance, there are several directions for further improvements.

The hand detection is still not reliable enough in the presence of sleeveless shirts and if the skin color of a person is too similar too its clothes. In these cases skin detection may be used to define regions of interest to which shape based hand detectors are applied.

Currently, a very simple person detection and tracking approach is used since within the current application scenario there is usually only on person present. For more complicated scenarios, more sophisticated approaches that can handle occlusions between users are required.





# Chapter 11

## Conclusion

### 11.1 Summary

The major goal of this work was to develop a hierarchical framework for the visual detection, description and recognition of humans. Thereby the focus was solely laid on appearance instead of motion cues, since they can be easily extracted from images and videos. The framework is largely inspired by the human visual perception and considers different channels of information (body, face) which are analyzed at various granularities (holistic, components). It is based on up-to-date techniques from several research fields such as image processing, machine learning, information fusion and graph theory which are combined in an original way to solve the individual tasks that constitute the complete framework. Finally, the framework was applied to several application scenarios to show its versatility.

The thesis started with a review of the most important application scenarios in the looking at people domain to illustrate the diversity of interests and approaches. For each of the applications it provided a comprehensive overview of the individual goals, challenges and conditions. Furthermore, it discussed the specific channels and characteristics that are considered.

Since the looking at people domain relies on techniques from several research fields including image processing, machine learning, information fusion and graph theory, fundamental techniques have been reviewed in concise way and links to the most important references are provided. Furthermore, the individual techniques are linked to the considered tasks where they are used.

In order to combine the application dependent approaches into a more generic approach a hierarchical framework for human analysis is proposed that has analogies to the scale space theory and focus of attention modeling. The basic idea is to detect and describe the appearance of humans at multiple levels including different channels (e.g. body and face), representations (e.g. holistic, components) and features (e.g. color, texture). In that way, both the detection and the description can be easily adapted according to internal (e.g. application specific interests) and external (e.g. environmental conditions) criteria. Further-

more, a hierarchical human model has been derived from anthropometrical models and the human visual perception.

For body recognition, an appearance based method has been developed that integrates a holistic and a component based representation as well as a large variety of color and texture features. Furthermore, complementary features and parts are fused by applying post mapping fusion in combination with feature selection. The experiments have shown that the proposed method works reliably in the presence of out of plane rotations. Furthermore, they have provided interesting insights regarding the suitability of the different representations and features as well as the tradeoff between a simple representation and complex visual features and vice versa.

For face detection, a novel component based method has been developed that can detect faces robustly even in the presence of partial occlusions. Due to the original approach it can provide additional occlusion information that can be used to improve the reliability of subsequent face analysis steps. Experiments have shown the superior performance in comparison to a holistic state-of-the-art method. Furthermore, the limits of both approaches with respect to the different challenges have been explored.

For face recognition, appearance based methods have been analyzed and extended to improve their performance in the presence of partial occlusions. Different representations (holistic, components and lophoscopic) have been considered and the fusion of multiple experts has been explored. The additional occlusion information from the face detection method has been used to improve the performance by selecting a subset of reliable experts for the fusion. Extensive experiments have shown considerable performance improvement using the adaptive fusion across a large variety of realistic occlusions.

An original system for multimodal person search and retrieval has been developed that combines face and voice analysis with content based search and retrieval techniques. It provides an efficient way to search for people within audiovisual documents such as video blogs, monologues and talks. Within the experiments various aspects of the system have been analyzed including the multimodal fusion between audio and video and the trade off between retrieval performance and required user interaction for the relevance feedback.

Based on the extracted human description an efficient system for visual person search and retrieval based on the query by visual thesaurus paradigm has been developed. Therefore, the original visual region thesaurus has been extended into a human visual thesaurus that distinguishes between different body and face representations. An adapted logical query composition scheme allows to combine these parts and the corresponding categories in a flexible way in order to support a large variety of search interests. For the creation of the thesaurus different visual features and clustering methods have been considered to achieve an optimal grouping of the humans within the database.

The appearance based description of the face and the body has been combined with a motion and shape based description of the hands to build a system for personalized human computer interaction. This system simultaneously recognizes the identities and gestures

of persons for advanced human computer interaction within an intelligent cash machine scenario. Information provided by the face analysis such as position and identity of faces improves the performance of the gesture recognition considerably.

In the following the general findings are summarized:

- So far human visual analysis including detection, description and recognition has been largely influenced by certain application scenarios.
- A universal human analysis framework needs to consider different channels (body, face, hands) and features (color, texture, shape, motion) to provide the appropriate scale based on external and internal criteria.
- Both holistic and component based approaches have distinct advantages and disadvantages and should be used in a complementary way.
- Holistic approaches are more suitable for lower resolutions and qualities and have a lower complexity
- Component based approaches are more complex but offer more flexibility with respect to internal and external criteria
- Combination of bottom-up visual processing and top-down modeling provides a good tradeoff between generality and reliability
- Search and retrieval performance can be improved by automatic grouping of similar objects, integration of the user through relevance feedback and multimodal fusion.

## 11.2 Major contributions

The major contributions and results of this work are shortly summarized below:

- Comprehensive overview of the looking at people domain with its major application scenarios
- Concise review of fundamental techniques from related research fields
- Proposed a generic framework for visual analysis of humans inspired by human visual perception and focus of attention
- Explored different representations, visual features and their fusion for appearance based body recognition
- Developed a component based face detection method for improved performance in the presence of occlusions

- Extended appearance based face recognition approaches with smart expert selection for improved occlusion handling
- Proposed an efficient system for multimodal person search that combines audiovisual analysis and content based retrieval techniques
- Developed an efficient system for visual person search by extending the query by visual thesaurus paradigm towards a human visual thesaurus
- Explored the benefits of personalized human computer interaction for an intelligent cash machine scenario

### 11.3 Outlook

Within this thesis a hierarchical framework for the visual analysis of humans within images and videos has been proposed and several aspects of such a framework have been explored. Nevertheless, there are still several directions to explore, which will be discussed below:

Taking a look at the human analysis framework a lot of aspects have been neglected. With respect to channels only face and body have been considered. For a really universal framework this needs to be extended towards limbs and hands. The focus of this work was laid on appearance based approaches that consider the color and the texture for the analysis. For the description of human behavior or activity motion and shape are required. Finally, since this work has worked mostly on images only the detection and recognition tasks were considered. For videos it is also important to include a reliable tracking that establishes the correspondence between humans and their body parts over time.

Once humans have been detected, tracked and described the question is how this meta-data can be stored in a universal and flexible way. The MPEG-7 content description standard [Manjunath et al., 2002] offers appropriate tools (description schemes and descriptors) to store the extracted data. It supports the hierarchical decomposition of spatial and spatio-temporal objects (humans) into parts and their description with a set of low level descriptors including color, texture, shape and motion characteristics. At the same time it also supports the description of mid and high level concepts such as the identity, gender, ethnicity, role and behavior of a person. The use of MPEG-7 for storing human analysis results has already been proposed for visual surveillance applications [Annesley et al., 2007] but may be extended to a universal description.

With respect to the individual parts of this work future work has already been discussed in detail within the corresponding chapters. Nevertheless, a short summary for each of the parts is provided below.

Although body detection has not been within the scope of thesis, a background segmentation method was used for some of the applications. While these motion based approaches are suitable for videos with static or controlled moving cameras they are not applicable for

images or videos with uncontrolled camera movements. In these cases, model based approaches similar to the developed face detection method are required. Several methods have been proposed for this purpose including the ones proposed by Leibe et al. [2008], Dalal [2006] and Wu and Nevatia [2007b]. The latter one not only detects, but also segments the humans within an image or video frame, which makes it a suitable replacement for the motion based approach which is used so far.

For the appearance based body recognition a rather simple model consisting of head, upper and lower body has been used, in order to provide a description more close to the human visual perception. For analyzing the motion of a human body a more elaborate articulated body model is required such as the one used in [Park et al., 2003]. Nevertheless, these articulated models can be easily mapped to the simplified model for the appearance analysis. One issue with the current model is the ambiguity between the skin and the hair within the head region. Therefore skin detection techniques may be applied to split the head region into these parts and treat them independently. By doing this the skin color may even be used as a soft biometric. Another thing which is interesting with respect to clothes is the detection and description of special features such as logos or appliqué which may help to improve the robustness. For their detection and description feature based approaches that detect and describe salient regions may be used.

Although the face detection approach has been developed with the goal of handling partial occlusions, it works quite well also for other challenges. This shows the higher robustness of a component based approach in comparison to holistic approaches. The same finding has been reported for body detection [Wu et al., 2008]. The major issue for a component based approach is actually to develop a suitable way to describe and consider the spatial arrangement of these components for the detection. This work has shown that structural pattern recognition techniques such as graph matching provide a suitable way to describe the relationship between these components. So the combination of statistical and structural pattern recognition techniques for object detection may be an interesting topic for further research.

The developed face recognition approach is basically an extension of the classical eigenface approach [Turk and Pentland, 1991a] towards different representations and an adaptable fusion step that improves the performance in the presence of occlusions by selecting reliable parts based on available occlusion information. In the mean time several adaptations of this approach have been proposed that extract the feature vector in a different way, exchange the used feature reduction technique or apply other classification methods to improve the performance. These techniques can easily be integrated within the developed approach, while preserving its occlusion awareness.

Apart from the personalized human computer interface all other applications were related to the search and retrieval of humans. For these applications it is very important to have an intuitive and reliable query interface. While query by example provides a good starting point if a representative sample is available, query by visual thesaurus provides and

interesting alternative through the combination of summarization and logical query composition. Both can be used as the starting point for the interactive search based on relevance feedback which may help to improve the performance considerably. Another interesting way that can be further explored is the combination of complementary information through multimodal fusion in order to improve the performance of the analysis task, as it has been shown for the multimodal person search. Finally, integration of the different analysis tasks within the personalized human computer interface has shown that the performance of usually independent modules can be greatly improved by exchanging relevant information.

Given such universal person analysis framework machines will be able to detect, track and recognize people and their behavior within an environment at different scales and react appropriately depending on application-specific or environment-dependent criteria.

# Appendix A

## Database overview

### A.1 Introduction

Like in any other research field comprehensive databases are the key for the development and evaluation of looking at people technologies. In order to develop robust methods they have to be as realistic as possible and include all possible challenges one may face in relevant application scenarios. Furthermore, for the comparative evaluation of different algorithms usually some sort of human annotated ground truth is required to compare it to the machine generated predictions. Although the acquisition and annotation of a high quality database is a very resource intensive task, the availability of publicly available databases is crucial for the advancement of the field.

**Modalities:** Databases may provide different modalities such image, video, or multimodal data. *Image* databases are usually used for appearance based analysis. In addition to that, *video* databases also support motion based analysis. *Multimodal* databases usually in the form of video and audio data may be used for joint audiovisual analysis.

**Channels:** Traditionally looking at people research has concentrated on individual channels due to specific application scenarios and sensor capabilities. The majority of available databases considers *faces* of humans. In order to provide faces in high resolutions short-distance or head-and-shoulder views are common among these databases. These databases are only suitable for face analysis such as face and expression recognition. Another group of databases focuses on the *body* of a human and provides a mid-distance views of the whole body. These databases are usually used for human motion analysis such as behavior and gait recognition. If the resolution is high enough they are also suitable for face analysis. The last group treats humans similar to general *objects* and usually provides a long-distance or wide angle view of an environment and the humans within. Since the resolution of the individual humans is usually quite low, these databases are only useful for object analysis such as object tracking and activity analysis.

**Scenario:** As already mentioned before, research within the looking at people research domain has concentrated on several application scenarios. For each of them databases for the development and evaluation have been created. In the following review we consider *biometrics* (BIO), *retrieval* (RET), *surveillance* (SUR), *smart room technologies* (SRT) and *human computer interaction* (HCI).

While it is quite difficult to get a complete overview of all publicly available databases for looking at people research, comprehensive surveys for some areas have been published. For face analysis the most comprehensive survey is given by Gross [2005], which reviews 27 face databases for face detection, face recognition and facial expression analysis.

The scope of this chapter is to provide a comprehensive survey of available databases for looking at people research and describe the databases that have been used throughout this work in more detail. The databases have been grouped according to their modality and each of the following sections will focus on one of them.

## A.2 Image databases

Most of the available image databases that are suitable for looking at people research, focus on the biometric scenario. Table A.1 provides an overview of available databases with the considered scenario and channel. The databases that have been used within this work are highlighted and will be described in more detail below.

### A.2.1 Neckermann Database

The Neckermann Database has been created in 2005 within the scope of this work. It consists of two sets of images that have been collected from the Neckermann online fashion store<sup>1</sup>.

It consists of two sets of color images with 42 and 100 images, respectively. All images have a common height of 480 pixels and contain a single human in mid-distance view. The database contains a large variety of costumes with different colors, textures and shapes. Some samples of the first set are shown in figure A.1

Originally, this database lacked any additional data. Binary foreground masks for the whole body have been created that serve as ground truth for body detection. Additional region maps for the individual body parts (head, upper body, lower body) have been created that can be used to evaluate body segmentation methods.

The first set of the Neckermann Database has been used for development and evaluation of the representation step within the body recognition module (see chapter 5).

---

<sup>1</sup><http://www.neckermann.de/>



Title	Scenario	Channel	URL
Free Character Database	RET	Humans	
Neckermann Database	RET	Humans	
AR Face Database	BIO	Faces	<a href="http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html">http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html</a>
BioId Database	BIO	Faces	<a href="http://www.Bodycan.de/support/downloads/facedb.php">http://www.Bodycan.de/support/downloads/facedb.php</a>
Caltech Face Database	BIO	Faces	<a href="http://www.vision.caltech.edu/html-files/archive.html">http://www.vision.caltech.edu/html-files/archive.html</a>
CAS-PEAL Database	BIO	Faces	<a href="http://www.jdl.ac.cn/peal/index.html">http://www.jdl.ac.cn/peal/index.html</a>
CMU Hyperspectral	BIO	Faces	<a href="http://www.ri.cmu.edu/pubs/pub_4110.html">http://www.ri.cmu.edu/pubs/pub_4110.html</a>
CMU PIE Database	BIO	Faces	<a href="http://www.ri.cmu.edu/projects/project418.html">http://www.ri.cmu.edu/projects/project418.html</a>
Cohn-Kanade Facial Expression Database	BIO	Faces	<a href="http://vasc.ri.cmu.edu/idb/html/face/facialexpression/index.html">http://vasc.ri.cmu.edu/idb/html/face/facialexpression/index.html</a>
CVL Face Database	BIO	Faces	<a href="http://lrv.fri.uni-lj.si/facedb.html">http://lrv.fri.uni-lj.si/facedb.html</a>
Equinox Infrared Face Database	BIO	Faces	<a href="http://www.equinoxsensors.com/products/HID.html">http://www.equinoxsensors.com/products/HID.html</a>
FERET Color Database	BIO	Faces	<a href="http://www.nist.gov/humanid/colorferet/">http://www.nist.gov/humanid/colorferet/</a>
FERET Database	BIO	Faces	<a href="http://www.nist.gov/humanid/feret/">http://www.nist.gov/humanid/feret/</a>
HRL Face Database	BIO	Faces	<a href="ftp://cvc.yale.edu/CVC/pub/images/hrlfaces">ftp://cvc.yale.edu/CVC/pub/images/hrlfaces</a>
HumanID Database	BIO	Faces	<a href="http://www.nd.edu/~cvrl/HID-data.html">http://www.nd.edu/~cvrl/HID-data.html</a>
JAFFE Database	BIO	Faces	<a href="http://www.mis.atr.co.jp/~mlyons/jaffe.html">http://www.mis.atr.co.jp/~mlyons/jaffe.html</a>
Korean Face Database	BIO	Faces	
MIT CBCL Face Database #1	BIO	Faces	
MIT Face Database	BIO	Faces	<a href="ftp://whitechapel.media.mit.edu/pub/images/">ftp://whitechapel.media.mit.edu/pub/images/</a>
MPIBC Face Database	BIO	Faces	<a href="http://faces.kyb.tuebingen.mpg.de/">http://faces.kyb.tuebingen.mpg.de/</a>
NIST MID Database	BIO	Faces	<a href="http://www.nist.gov/RETD/nistsd18.htm">http://www.nist.gov/RETD/nistsd18.htm</a>
ORL Face Database	BIO	Faces	<a href="http://www.uk.reRET.att.com/facedatabase.html">http://www.uk.reRET.att.com/facedatabase.html</a>
Shimon Edelman Face Database	BIO	Faces	<a href="ftp://ftp.wisdom.weizmann.ac.il/pub/FaceBase/">ftp://ftp.wisdom.weizmann.ac.il/pub/FaceBase/</a>
UMIST Face Database	BIO	Faces	<a href="http://images.ee.umist.ac.uk/danny/database.html">http://images.ee.umist.ac.uk/danny/database.html</a>
University of Maryland Database	BIO	Faces	<a href="http://www.umiacs.umd.edu/users/yaser/DATA/index.html">http://www.umiacs.umd.edu/users/yaser/DATA/index.html</a>
University of Oulu Face Database	BIO	Faces	<a href="http://www.ee.oulu.fi/reRET/imag/color/pbfd.html">http://www.ee.oulu.fi/reRET/imag/color/pbfd.html</a>
UPC Face Database	BIO	Faces	<a href="http://gps-tsc.upc.es/GTAV/ReRETAreas/UPCFaceDatabase/GTAVFaceDatabase.htm">http://gps-tsc.upc.es/GTAV/ReRETAreas/UPCFaceDatabase/GTAVFaceDatabase.htm</a>
VISNET Face Database	BIO	Faces	<a href="http://www.visnet-noe.org/">http://www.visnet-noe.org/</a>
Yale Face Database	BIO	Faces	<a href="ftp://plucky.cs.yale.edu/CVC/pub/images/yalefaces/">ftp://plucky.cs.yale.edu/CVC/pub/images/yalefaces/</a>
Yale Face Database B	BIO	Faces	<a href="http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html">http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html</a>

**Table A.1:** Overview of available image databases for looking at people research with considered scenario and channel.

## A.2.2 Free Character Database

The Free Character Database was created in 2007 within the scope of this work. It consists of images collected from Got3d<sup>2</sup>, a online store for 3D models and real textures. More specifically, it is one of the sets that is available for free.

It consists of 216 color images captured in two sessions with resolutions of 1920x2560

<sup>2</sup><http://www.got3d.com/>



**Figure A.1:** Visual samples of the Neckermann Database.



**Figure A.2:** Visual samples of the Free Character Database.

and 2848x4288 pixels, respectively. The images contain single humans in mid distance view in front of a uniform background. The database contains 22 identities with 54 costumes (from sportive over casual to business) in 4 different views (front, left, right, rear). Figure A.2 provides some visual samples of the database.

Originally the database did not contain any additional data. Manually annotated labels include the identities, costumes, views and sessions. Furthermore, pixel-wise segmentation maps were created for the whole, the upper and the lower body as well as for the head. For the frontal views facial points (left eye, right eye, nose and mouth) were annotated.

The Free Character Database has been used for the development and evaluation of the body recognition module (see chapter 5) and the visual person search application (see chapter 9).

### A.2.3 AR Face Database

The AR Face Database<sup>3</sup> was created Martinez and Benavente [1998].

It consists of 4896 color images with a resolution of 786x576 pixels. The images contain single faces in frontal view in front of a uniform background. In contrast to what is mentioned on the website, the database comprises 136 subjects (76 male, 69 female) with 26 images per subject. Variations include 4 illuminations (neutral, left, right, frontal), 3 occlu-

<sup>3</sup>[http://cobweb.ecn.purdue.edu/~aleix/aleix\\_face\\_DB.html](http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html)



**Figure A.3:** Visual samples of the AR Face Database.



**Figure A.4:** Visual samples of the UMIST Face Database.

sions (none, glasses, scarf), 2 sessions (14 days difference) and 4 expressions (neutral, happy, angry, tired). Some visual samples are provided in figure A.3.

Originally the database did not contain any additional information apart from labels that can be extracted from the file names including gender, identity, variation. The FGNET project further provided a markup<sup>4</sup> of 22 facial features for the variations (1,2,3,5). For this work the annotation was extended with labels for the different types of variation and a markup of 4 facial features (right eye, left eye, nose, mouth) for the variations (1,8,11,14,21,24).

The AR Face Database has been used for the development and evaluation of the face detection and face recognition modules, described in chapter 6 and 7, respectively.

#### A.2.4 UMIST Face Database

The UMIST Face Database (now Sheffield Face Database)<sup>5</sup> was developed by Graham and Allinson [1998] in 1998.

It consists of 564 grayscale images of approximately 220x220 pixels resolution. The images contain single faces of 20 subjects in a range of poses between frontal and profile view. It is important to note that the number of images per subject varies quite considerably between 24 and 84.

Apart from the identities which can be derived from the folder structure no additional information is available. Thus this database has only been used for subjective assessment.

Within this work the UMIST Database has been used to assess the robustness of the

<sup>4</sup>[http://www-prima.inrialpes.fr/FGnet/data/05-ARFace/tarfd\\_markup.html](http://www-prima.inrialpes.fr/FGnet/data/05-ARFace/tarfd_markup.html)

<sup>5</sup><http://www.shef.ac.uk/eee/research/vie/research/face.html>



**Figure A.5:** Visual samples of the VISNET II Face Database.

face detection module (see chapter 6) with respect to out-of-plane rotations and different resolutions.

### A.2.5 VISNET II Face Database

Due to the lack of suitable face databases with realistic occlusions the VISNET II Face Database has been created in 2007 within the scope of the European Network of Excellence VISNET II<sup>6</sup>.

It consists of 4070 color images of 720x576 pixels resolution that were captured in an office environment with standard illumination and frontal pose. The database comprises 37 subjects (34 male, 3 female). Variations are restricted to 3 expressions, 7 occlusions and 1 non occlusion. For each of the variations 10 images have been captured to support the use of different training and testing sets. Figure A.5 provides some visual samples.

Several annotations were added to the database, including labels for the different identities, variations, and samples. Furthermore, a markup of 4 facial features (left eye, right eye, nose and mouth) is provided together with an occlusion flag that marks this part as occluded or not.

The VISNET II Face Database has been used for the development and evaluation of the face detection and face recognition modules, described in chapter 6 and 7.

## A.3 Video databases

Most of the available video databases that are suitable for looking at people research focus on the surveillance scenario. Table A.2 provides an overview of available databases with the considered scenario and channel. The databases that have been used within this work are highlighted and will be described in more detail below.

<sup>6</sup><http://www.visnet-noe.org/>

Title	Scenario	Channel	URL
NLPR Gait Database	BIO	Humans	<a href="http://nlpr-web.ia.ac.cn/english/irds/gaitdatabase.htm">http://nlpr-web.ia.ac.cn/english/irds/gaitdatabase.htm</a>
NUE SRT Database	SRT	Humans	<a href="http://www.nue.tu-berlin.de/">http://www.nue.tu-berlin.de/</a>
VISNET Cash Machine Database	HCI	Humans	<a href="http://www.visnet-noe.org/">http://www.visnet-noe.org/</a>
ASTAR Video Dataset	SUR	Objects	<a href="http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html">http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html</a>
ATON Video Dataset	SUR	Objects	<a href="http://cvrr.ucsd.edu/aton/shadow/">http://cvrr.ucsd.edu/aton/shadow/</a>
AVSS 2005 Dataset	SUR	Objects	<a href="http://www-dsp.elet.polimi.it/avss2005">http://www-dsp.elet.polimi.it/avss2005</a>
AVSS 2007 Dataset	SUR	Objects	<a href="http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html">http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html</a>
CAVIAR Test Case Scenarios	SUR	Objects	<a href="http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/">http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/</a>
ITEA CANDELA Dataset	SUR	Objects	<a href="http://www.hitech-projects.com/euprojects/candela/">http://www.hitech-projects.com/euprojects/candela/</a>
PETS 2000 Dataset	SUR	Objects	<a href="http://ftp.pets.rdg.ac.uk/cs/PETS2000/">http://ftp.pets.rdg.ac.uk/cs/PETS2000/</a>
PETS 2001 Dataset	SUR	Objects	<a href="http://www.cvg.cs.rdg.ac.uk/PETS2001/">http://www.cvg.cs.rdg.ac.uk/PETS2001/</a>
PETS 2002 Dataset	SUR	Objects	<a href="http://www.cvg.cs.rdg.ac.uk/PETS2002/">http://www.cvg.cs.rdg.ac.uk/PETS2002/</a>
PETS 2005 Dataset	SUR	Objects	<a href="http://www.cvg.cs.rdg.ac.uk/PETS2005/">http://www.cvg.cs.rdg.ac.uk/PETS2005/</a>
PETS 2006 Dataset	SUR	Objects	<a href="http://www.cvg.rdg.ac.uk/PETS2006/data.html">http://www.cvg.rdg.ac.uk/PETS2006/data.html</a>
PETS-ECCV 2004 Dataset	SUR	Objects	<a href="http://www-prima.inrialpes.fr/PETS04/caviar_data.html">http://www-prima.inrialpes.fr/PETS04/caviar_data.html</a>
PETS-ICVS 2003 Dataset	SUR	Objects	<a href="http://www-prima.inrialpes.fr/FGnet/data/08-Pets2003/data/">http://www-prima.inrialpes.fr/FGnet/data/08-Pets2003/data/</a>
PETS-RISA Dataset	SUR	Objects	<a href="http://ftp.pets.rdg.ac.uk/cs/PETS-RISA/">http://ftp.pets.rdg.ac.uk/cs/PETS-RISA/</a>
VS-PETS 2003 Dataset	SUR	Objects	
VSSN 2006 Dataset	SUR	Objects	<a href="http://mmc36.informatik.uni-augsburg.de/VSSN06_OSAC/">http://mmc36.informatik.uni-augsburg.de/VSSN06_OSAC/</a>

**Table A.2:** Overview of available video databases for looking at people research with considered scenario and channel.

### A.3.1 VISNET II Cash Machine Database

The VISNET II Cash Machine Database was developed in 2007 within the European Network of Excellence VISNET II<sup>7</sup> for the development of an intelligent cash machine.

The databases consists of two sets of videos with a common resolution of 360x288 pixels and a framerate of 25 fps. The first set comprises 420 videos of 14 *individual gestures*, 10 subjects and 3 samples each and may be used to train and evaluate gesture recognition for individual gestures. The second set contains 150 videos of 5 *scenarios* consisting of multiple gestures, 10 subjects and 3 samples each. It presents a more realistic scenario and may be used to evaluate gesture recognition for continuous gestures. A sample for each of the users in the database can be seen in figure A.6.

The database has been annotated with labels regarding the identity of each person, and the present gestures. Furthermore, the body and other interesting body parts (face, hands) have been annotated in form of bounding boxes for every 25th frame. This allows not only to assess the performance of face and gesture recognition but also of detection and tracking of the whole body and individual body parts.

The VISNET II Cash Machine Database has been used for the development and the evaluation of the personalized human computer interaction framework described in chapter 10.

<sup>7</sup><http://www.visnet-noe.org/>





**Figure A.6:** Visual samples of the VISNET II Cash Machine Database.

Title	Scenario	Channel	URL
AMI Corpus	RET	Humans	<a href="http://corpus.amiproject.org/">http://corpus.amiproject.org/</a>
M4 Meeting Corpus	RET	Humans	<a href="http://www.idiap.ch/mmm/corpora/m4-corpus">http://www.idiap.ch/mmm/corpora/m4-corpus</a>
BANCA Database	BIO	Faces	<a href="http://www.ee.surrey.ac.uk/ReRET/VSSP/banca/">http://www.ee.surrey.ac.uk/ReRET/VSSP/banca/</a>
CUAVE Database	BIO	Faces	<a href="http://www.ece.clemson.edu/speech/cuave.htm">http://www.ece.clemson.edu/speech/cuave.htm</a>
VALID Database	BIO	Faces	<a href="http://ee.ucd.ie/validdb/">http://ee.ucd.ie/validdb/</a>
XM2VTS Database	BIO	Faces	<a href="http://www.ee.surrey.ac.uk/ReRET/VSSP/xm2vtsdb/">http://www.ee.surrey.ac.uk/ReRET/VSSP/xm2vtsdb/</a>

**Table A.3:** Overview of available multimodal (audiovisual) databases for looking at people research with considered scenario and channel.

## A.4 Multimodal databases

Most of the available multimodal databases suitable for looking at people research focus on the biometric scenario. Table A.3 provides an overview of available databases with the considered scenario and channel. The databases that have been used within this work are highlighted and will be described in more detail below.

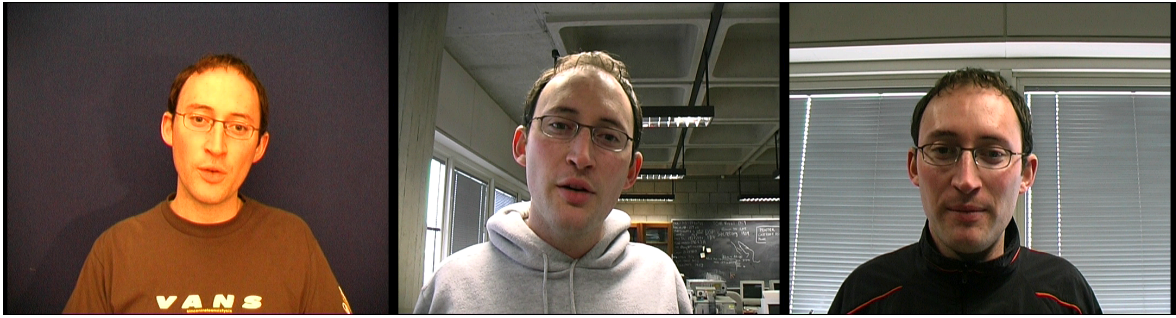
### A.4.1 VALID Database

The VALID Database<sup>8</sup> was primarily created for the development and evaluation of multimodal biometric systems.

It consists of 1060 audiovisual sequences containing individual persons in head and shoulder view either saying a short sentence or counting numbers. Each of the 106 individuals (27 female, 79 male) is captured in 5 environments (1 studio, 4 office), leading to 10 files for each of them. Both the acoustical (noise, reverberation) and visual characteristics (illumination, background) of the environments are quite diverse, making the data even more realistic for the given application scenario. Figure A.7 shows some visual samples of the same individual within the different environments.

Originally the database did not contain any annotation apart from the identity, the environment and the activity that can be derived from the file name. The same markup as before

<sup>8</sup><http://ee.ucd.ie/validdb/>



**Figure A.7:** Visual samples of the VALID Database.

of 4 facial features (right eye, left eye, nose, mouth) has been added to the images to support the evaluation of face detection methods.

Within this work the VALID Database has been used for the development and evaluation of the multimodal person search application described in chapter 8.





## Appendix B

# Evaluation methodologies

The goal of evaluation is to assess the performance of a method that achieves a certain task and make it comparable to other methods. With respect to machine learning the evaluation may consider the following criteria

- Accuracy with respect to human perception
- Efficiency in creating and applying the model
- Robustness regarding noise and missing observations
- Scalability for a large variety of conditions
- Interpretability provided by the model
- Compactness of the model or rules

Although other criteria have been also considered in this work, the major focus is laid on the accuracy of the developed methods with respect to a human-defined ground truth.

Several tasks have been addressed within this work including segmentation, detection, recognition, retrieval and clustering. Since the evaluation methods for these tasks share some common ideas and tools, these will be discussed first and the specifics of the evaluation for each tasks are discussed later.

### B.1 Confusion matrix

A confusion matrix [Provost et al., 1998] is a common tool for the evaluation of unsupervised and supervised learning tasks. It provides a comparison of ground truth and predicted classes (labels) against each other.

Depending on the task one can distinguish between unary or binary and n-ary confusion matrices which are shown figure B.1 along with the different numbers that can be derived from it:

		Prediction (PR)	
		P	N
Ground truth (GT)	P	TP	FN
	N	FP	TN

(a)

		Prediction (PR)			
			1	2	3
Ground truth (GT)	1	TN	TN	FP	TN
	2	TN	TN	FP	TN
	3	FN	FN	TP	FN
	4	TN	TN	FP	TN

(b)

**Figure B.1:** Confusion matrix as a common tool for unsupervised and supervised learning: (a) Unary/binary case, and (b) n-ary case.

**True positives (TP):** The number of correct predictions that an observation is positive.

**True negatives (TN):** The number of correct predictions that an observation is negative.

**False positives (FP):** The number of false predictions that an observation is positive. Also referred to as type I error in statistics.

**False negatives (FN):** The number of false predictions that an observation is negative. Also referred to as type II error in statistics.

Furthermore, one can distinguish between the number of true (correct) predictions  $T$  that correspond to the sum of the elements along the diagonal and the number of false (incorrect) predictions  $F$  that correspond to the sum of the elements in the upper and lower triangular part of the confusion matrix.

The major difference between unary and binary classification problems is that for the former  $TN$  is usually not very well defined and usually discarded from the evaluation.

## B.2 Detection evaluation

The evaluation of any detection problem is usually based on a set of ground truth  $G = \{g_i : i = 1, \dots, n\}$  and a set of predicted objects  $P = \{p_j : j = 1, \dots, m\}$ , which may contain different number of objects.

In the first step a  $(n \times m)$ -distance matrix  $D$  is derived, that compares the ground truth and the predicted objects according to some object-specific criterion. It is usually based on the position and the size of the objects, but also the orientation can be considered. Within this work a generic criteria based on the objects bounding boxes is used [Everingham et al.,

2005]. The dissimilarity  $d_{ij}$  between two bounding boxes  $g_i$  and  $p_j$  is defined as

$$d_{ij} = \frac{\text{area}(g_i \cap p_j)}{\text{area}(g_i \cup p_j)} \quad (\text{B.1})$$

and they are considered as a match if  $d_{ij} > 0.5$ .

Based on the distance matrix  $D$  a match matrix  $M$  is derived by solving the one-to-one assignment problem between the ground truth and the predicted objects using the Hungarian algorithm [Munkres, 1957].

From the match matrix  $M$  a confusion matrix is derived by considering the sum of assignments as true positives, the ground truth objects without an assignment as false negatives and the predicted objects without an assignment as false positives.

Given the confusion matrix several detection measures are derived, including the true positive rate (TPR), the false positive rate (FPR), and the false negative rate (FNR) defined as

$$TPR = \frac{TP}{TP + FN} \quad (\text{B.2})$$

$$FPR = \frac{FP}{TN + FP} \quad (\text{B.3})$$

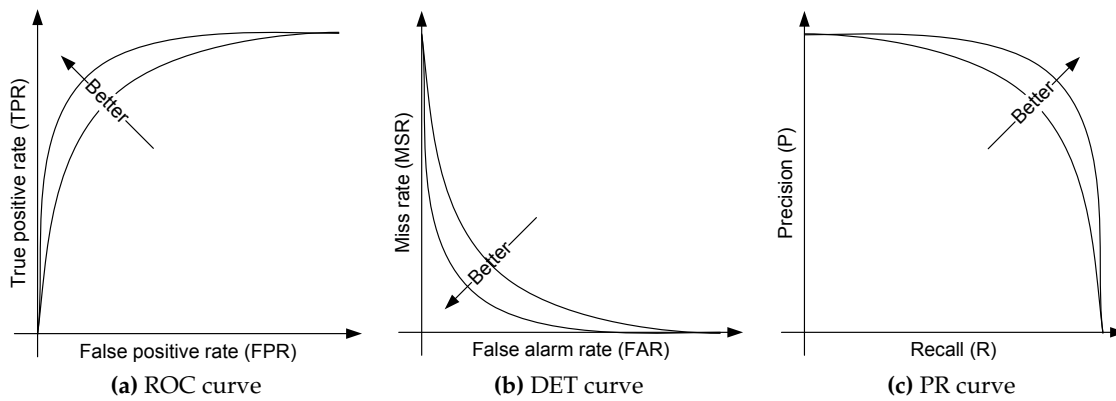
$$FNR = \frac{FN}{TP + FN} \quad (\text{B.4})$$

Detection approaches usually contain some thresholding operation, that either accepts or rejects an image region as the object of interest. By varying the corresponding threshold  $t$  is possible to achieve a tradeoff between the number of false positives and false negatives. A low threshold usually decreases the number of false negatives while increasing the number of false positives and vice versa. For a better comparison across a large spectrum of operating points detection approaches are usually evaluated for a set of thresholds. Each considered threshold leads then to an individual confusion matrix along with the measures defined above. By plotting the the individual measures pair wise two types of curves can be derived.

### B.2.1 Receiver operating characteristic (ROC) curve

ROC curves plot the TPR versus the FPR parametrically as a function of the detection threshold as shown in figure B.2(a). ROC curves are threshold independent, allowing for a performance comparison of different methods under similar conditions or of a single system under different conditions. The closer a curve is to the upper left corner  $(FPR, TPR) = (0, 1)$  the better the performance of the corresponding method.

Within this work ROC curves have been used for face and facial component detection evaluation in chapter 6.



**Figure B.2:** Common evaluation curves for unary/binary classification evaluation such as detection and retrieval tasks.

### B.2.2 Detection error tradeoff (DET) curve

In contrast to ROC curves, DET curves plot error rates on both axes namely FNR versus FPR parametrically as a function of the detection threshold as shown in figure B.2(b). This allows for uniform treatment of both types of errors. The closer the curve is to the lower right corner  $(FPR, FNR) = (0, 0)$  the better the performance of the corresponding method.

## B.3 Retrieval evaluation

Many different measures for information retrieval have been proposed. In any case the goal is usually to find a set of documents that are relevant for the user. This can be either achieved by ranking all available documents according to some criteria or returning a unranked set of documents. Within this work both cases are considered.

Based on a ranked list of available documents and a list of relevant items the evaluation is either based on *precision/recall* or *rank* measures.

### B.3.1 Precision recall (PR) curves

PR curves are comparable to receiver operating characteristic (ROC) and detection error tradeoff (DET) curves, since they consider retrieval as a binary classification problem and plot measures derived from the confusion matrix parametrically as a function of the result set size. In contrast to the two other curves, precision (P) and precision (P) are considered here, defined as

$$R = \frac{TP}{TP + FN} \quad (\text{B.5})$$

$$P = \frac{TP}{TP + FP} \quad (\text{B.6})$$

Similar to the threshold for the detection evaluation (see section B.2) varying the result

set size leads to a tradeoff between recall (R) and P. While a smaller result set size usually leads to a higher precision, a higher recall is usually achieved with a larger result set size as shown in figure B.2(c). The closer the curve is to the upper right corner  $(R, P) = (1, 1)$  the better the performance of the corresponding method.

By assuming equal cost for both types of error the precision and the recall can be combined into the f-measure (F) which is defined as

$$F = \frac{2PR}{P + R} \quad (\text{B.7})$$

As another measure, the average precision (AP)  $\bar{P}$  combines the precision and the ranking of relevant items into a single measure by averaging the precision for the relevant items which can be written as

$$\bar{P} = \frac{1}{N_R} \sum_{i=1}^{N_R} P_i = \frac{1}{N_R} \sum_{i=1}^{N_R} \frac{TP_i}{TP_i + FP_i} \quad (\text{B.8})$$

To get an average precision of 1, the system must retrieve all relevant documents at the top without any non relevant document in between.

Within this work precision, recall and the f-measure have have been used for the evaluation of the multimodal person search system (chapter 8) and the face detection and classification (chapter 6)

### B.3.2 Ranks

Rank measures follow a slightly different approach by directly considering the position of the relevant documents within the ranked list.

A very simple and robust measure is the best rank  $Rk_1$  which is simply defined as the rank of the first relevant image. It is a common measure for text and content based image retrieval.

The normalized average rank (NAR)  $\widetilde{Rk}$  is an extension of the average rank (AR)  $\overline{Rk}$  [Gargi and Kasturi, 1999] proposed by Mueller et al. [1999] to remove its dependency on the collection size  $N$  and the number of relevant items  $N_R$ . It is defined as

$$\widetilde{Rk} = \frac{1}{NN_R} \left( \sum_{i=1}^{N_R} R_i - \frac{N_R(N_R - 1)}{2} \right) \quad (\text{B.9})$$

Within this work rank measures have been used for the evaluation of the multimodal person search system described in chapter 8.

## B.4 Recognition evaluation

Recognition can be evaluated as a supervised learning problem with the goal to predict the class for a given test sample based on a classifier that has been created from several training

samples. One can distinguish between different recognition tasks, including verification, identification and watchlist. While the first is a typical binary classification problem and the second is a n-ary classification problem, the third is a combination of both. Within this work only identification tasks are considered.

Identification tasks can be evaluated in two different ways which are related to each other. While recognition and error rates are based on a certain classifier, cumulative match characteristic (CMC) curves consider the task as a simple matching problem.

#### B.4.1 Recognition and error rate

Given a set of training samples with associated ground truth labels, a classifier is built. It is applied to a set of test samples which leads to a set of predicted labels. These are compared to the ground truth labels of the corresponding samples which leads to a n-ary confusion matrix (see section B.1). From this matrix the number of true  $T$  and false predictions  $F$  can be derived and used to compute the recognition and the error rate defined as

$$RR = \frac{T}{T + F} \quad (B.10)$$

$$ER = \frac{F}{T + F} = 1 - RR \quad (B.11)$$

Within this work the recognition rate has been considered for the face recognition evaluation in chapter 7 and the occlusion classification evaluation in chapter 6.

#### B.4.2 Cumulative match characteristic (CMC) curve

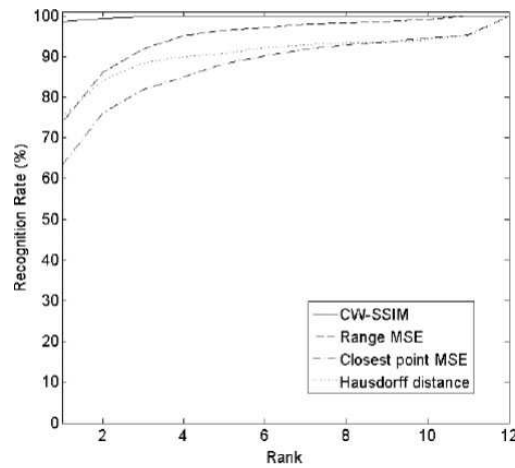
The cumulative match characteristic (CMC) curve [Johnson et al., 2003] matches the testing samples directly against the training samples. For each testing sample the training samples are ranked with increasing distance. Based on the ground truth label of the probe the rank of the corresponding gallery sample is determined. By iterating over the probe samples the probability of identification can be computed across the different ranks and plotted as CMC curve as shown in figure B.3.

Within this work the CMC curve has been used for the evaluation of the body recognition approach in chapter 5.

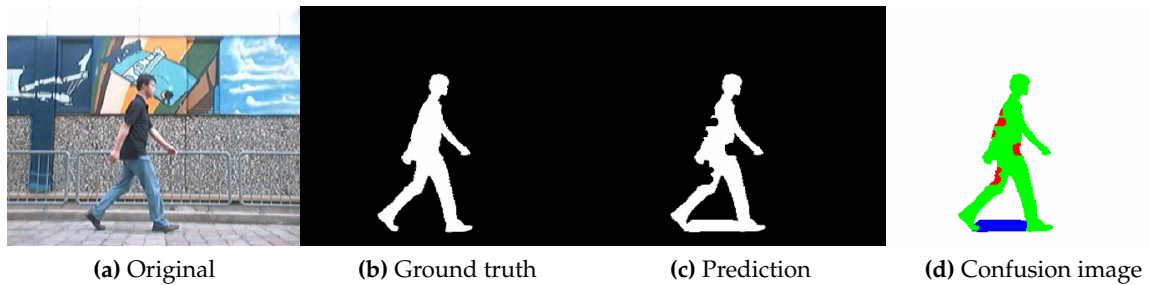
### B.5 Segmentation evaluation

The goal of most image segmentation tasks is to partition an image or image region into a set of non-overlapping regions. Several methodologies have been proposed for the evaluation including pixel based, object based and perceptual metrics [Renno et al., 2006]. Depending on the number of regions one can distinguish between binary and n-ary segmentation.

Within this work a pixel based approach was adopted which can be seen as a pixel-wise classification problem. Given ground truth and predicted segmentation map either a binary



**Figure B.3:** Cumulative match characteristic curve for the evaluation of recognition tasks.



**Figure B.4:** Illustration of the segmentation evaluation process. (a) Original image (b) Binary ground truth mask (c) binary prediction mask (d) confusion image with the true positives (green), false positives (blue), false negatives (red) and true negatives (white).

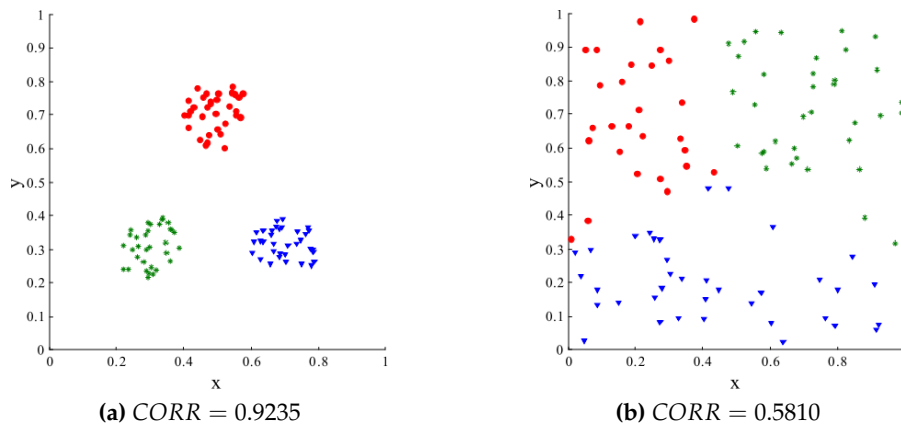
of n-ary confusion matrix (described in section B.1) is built. An example of that process for a binary foreground segmentation problem is illustrated in figure B.4 where the colors of confusion image in figure B.4(d) correspond to the colors of the binary confusion matrix in figure B.1(b).

Based on that confusion matrix one can compute several quality measures. For a binary segmentation problem the precision (P), recall (R) and f-measure (F) from the retrieval evaluation (defined in section B.3.1) are adopted. For n-ary segmentation problems the recognition rate (RR) and the error rate (ER) from the recognition evaluation (defined in section B.4.1) are used.

## B.6 Clustering evaluation

Clustering evaluation deals with measuring the goodness of the partitions produced by a clustering method [Jain et al., 1999] (see section 3.3.4). Measures of cluster validity can be grouped into three different classes [Tan et al., 2005]

**Internal:** Measure the goodness of a clustering by only using the data and clusters with-



**Figure B.5:** Clustering evaluation via correlation [Tan et al., 2005].

out any external information. Well known measure are sum of squared error (SSE), correlation, similarity matrix, cohesion and separation, and the silhouette coefficient.

**External:** Measure the goodness of a clustering by comparing the clusters with externally supplied class labels. Well known measures include entropy and purity.

**Relative:** Measure to compare two different clusterings with each other. Often internal or external measures are used for the comparison.

Within this work internal and external clustering evaluation methods have been used in chapter 9. The used methods are explained in more detail below.

### B.6.1 Internal measures

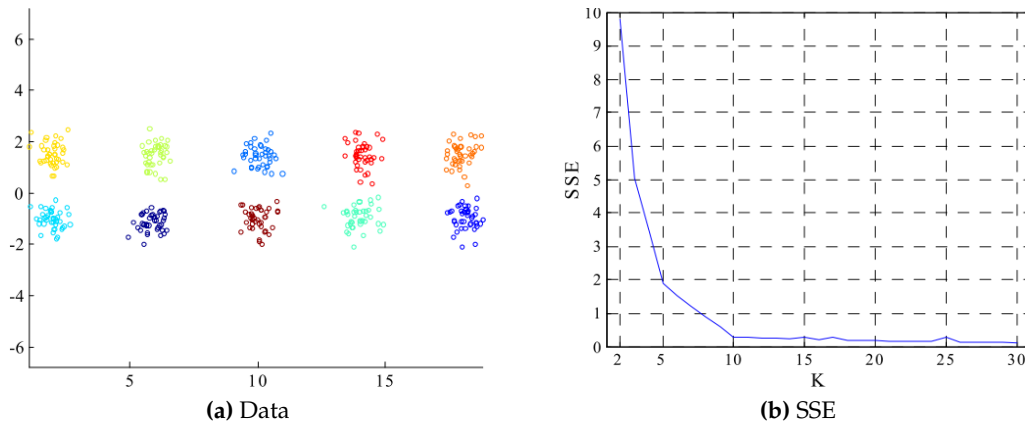
As mentioned before internal measures do not require any external information in form of ground truth class labels for the evaluation. They do only consider the data samples and their predicted cluster labels provided by the clustering method.

#### Correlation

Measuring cluster validity via correlation (CORR) [Tan et al., 2005] is based on two matrices. While the proximity matrix contains the pair wise distances between the data points, the incidence matrix is a binary matrix with its entries equal to 1 if two data points belong to the same cluster. The correlation between these two matrices indicates how close data points are that belong to the same cluster. This is illustrated for an example in figure B.5.

While correlation is a good measure for distance based clusterings, it provides a rather poor evaluation for some density or contiguity based clusterings.





**Figure B.6:** Estimating the optimal number of clusters via the sum of squared errors (SSE) [Tan et al., 2005].

### Cohesion and separation

Another way to measure cluster validity is to consider cohesion (COH) and separation (SEP) [Tan et al., 2005]. While the former measures how closely related are objects within clusters, the second measures how distinct clusters are from each other. Both have been defined in several ways but the following ones are the most commonly used

$$COH = \sum_{c=1}^C \sum_{x \in X_c} (x - \mu_c)^2 \quad (B.12)$$

$$SEP = \sum_{c=1}^C N_c (\mu - \mu_c)^2 \quad (B.13)$$

The cohesion which is equal to the sum of squared error (SSE) can also be used to estimate the optimal number of clusters as it is shown in figure B.6.

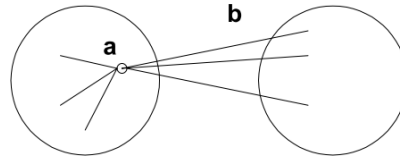
### Silhouette coefficient

The silhouette coefficient (SIL) [Tan et al., 2005] combines the ideas of cohesion and separation but not only for clusterings, but also individual clusters and data points.

For an individual point  $x$  it computes the average distance of  $x$  to the points in its cluster  $a(x)$  and the minimum average distance of  $x$  to points in the other clusters  $b(x)$ . This idea is illustrated in figure B.7. Based on that, the silhouette coefficient which lies in the interval  $[0, 1]$  is defined as

$$s(x) = 1 - a(x)/b(x) \quad (B.14)$$

It is also possible to compute silhouette coefficient for individual clusters or the whole clustering by averaging.



**Figure B.7:** Illustration of the silhouette coefficient for a single data point and two clusters [Tan et al., 2005].

### B.6.2 External measures

In contrast to internal measures, external measures rely on ground truth class labels which are compared to the predicted cluster labels of the corresponding data points. This is comparable to the  $n$ -ary classification evaluation, where a  $(n \times n)$  confusion matrix is built. Since the number of classes  $n$  and the number of clusters  $m$  may not be the same a  $m \times n$  confusion matrix is used here.

#### Class and cluster entropies

Given this confusion matrix  $C = (c_{ij})$  between predicted cluster labels  $i$  and ground truth class labels  $j$ , cluster and class entropies are computed. The former is defined as

$$e_i^m = - \sum_{j=1}^n p_{ij} \log_2 p_{ij} \text{ with } p_{ij}^m = \frac{c_{ij}}{\sum_{i=1}^m c_{ij}} \quad (\text{B.15})$$

while the latter is computed by exchanging the rows and columns of the confusion matrix as

$$e_j^n = - \sum_{i=1}^m p_{ij} \log_2 p_{ij} \text{ with } p_{ij}^n = \frac{c_{ij}}{\sum_{j=1}^n c_{ij}} \quad (\text{B.16})$$

Both can be averaged across the clusters or classes and combined into an overall entropy  $e = r \cdot e^m + (1 - r) \cdot e^n$  based on the ratio  $r$  which is commonly set to 0.5.

#### Purity and coverage

Similar to the cluster and class entropies, the cluster purity (PUR) and the class coverage (COV) are defined as

$$PUR_j = \max_i p_{ij}^m \quad (\text{B.17})$$

$$COV_i = \max_j p_{ij}^n \quad (\text{B.18})$$

with  $p_{ij}^m$  and  $p_{ij}^n$  defined above. Again both can be averaged across the clusters or classes and combined into the overall quality  $QUAL = r \cdot PUR + (1 - r) \cdot COV$  based on the ratio  $r$  which is commonly set to 0.5.

# Bibliography

- E. Aarts. Ambient intelligence drives open innovation. *ACM Interactions*, 12(4):66–68, 2005.
- T. Adamek and N. O'Connor. Using dempster-shafer theory to fuse multiple information sources in region-based segmentation. In *IEEE International Conference on Image Processing (ICIP)*, volume 2, pages 269–272, Sep 2007.
- T. Adamek, N. O'Connor, and N. Murphy. Region based segmentation of images using syntactic visual features. In *WIAMIS 2005 - 6th International Workshop on Image Analysis for Multimedia Interactive Services*, 2005.
- Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):721–732, 1997.
- J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Articulated and elastic non-rigid motion: A review. In *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, 1994.
- A. Alatan and L. Onural. Image sequence analysis for emerging interactive multimedia services -the european cost 211 framework. *IEEE Transactions on circuits and systems for video technology*, 8(7), November 1998.
- J. Annesley, J. Orwell, and J.-P. Renno. Evaluation of mpeg7 color descriptors for visual surveillance retrieval. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pages 105–112, Oct 2005.
- J. Annesley, A. Colombo, J. Orwell, and S. Velastin. A profile of mpeg-7 for visual surveillance. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 482–487, Sept. 2007. doi: 10.1109/AVSS.2007.4425358.
- S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line nonlinear/non-gaussian bayesian tracking. *IEEE TSP*, 2002.
- K. Baek, B. A. Draper, J. R. Beveridge, and K. She. Pca vs. ica: A comparison on the feret data set, 2002.

- W. Bailer, P. Schallauer, H. B. Haraldson, and H. Rehatschek. Optimized mean shift algorithm for color segmentation in image sequences. In *SPIE Image and Video Communications and Processing*, 2005. ISBN 0-8194-5658-6.
- M. Balcells Capilades. An appearance based approach for consistent labeling of humans and objects in videos. *Pattern Analysis and Applications (PAA)*, 2004.
- A. Banerjee and R. N. Dave. Validating clusters using the hopkins statistic. In *IEEE International Conference on Fuzzy Systems*, 2004.
- J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *Int. J. Comput. Vis.*, 12(1):42?77, 1994.
- E. Bart, E. Byvatov, and S. Ullman. View invariant recognition using corresponding object fragments. In *International Conference on Computer Vision (ICCV)*, pages 152–165, 2004.
- M. Bartlett and T. Sejnowski. Independent components of face images: A representation for face recognition. In *4th Annual Joint Symposium on Neural Computation*, 1997.
- M. S. Bartlett, H. M. Lades, and T. J. Sejnowski. Independent component representations for face recognition. *Proceedings of the SPIE*, 1998.
- R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2003.
- P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision (IJCV)*, 1998.
- P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):711–720, 1997.
- R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.
- S. Belongie, J. Malik, and J. Puzicha. Matching shapes. In *Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001. Proceedings*, volume 1, 2001.
- B. Berlin and P. Kay. Basic color terms: Their universality and evolution. University of California Press, 1969.
- D. J. Beymer. Face recognition under varying pose. Technical report, Massachusetts Institute of Technology, 1993.
- J. C. Bezdek. Pattern recognition with fuzzy objective function algorithms. Plenum Press, 1981.

- S. T. Birchfield and S. Rangarajan. Spatiograms versus histograms for region based tracking. In *CVPR*, 2005.
- M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.
- V. Blanz, S. Romdhani, and T. Vetter. Face identification across different poses and illuminations with a 3d morphable model. In *IEEE International Conference on Automatic Face and Gesture Recognition*, page 202–207, 2002.
- N. Boujemaa, J. Fauqueur, and V. Gouet. What’s beyond query by example? In *ICISP*, 2003.
- R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A linguistic feature vector for the visual interpretation of sign language. In *European Conference on Computer Vision (ECCV)*, 2004.
- K. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer Vision and Image Understanding*, 2006.
- D. Brien, editor. *Dictionary of British Sign Language/English*. Faber and Faber, 1992.
- R. Brunelli and T. Poggio. Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 15(10):1042–1052, Oct 1993. doi: 10.1109/34.254061.
- H. Bunke. Graph matching: Theoretical foundations, algorithms, and applications, 2000.
- H. Bunke. Recent advances in structural pattern recognition with applications to visual form analysis. In C. Arcelli, L. Cordella, and G. Sanniti di Baja, editors, *Visual Form 2001*, pages 11–23. Springer Verlag, 2001.
- H. Bunke, S. Günter, and X. Jiang. Towards bridging the gap between statistical and structural pattern recognition. In S. Singh, N. Murshed, and W. Kropatsch, editors, *Advances in Pattern Recognition*, pages 1–11. Springer Verlag, 2002.
- S.-F. Chang, T. Sikora, and A. Puri. Overview of the mpeg-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 11(6):688–695, 2001.
- T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal person recognition using unconstrained audio and video. In *International Conference on Audio and Video Based Biometric Authentication*, 1999.
- R. T. Collins, T. Lipton, A. J. and Kanade, H. Fujiyoshi, D. Duggins, et al. A system for video surveillance and monitoring. Technical Report CMU-RI-TR-00-12, Carnegie Mellon University (CMU), Pittsburgh, USA, 2000.

- A. J. Colmenarez and T. S. Huang. Face detection with information-based maximum discrimination. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997.
- T. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *European Conference on Computer Vision (ECCV)*, 1998.
- T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(6):681–685, 2001.
- I. Craw, H. Ellis, and J. R. Lishman. Automatic extraction of face-feature. *Pattern Recog. Lett.*, 1987.
- M. Crucianu, M. Ferecatu, and N. Boujemaa. Relevance feedback for image retrieval: A short survey. Technical report, INRIA Rocquencourt, 2004.
- R. Cucchiara. Multimedia surveillance systems. In *International Workshop on Video Surveillance and Sensor Networks (VSSN)*, 2005.
- R. Cutler and L. Davis. Robust periodic motion and motion symmetry detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 615–622, 2000.
- L. da Vinci. Vitruvian man, 1487.
- Y. Dai and Y. Nakano. Face texture model based on sgld and its application in face detection in a color scene. *Pattern Recognition (PR)*, 296:1007–1017, 1996.
- N. Dalal. *Finding People in Images and Video Sequences*. thesis, INRIA Rhone-Alpes, 2006.
- N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision (ECCV)*, 2006.
- R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 2008.
- P. Delacourt and C. J. Welekens. Distbic: A speaker-based segmentation for audio data indexing. *Speech Communication*, 32:111–126, 2000.
- T. Deselaers. Features for image retrieval. Master’s thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 2003.
- T. Dietterich. Ensemble methods in machine learning. In F. Kittler, J. Roli, editor, *Multiple Classifier Systems*, volume 1857 of *LNCS*. Springer, 2001.
- R. O. Duda, H. P., and E. Stork. *Pattern Classification*. Wiley, 2 edition, 2002.

- G. J. Edwards, T. F. Cootes, and C. J. Taylor. Face recognition using active appearance models. In *European Conference on Computer Vision (ECCV)*, 1998.
- M. J. Egenhofer. Spatial query by sketch. In *Proceedings of the IEEE Symposium on Visual Languages*, 1996.
- A. Ekin and A. M. Tekalp. Automatic soccer video analysis and summarization. In *EI*, 2003.
- A. M. Elgammal and L. S. Davis. Probabilistic framework for segmenting people under occlusion. In *ICCV*, 2001.
- I. A. Essa. Computers seeing people. *AI Magazine*, 1999.
- M. Everingham, L. V. Gool, C. Williams, and A. Zisserman. The pascal visual object classes challenge results. In *Visual Object Recognition Challenge (VOC)*, Apr 2005.
- L. G. Farkas. *Anthropometry of the Head and Face*. Raven Press, 1994.
- P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.
- R. Feraud, O. Bernier, and D. Collobert. A constrained generative model applied to face detection. *Neural Processing Letters*, 5:73–81, 1997.
- R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.
- R. Fisher. *Hypertext Image Processing Reference*. University of Edinburgh, 2004.
- Y. Freund and R. E. Schapire. A decision theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory: Eurocolt 95*. Springer, 1995.
- B. V. Funt and G. D. Finlayson. Color constant color indexing. *TPAMI*, 1995.
- E. Galmar and B. Huet. Graph based spatio temporal region extraction. In *International Conference for Image Analysis and Recognition*, 2006.
- U. Gargi and R. Kasturi. Image database querying using a multi-scale localized color representation. *CBA*, 1999.
- F. Ge, S. Wang, and T. Liu. Evaluating edge detection through boundary detection. *EURASIP Journal of Applied Signal Processing*, pages 1–15, 2005.
- A. S. Georgiades, P. N. Belhumeur, , and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23:643–660, 2001.

- N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Computer Society Conference Computer Vision and Pattern Recognition*, 2006.
- R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 3 edition, 2007.
- C. C. Gotlieb and H. E. Kreyzig. Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics and Image Processing*, 1990.
- D. B. Graham and N. M. Allinson. Characterizing virtual eigensignatures for general purpose face recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*, pages 446–456. Springer, 1998.
- D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, October 2007.
- R. Gross. Face databases. In A. S. Li, editor, *Handbook of Face Recognition*. Springer, 2005.
- R. Gross, I. Matthews, and S. Baker. Appearance based face recognition and light fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002.
- G. Guo and S. Z. Li. Content based audio classification and retrieval by support vector machines. *IEEE Transactions On Neural Networks*, 14(1), Jan 2003.
- G. D. Guo, S. Z. Li, and K. L. Chan. Face recognition by support vector machines. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2000.
- S. Gupte, O. Masoud, and N. P. Papanikolopoulos. Vision based vehicle classification. In *IEEE International Conference on Intelligent Transport Systems*, 2000.
- R. Gutierrez. *Introduction To Pattern Recognition*. Wright State University (WRS), 2002.
- N. Habili, C. C. Lim, and A. Moini. Segmentation of the face and hands in sign language video sequences using color and motion cues. *IEEE TCSVT*, 2004.
- D. L. Hall and J. Llinas. *Handbook of Multisensor Data Fusion*. CRC Press, 2001.
- A. Hampapur, L. Brown, J. Connell, S. Pankanti, A. Senior, and Y. Tian. Smart surveillance: Applications, technologies and implications. In *IEEE Pacific-Rim Conference On Multimedia*, 2003.
- A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, S. Pankanti, A. Senior, C.-F. Shu, and Y. L. Tian. Smart video surveillance. *IEEE SP*, 2005.
- D. Hansen, B. Mortensen, P. Duizer, J. Andersen, and T. Moeslund. Automatic annotation of humans in surveillance video. In *International workshop on Video Processing and Recognition*, Montreal, Canada, May 28-30 2007.



- R. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5): 786–804, 1979.
- I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22:809–830, Aug 2000.
- C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, page 147?152, 1988.
- B. Heisele, P. Ho, J. Wu, and T. Poggio. Face recognition: component-based versus global approaches. *Computer Vision and Image Understanding*, 91(1-2):6–21, 2003.
- T. Heseltine, N. Pears, and J. Austin. Evaluation of image preprocessing techniques for eigenface based face recognition. In *Proceedings of the International Conference on Image and Graphics*, 2002.
- T. Heseltine, N. Pears, J. Austin, and Z. Chen. Face recognition: A comparison of appearance-based approaches. In *Proc. Digital Image Computing: Techniques and Applications*, 2003.
- M. Hähnel, D. Klünder, and K.-F. Kraiss. Color and texture features for person recognition. In *International Joint Conference on Neural Networks (IJCNN)*, Jul 2004.
- E. Hjelmås and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding (CVIU)*, 83(3):236–274, 2001.
- L. Hong, A. K. Jain, and S. Pankanti. Can multibiometrics improve performance. In *IEEE Workshop on Automatic Identification Advanced Technologies (AutoID)*, 1999.
- K. Hotta, T. Kurita, and T. Mishima. Scale invariant face detection method using higher-order local auto-correlation features extracted from log-polar image. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 1998.
- P. Howarth and S. Rüger. Evaluation of texture features for content-based image retrieval. In *International Conference on Image and Video Retrieval*, 2004.
- R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(5):696–706, 2002.
- W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviours. *IEEE Transactions on Systems, Man and Cybernetics*, 2004.
- J. Huang, B. Heisele, and V. Blanz. Component based face recognition with 3d morphable models. In *International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, 2003.

- K. Ignasiak, M. Morgos, and S. Ongkittikul. Architecture of information system for intelligent cash machine. In *International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, Barcelona, Spain, Jul 2007.
- M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *European Conference on Computer Vision (ECCV)*, 1996.
- M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision (IJCV)*, 1998.
- K. Jack. Video demystified. Independent Pub Group (Computer), 1996.
- D. W. Jacobs, P. N. Belhumeur, and R. Basri. Comparing images under variable illumination. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998.
- G. Jaffre and P. Joly. Costume a new feature for automatic video content indexing. In *Recherche d'Information Assistee par Ordinateur*, pages 314–325, Avignon, France, april 2004.
- A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 2005a.
- A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice Hall, 1989.
- A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 1999.
- A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:4–37, 2000.
- A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1), Jan 2004.
- A. K. Jain, S. C. Dass, and K. Nandakumar. Can soft biometric traits assist user recognition? In *SPIE Defense and Security Symposium*, apr 2005b.
- A. K. Jain, A. Ross, and S. Pankanti. Biometrics: A tool for information security. *IEEE Transactions on Information Forensics and Security*, 1(2):125–143, Jun 2006.
- S. H. Jeng, H. Y. M. Liao, C. C. Han, M. Y. Chern, and Y. T. Liu. Facial feature detection using geometrical face model: An efficient approach. *Pattern Recog.*, 31, 1998.
- A. Y. Johnson, J. Sun, and A. F. Bobick. Predicting large population data cumulative match characteristic performance from small population data. In *International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, 2003.
- S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 2:241–254, 1967.

- T. Kadir, R. Bowden, E. Ong, and A. Zisserman. Minimal training, large lexicon, unconstrained sign language recognition. In *BMVC*, 2004.
- M. Karaman. *Towards Robust Object Segmentation In Video Sequences And Its Applications*. PhD thesis, Technical University of Berlin, 2009.
- M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. In *International Conference on Computer Vision (ICCV)*, 1987.
- K. Kim, K. Jung, and H. J. Kim. Face recognition using kernel principal component analysis. *IEEE Signal Processing Letters*, 9(2), 2002.
- J. Kittler, H. Mhamad, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1998.
- R. Kjeldsen and J. Kender. Finding skin in color images. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 312–317, 1996.
- M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic linkarchitecture. *Computers, IEEE Transactions on*, 1993.
- A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic tracking, coding and reconstruction of human faces, using flexible appearance models. *IEEE Electron. Lett.*, 30:1578–1579, 1994.
- A. Lanitis, C. J. Taylor, and T. F. Cootes. An automatic face identification system using flexible appearance models. *Image and Vision Computing (IVC)*, 13(5):393–401, 1995.
- S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks (TNN)*, 8:98–113, 1997.
- C. H. Lee, J. S. Kim, and K. H. Park. Automatic human face location in a complex background. *Pattern Recog.*, 29:1877–1889, 1996.
- B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2008.
- T. K. Leung, M. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *International Conference on Computer Vision (ICCV)*, 1995.
- M. S. Lew, N. Sebe, C. D. Liff, and R. Jain. Content based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.
- R. Lienhart, L. Liang, and A. Kuranov. An extended set of haar-like features for rapid object detection. In *International Conference on Image Processing (ICIP)*, 2002.

- R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Pattern Recognition Symposium of the German Association for Pattern Recognition (DAGM)*, 2003.
- S.-H. Lin, S.-Y. Kung, and L.-J. Lin. Face recognition/detection by probabilistic decision-based neural network. *IEEE Transactions on Neural Networks*, 8:114–132, 1997.
- Y. Lin, T. Liu, and C. Fuh. Fast object detection with occlusions. In *European Conference on Computer Vision (ECCV)*, 2004.
- T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998.
- A. J. Lipton. Local application of optic flow to analyze rigid versus nonrigid motion. In *International Conference on Computer Vision (ICCV)*, 1999.
- A. J. Lipton, H. Fujiyoshi, and R. S. Patil. Moving target classification and tracking from real-time video. *IEEE Workshop on Applications of Computer Vision*, 1998.
- Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, Jan 2007.
- D. Lowe. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.
- X. Lu. Image analysis for face recognition, May 2003.
- X.-G. Lv, J. Zhou, and C.-S. Zhang. A novel algorithm for rotated human face detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000.
- J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967.
- B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 18(8):837–842, 1996.
- B. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 11(6):703–715, 2001.
- B. S. Manjunath, R. Chellappa, and C. V. D. Malsburg. A feature based approach to face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1992.
- B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7*. John Wiley & Sons Ltd., 2002.

- A. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *Transactions on Pattern Analysis and Machine Intelligence*, 24, 2002.
- A. Martinez and R. Benavente. The AR face database. Technical Report CVC TR-24, Purdue University, June 1998.
- S. McKenna and S. Gong. Recognising moving faces. In H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*. Springer, 1998.
- S. McKenna, S. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998.
- S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *CVIU*, 80(1):42–56, October 2000.
- D. Meyer, J. Denzler, and H. Niemann. Model based extraction of articulated objects in image sequences for gait analysis. In *International Conference on Image Processing*, page 78781. IEEE, 1998.
- V. Mezaris, I. Kompatsiaris, and M. Strintzis. An ontology approach to object based image retrieval. In *International Conference on Image Processing (ICIP)*, pages 511–514, 2003.
- D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.
- U. Mönich. Objektdetektion basierend auf komponenten und ihrer topologie. Diplomarbeit, Technisch Universität Berlin, Berlin, Germany, May 2005.
- B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. In *Automatic Systems for the Identification and Inspection of Humans, SPIE'94*, volume 2257, 1994.
- B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.
- B. Moghaddam, W. Wahid, and A. Pentland. Beyond eigenfaces: probabilistic matching for face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 1998.
- A. Mohan, C. Papageorgiou, and T. Poggio. Example based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4), Apr 2001.
- M. Montague and J. A. Aslam. Relevance score normalization for metasearch. In *International Conference on Information and Knowledge Management*, page 427433, Atlanta, USA, 2001.

- H. Mueller, W. Mueller, D. McG. Squire, and T. Pun. Performance evaluation in content based image retrieval: Overview and proposals. Technical report, University of Geneva, 1999.
- J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, 5(1):32?38, 1957.
- H. Murase and S. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14(1):5?24, 1995.
- C. Nakajima, M. Pontil, and T. Poggio. People recognition and pose estimation in image sequences, 2000.
- C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full body person recognition system. *Pattern Recognition*, 2003.
- K. Nandakumar. Integration of multiple cues in biometric systems. Master's thesis, Michigan State University, 2005.
- K. Nandakumar. *Multibiometric Systems: Fusion Strategies and Template Security*. PhD thesis, Michigan State University (MSU), 2008.
- A. V. Nefian. *Statistical Approaches to Face Recognition*. PhD thesis, Gerogia Institute of Technology, 1996.
- A. V. Nefian and M. H. Hayes. Hidden markov models for face recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.
- T. Ojala, M. Pietikainen, and T. Maenpaa. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision (ECCV)*, 2000.
- S. Ongkittikul, S. Worrall, and A. Kondoz. Enhanced hand tracking using the k-means embedded particle filter with mean-shift vector re-sampling. In *Visual Information Engineering Conference (VIE)*, 2008.
- E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–136, Juna 1997.
- C. P. Papageorgiou. Object and pattern detection in video sequences. Master's thesis, Massachusetts Institute of Technology (MIT), 1997.
- N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22:266–280, Mar 2000.

- D. K. Park, Y. S. Jeon, C. S. Won, and S.-J. Park. Efficient use of local edge histogram descriptor. In *ACM International Workshop on Standards, Interoperability and Practices*, page 52–54, 2000.
- S. Park and J. K. Aggarwal. Recognition of human interaction using multiple features in grayscale images. In *Proceedings of the International Conference on Pattern Recognition*, pages 51–54, 2000.
- S. Park and J. K. Aggarwal. Segmentation and tracking of interacting human body parts under occlusion and shadowing. In *IEEE Workshop on Motion and Video Computing*, pages 105–111, 2002.
- S. Park, J. Park, and J. K. Aggarwal. Video retrieval of human interactions using model-based motion tracking and multi-layer finite state automata. In *CIVR'03*, 2003.
- G. Pass and R. Zabih. Comparing images using color coherence vectors. In *ACM Conference on Multimedia*, 1996.
- P. Penev and J. Atick. Local feature analysis: A general statistical theory for object representation, 1996.
- A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994.
- F. Pereira, A. Vetro, and T. Sikora. Multimedia retrieval and delivery: Essential metadata challenges and standards. *Proceedings of the IEEE*, 96(4):721–744, 2008.
- J. Pers, G. Vuckovic, B. Dezman, and S. Kovacic. Human activities at different levels of detail, 2003.
- N. Peterfreund. Robust tracking of position and velocity with kalman snakes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22:564–569, Jun 2000.
- P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 947–954, Jun 2005. doi: 10.1109/CVPR.2005.268.
- C. Poynton. Color faq, 2008.
- F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms, 1998.
- A. Rama and F. Tarres. Lophoscopic PCA: A novel method for face recognition. In *Workshop on Image Analysis for Multimedia Interactive Services*, Montreux, Switzerland, April 2005.

- A. Rama and F. Tarres. Partial lda vs partial pca. In *International Conference on Multimedia and Expo (ICME)*, 2006.
- C. S. Regazzoni, V. Ramesh, and G. L. Foresti. Scanning the issue technology. *Proceedings Of The IEEE: Special Issue on Video Communications, Processing, and Understanding for Third Generation Surveillance Systems*, 89(10), OCTOBER 2001.
- J. Renno, N. Lazarevic-McManus, D. Makris, and G. Jones. Evaluating motion detection algorithms: Issues and results. In *Sixth IEEE International Workshop on Visual Surveillance*, May 2006.
- D. Roth, M.-H. Yang, and N. Ahuja. A snow based face detector. *Advances in Neural Information Processing Systems (NIPS)*, 2000.
- S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
- H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1996.
- H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20(1):23–28, January 1998a.
- H. A. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network based face detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998b.
- P. Salembier, N. O'Connor, P. Correia, and F. Pereira. Hierarchical visual description schemes for still images and video sequences. In *International Conference on Image Processing (ICIP)*, 1999.
- S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *ICTAI*, 2004.
- F. S. Samaria. *Face Recognition Using Hidden Markov Models*. PhD thesis, University of Cambridge, 1994.
- C. Sanderson. *Automatic Person Verification Using Speech and Face Information*. PhD thesis, Griffith University, Queensland, Australia, 2002.
- H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998.
- H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 746–751, 2000.



- B. Schoelkopf, A. Smola, and K. Mueller. Non linear component analysis as a kernel eigenvalue problem. *Neural Computing*, 10:1299–1319, 1998.
- B. Schoelkopf, A. Smola, and K. Mueller. Kernel principal component analysis. In *Advances in Kernel Methods Support Vector Learning*. MIT Press, 1999.
- G. Shakhnarovich and B. Moghaddam. Face recognition in subspaces. In S. Z. Li and A. K. Jain, editors, *Handbook of Face Recognition*. Springer Verlag, December 2004.
- A. Shashua. On photometric issues in 3d visual recognition from single 2d image. *International Journal of Computer Vision (IJCV)*, 21(99-122), 1997.
- S. Siggelkow. *Feature Histograms for Content-Based Image Retrieval*. PhD thesis, Albert-Ludwigs-Universität Freiburg, 2002.
- P. Sinha. Object recognition via image invariants: A case study. *Investigative Ophthalmology and Visual Science*, 35(4):1735–1740, 1994.
- P. Smets, E. Mamdami, D. Dubois, and H. Prade. *Non Standard Logics for Automated Reasoning*. ISBN 0126495203. Academic Press, Harcourt Brace Jovanovich Publisher, 1988.
- A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish. Clustering speakers by their voices. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 757–760, May 1998. doi: 10.1109/ICASSP.1998.675375.
- J. Sporring, M. Nielsen, L. Florack, and P. Johansen. *Gaussian Scale-Space Theory*. Kluwer Academic Publishers, 1997.
- C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 246–252, 1999.
- C. Stauffer and W. E. L. Grimson. Similarity templates for detection and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.
- S. S. Stevens. On the psychophysical law. *Psychological Review*, 1957.
- M. Stricker and M. Orengo. Similarity of color images. In *EI*, 1995.
- A. Sturn. Cluster analysis for large scale gene expression studies. Master thesis, Graz University of Technology, 2000.
- Z. Su, H. Zhang, S. Li, and S. Ma. Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning. *IEEE Transactions on Image Processing*, 12(8):924–937, Aug 2003.

- K.-K. Sung and T. Poggio. Example based learning for view based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20(1), January 1998.
- M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1): 11–32, 1991.
- B. Takacs and H. Wechsler. Detection of faces and facial landmarks using iconic filter banks. *Pattern Recog.*, 30, 1997.
- H. Tamura, S. Mori, and T. Yamakawi. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics (TSMC)*, 1978.
- P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, Jul 2005.
- L. F. Teixeira and L. Corte-Real. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters*, 2008.
- J.-C. Terrillon, M. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, 2000.
- L. Torres, L. Lorente, and J. Vila. Automatic face recognition of video sequences using self-eigenfaces. In *International Symposium on Image/Video Communication over Fixed and Mobile Networks*, Rabat, Morocco, April 17-20 2000.
- M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1): 71–86, 1991a.
- M. Turk and A. Pentland. Face recognition using eigenfaces. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, 1991b.
- T. Tuytelaars and K. Mikolajczyk. A survey on local invariant features, May 2006.
- V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 2nd edition, 2000.
- T. Vetter and V. Blanz. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003.
- P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518, 2001a.
- P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 57(2), May 2004.

- P. A. Viola and M. J. Jones. Robust real-time object detection. In *IEEE Workshop on Statistical and Computational Theories of Computer Vision*, 2001b.
- C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *CVIU*, 81(3), 2001.
- U. von Luxburg. *Statistical Learning with Similarity and Dissimilarity Functions*. PhD thesis, Technische Universität Berlin, 2004.
- J. Z. Wang, G. Wiederhold, O. Firschein, and S. X. Wei. Wavelet based image indexing techniques with partial sketch retrieval capability. In *Fourth Forum on Research and Technology Advances in Digital Libraries*, 1997.
- M. Weiser. The computer for the twenty-first century. *Scientific American*, 265(3):94–104, 1991.
- L. Wiskott, J.-M. Fellous, N. Krueger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7), JULY 1997.
- S. Won and D. K. Park. Image block classification and variable block size segmentation using a model-fitting criterion. *Optical Engineering*, 36:2204–2209, 1997.
- B. Wu. *Part based Object Detection, Segmentation, and Tracking by Boosting Simple Feature based Weak Classifiers*. PhD thesis, USC, 2008.
- B. Wu and R. Nevatia. Detection and tracking of multiple partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 2007a.
- B. Wu and R. Nevatia. Simultaneous object detection and segmentation by boosting local shape feature based classifier. In *CVPR*, 2007b.
- B. Wu, R. Nevatia, and Y. Li. Segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. In *CVPR*, 2008.
- Y. Wu, B. Tseng, and J. Smith. Ontology based multi-classification learning for video concept detection. In *International Conference on Multimedia and Expo (ICME)*, volume 2, pages 1003–1006 Vol.2, June 2004. doi: 10.1109/ICME.2004.1394372.
- L. Xu, A. Krzyzak, and C. Y. Suen. Associative switch for combining multiple classifiers. *Journal of Artificial Neural Networks*, 1(1):77–100, 1994.
- G. Yang and T. S. Huang. Human face detection in complex background. *Pattern Recognition*, 27(1):53–63, 1994.

- J. Yang, D. Zhang, A. Frangi, and J. Yu Yang. Two dimensional pca: a new approach to appearance based face representation and recognition. *2004*, 26(1):131–137, Jan 2004. doi: 10.1109/TPAMI.2004.1261097.
- M.-H. Yang. Face recognition using extended isomap. In *International Conference on Image Processing (ICIP)*, volume 2, page 117–120, 2002.
- M.-H. Yang, N. Ahuja, and D. Kriegman. Face detection using mixtures of linear subspaces. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2000.
- M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(1):34–58, January 2002.
- A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 2006.
- K. C. Yow and R. Cipolla. Feature based human face detection. *Image and Vision Computing (IVC)*, 15(9), 1997.
- A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision (IJCV)*, 8(2):99–111, 1992.
- W. Zhao and R. Chellappa. Sfs based view synthesis for robust face recognition. In *International Conference on Automatic Face and Gesture Recognition*, 2000.
- W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 1998.
- W. Zhao, R. Chellappa, and P. J. Phillips. Subspace linear discriminant analysis for face recognition. Technical report, University of Maryland, 1999.
- W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey. survey, University of Maryland, 2000.
- W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 4(35):399–458, 2003.
- S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding (CVIU)*, 2003.
- S. K. Zhou and R. Chellappa. Biometrics for surveillance. In *International Conference on Image Processing (ICIP)*, 2006.

# Publications

- L. Goldmann, M. Karaman, and T. Sikora. Human body posture recognition using mpeg-7 descriptors. In *Electronic Imaging (EI)*, San Jose, USA, Jan 2004.
- L. Goldmann, M. Krinidis, N. Nikolaidis, S. Asteriadis, and T. Sikora. An integrated system for face detection and tracking. In *Workshop on Immersive Communication and Broadcast Systems (ICOB)*, Berlin, Germany, October 27-28 2005.
- L. Goldmann, M. Karaman, J. T. Saez Minquez, and T. Sikora. Appearance based person recognition for surveillance applications. In *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2006a.
- L. Goldmann, U. Mönich, and T. Sikora. Robust face detection based on components and their topology. In *Electronic Imaging (EI)*, 2006b.
- L. Goldmann, A. Samour, M. Karaman, and T. Sikora. Extracting high level semantics by means of speech, audio, and image primitives in surveillance applications. In *International Conference on Image Processing (ICIP)*, 2006c.
- L. Goldmann, A. Samour, and T. Sikora. Multimodal analysis for universal smart room applications. In *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2006d.
- L. Goldmann, L. Thiele, and T. Sikora. Online image retrieval system using long term relevance feedback. In *International Conference on Image and Video Retrieval (CIVR)*, 2006e.
- L. Goldmann, U. J. Mönich, and T. Sikora. Components and their topology for robust face detection in the presence of partial occlusions. *IEEE Transactions on Information Forensics and Security, Special Issue on Human Detection and Recognition*, 2(3), 2007a. doi: 10.1109/TIFS.2007.902019.
- L. Goldmann, A. Samour, and T. Sikora. Towards person google: Multimodal person search and retrieval. In *International Conference on Semantics and Digital Media Technologies (SAMT)*, 2007b.

- L. Goldmann, T. Adamek, P. Vajda, M. Karaman, R. Mörzinger, E. Galmar, T. Sikora, N. O'Connor, T. Ha-Minh, T. Ebrahimi, P. Schallauer, and B. Huet. Towards fully automatic image segmentation evaluation. In *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, 2008a.
- L. Goldmann, T. Adamek, P. Vajda, M. Karaman, R. Mörzinger, E. Galmar, T. Sikora, N. O'Connor, T. Ha-Minh, T. Ebrahimi, P. Schallauer, and B. Huet. Towards fully automatic image segmentation evaluation. In *CVPR*, 2008b.
- L. Goldmann, A. Rama, T. Sikora, and F. Tarres. On the detection and localization of facial occlusions and its use within different scenarios. In *International Workshop on Multimedia Signal Processing (MMSP)*, 2008c.
- M. Karaman, L. Goldmann, D. Yu, and T. Sikora. Comparison of static background segmentation methods. In *Visual Communications and Image Processing (VCIP)*, 2005.
- M. Karaman, L. Goldmann, and T. Sikora. A new segmentation approach using gaussian color model and temporal presentation. In *Electronic Imaging (EI)*, 2006.
- M. Karaman, L. Goldmann, and T. Sikora. Improving object segmentation by reflection detection and removal. In *EI*, 2009.
- A. Koutsia, N. Grammalidis, K. Dimitropoulos, M. Karaman, and L. Goldmann. Football player tracking from multiple views using a novel background segmentation algorithm and multiple hypothesis tracking. In *2nd International Conference on Computer Vision Theory and Applications*, 2007.
- A. Rama, L. Goldmann, F. Tarres, and T. Sikora. More robust face recognition by considering occlusion information. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Sep 2008.
- D. Rodriguez, L. Goldmann, S. Ongkittikul, M. Karaman, and T. Sikora. A system for personalized human computer interaction. In *50th International Symposium Electronics in Marine (ELMAR)*, Sep 2008.
- A. Samour, L. Goldmann, and T. Sikora. Towards person google: Multimodal person search and retrieval. In *K-Space PhD Jamboree (KSPJ)*, 2007a.
- A. Samour, M. Karaman, L. Goldmann, and T. Sikora. Video to the rescue of audio: Shot boundary assisted speaker change detection. In *EI*, 2007b.
- P. Wilkins, T. Adamek, P. Ferguson, M. Hughes, G. J.F.Jones, G. Keenan, K. McGuinness, J. Malobabic, N. E. O'Connor, D. Sadlier, A. F. Smeaton, R. Benmokhtar, E. Dumont, B. Huet, B. Merialdo, E. Spyrou, G. Koumoulos, Y. Avrithis, R. Moerzinger, P. Schallauer, W. Bailer, Q. Zhang, T. Piatrik, K. Chandramouli, E. Izquierdo, L. Goldmann, M. Haller,

- T. Sikora, P. Praks, J. Urban, X. Hilaire, and J. M. Jose. K-space at trecvid 2006. In *TREC Video Retrieval Evaluation (TRECVID)*, 2006.
- P. Wilkins, T. Adamek, D. Byrne, G. J.F.Jones, H. Lee, G. Keenan, K. McGuinness, N. E. O'Connor, A. F. Smeaton, A. Amin, Z. Obrenovic, R. Benmokhtar, E. Galmar, B. Huet, S. Essid, R. Landais, F. Vallet, G. T. Papadopoulos, S. Vrochidis, V. Mezaris, I. Kompat-siaris, E. Spyrou, Y. Avrithis, R. Mörzinger, P. Schallauer, W. Bailer, T. Piatrik, K. Chandramouli, E. Izquierdo, M. Haller, L. Goldmann, A. Samour, A. Cobet, T. Sikora, and P. Praks. K-space at trecvid 2007. In *TREC Video Retrieval Evaluation (TRECVID)*, 2007.
- P. Wilkins, T. Adamek, D. Byrne, G. J.F.Jones, H. Lee, G. Keenan, K. McGuinness, N. E. O'Connor, A. F. Smeaton, A. Amin, Z. Obrenovic, R. Benmokhtar, E. Galmar, B. Huet, S. Essid, R. Landais, F. Vallet, G. T. Papadopoulos, S. Vrochidis, V. Mezaris, I. Kompat-siaris, E. Spyrou, Y. Avrithis, R. Mörzinger, P. Schallauer, W. Bailer, T. Piatrik, K. Chandramouli, E. Izquierdo, M. Haller, L. Goldmann, A. Samour, A. Cobet, T. Sikora, and P. Praks. K-space at trecvid 2008. In *TREC Video Retrieval Evaluation (TRECVID)*, 2008.