

# Exploiting Motion Information for Video Analysis in Sequences with Moving Camera

vorgelegt von  
Dipl.-Ing.  
Marina Georgia Arvanitidou  
geboren in Thessaloniki

von der Fakultät IV – Elektrotechnik und Informatik  
der Technischen Universität Berlin  
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften  
- Dr.-Ing. -

genehmigte Dissertation

Promotionausschuss:

Vorsitzender:	Prof. Dr. Sebastian Möller	Technische Universität Berlin
Gutachter:	Prof. Dr. Thomas Sikora	Technische Universität Berlin
Gutachter:	Prof. Dr. Athanassios Skodras	University of Patras
Gutachter:	Dr. Lutz Goldmann	incube Labs

Tag der wissenschaftlichen Aussprache: 11. November 2014

Berlin 2015





# Abstract

---

The processing of low-level information towards the extraction of high-level information, and specifically object segmentation, constitutes a great challenge for image processing and computer vision. With this respect, motion estimation is a task of major importance, since motion constitutes arguably one of the most valuable underlying clues in image sequences. Furthermore, motion is shown to be strongly connected with the human visual system. Further possibilities are thus emerging in the field of visual quality assessment for developing appropriate motion exploitation strategies that are aligned with the human visual system. This thesis focuses on exploiting motion for video analysis in image sequences captured by a moving camera and provides an appropriate evaluation framework.

Firstly, motion induced by the camera movement has to be distinguished from motion resulting from the moving content itself. Therefore, the first part of the thesis is devoted to global motion estimation, i.e. the estimation of background motion. Outlier regression techniques are employed for the formulation of parametric models for global motion. It is shown that this modelling benefits from the consideration of block information since it implicitly contains information regarding foreground objects that move independently of the background region. Moreover, the parametric modelling of global motion is shown to have a positive influence towards enhancing conventional motion prediction.

The second part of the thesis deals with object segmentation. A short-term object segmentation scheme that exploits bidirectional information for change detection is built, based on parametric modelling of global motion. Aspects related to the thresholding procedure, namely the spatial location of foreground candidates and the optimal selection of the involved parameters are examined. Thus, robust segmentation performance is achieved avoiding heuristics and training algorithms for parameter selection. Furthermore, background classification inconsistencies occurring during the independent calculation of segmentation masks over time are addressed using adaptive filtering according to foreground motion.

Finally, the exploitation of motion features and object-knowledge on video quality assessment is investigated. Existing objective quality assessment algorithms often rely on the calculation of quality scores ignoring such higher-level information. Thus, possibilities of improving objective video quality assessment models' performance are herein examined. Specifically, the contributions on objective video quality assessment are threefold; building a content-aware video quality assessment approach that accounts for moving objects, formulating a saliency model that exploits motion features on spatial level and furthermore proposing an approach for consideration of global motion in the temporal dimension that leads to accuracy improvement.



# Zusammenfassung

---

Die Verarbeitung von Informationen auf niedriger Hierarchieebene zur Extraktion von Information auf höheren Ebenen und insbesondere die Objektsegmentierung stellt eine große Herausforderung in der Bildverarbeitung dar. In dieser Hinsicht ist die Bewegungsschätzung ein besonders wichtiger Prozess, da Bewegung wohl einer der bedeutungsvollsten Aspekte in Bildsequenzen ist.

Dabei ist Bewegung von besonderer Bedeutung im menschlichen visuellen System. Weitere Möglichkeiten und Notwendigkeiten ergeben sich daher auf dem Gebiet der objektiven Qualitätsevaluierung von optischen Darstellungen, wobei die Entwicklung von geeigneten Verfahren angestrebt wird, welche in Einklang mit der subjektiven Wahrnehmung sind.

Der Schwerpunkt dieser Arbeit liegt in der Verwendung der Bewegung zur Analyse von Bildsequenzen, welche mit einer bewegten Kamera aufgenommen sind sowie in der Entwicklung geeigneter Methoden zu deren Evaluierung.

Dabei wird zunächst zwischen der Bewegung, welche durch eine bewegte Kamera verursacht wird, und der Bewegung der Objekte innerhalb des Bildes unterschieden. Der erste Teil der Arbeit befasst sich mit der globalen Bewegungsschätzung, also der Schätzung der Bewegung des Bildhintergrundes. Regressionsmethoden werden eingesetzt, um parametrische Modelle für die globale Bewegung zu formulieren. Es wird dabei gezeigt, dass die Berücksichtigung der Blockgröße für den Modellierungsprozess besonders hilfreich ist, da diese implizite Information über die sich unabhängig vom Hintergrund bewegenden Vordergrundobjekte beinhaltet. Des Weiteren wird gezeigt, dass die parametrische Modellierung der globalen Bewegung einen positiven Einfluss auf die Qualität der Bewegungsprädiktion hat.

Im zweiten Teil der Arbeit wird die Objektsegmentierung untersucht. Eine Objektsegmentierung, welche auf der parametrischen Modellierung der kurzzeitigen globalen Bewegung basiert, nutzt bidirektionale Information aus, um Änderungen zu detektieren. Aspekte bei der Wahl eines geeigneten Schwellwertes für die Detektion, nämlich die räumliche Lage der Vordergrundkandidaten sowie die Optimierung der Parametrisierung werden untersucht. So wird eine robuste Segmentierung erreicht, ohne auf heuristische Methoden und Trainingsalgorithmen zurückgreifen zu müssen. Weiterhin werden Fehlklassifikationen des Hintergrundes, welche während der unabhängigen Berechnung der Segmentierungsmasken über die Zeit auftreten, durch adaptive Filterung behandelt, welche sich an die Bewegung des Vordergrundes anpasst.

Letztlich werden die resultierenden Bewegungseigenschaften und die Kenntnis über die Objekte für die Qualitätsevaluierung untersucht. Existierende Algorithmen zur objektiven Bewertung der Qualität basieren auf der Berechnung von Qualitätsindikatoren, welche Information auf höheren Ebenen meist nicht berücksichtigen. Deshalb werden hier entsprechende Verbesserungen untersucht. Im Einzelnen sind die Beiträge folgende: Aufbau einer inhaltsbezogenen Methode zur Bewertung der Videoqualität, welche für sich bewegende Objekte einsetzbar ist, Formulierung eines Modells, welches visuelle Wichtigkeit und räumliche Bewegungseigenschaften nutzt sowie die Entwicklung einer Methode zur Berücksichtigung der globalen Bewegung in der zeitlichen Dimension, welche die Genauigkeit der Qualitätsbewertung und die Konsistenz mit subjektiver Evaluierung erhöht.

# Acknowledgments

---

For the accomplishment of this thesis I received advice, support and help from many people around me. I am grateful to all of them for their contribution they made to this work. First, I would like to express my sincere gratitude to my supervisor Thomas Sikora, for giving me the opportunity to work on this thesis at the Communication Systems Group of the Technical University and his support all these years. I would also like to thank Athanassios Skodras and Lutz Goldmann for accepting to review this thesis so promptly and for their precious feedback and comments.

Special thanks go to my colleagues Rubén Heras Evangelio and Savvas Argyropoulos for their valuable feedback on the thesis and the fruitful discussions we have had. I am also grateful to Florian Kaiser for the proofreading of the thesis, for being the kindest roommate for some years at NUe and also for being a precious and supportive friend. I also thank my colleague Tobias Senst for his valuable comments on the preparations of this thesis and all NUe colleagues for the nice working environment. I am also grateful to Birgit Boldin and Wiryadi for their kind support in the organisational matters.

Finally, I would like to thank my friends, especially Morfoula, Eleni, Nikos, Tasos, Engin, Cagri, Maria, Erkal with whom I shared nice and difficult moments, and made everything funnier and easier. My deepest gratitude goes to my parents for their love and affection that I receive so generously. To my sisters Efi, Lena, Irini and also Dimitris who were always there for me and make me feel really lucky to have them in my life! And to Konstantinos, without whose love and support I wouldn't be able to carry out this work.



# Contents

<b>Abstract</b>	<b>i</b>
<b>Zusammenfassung</b>	<b>iii</b>
<b>Acknowledgments</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation and objectives . . . . .	1
1.2 Contributions and organisation of the thesis . . . . .	5
<b>2 Global Motion Analysis</b>	<b>9</b>
2.1 Introduction . . . . .	10
2.1.1 Motion modelling in video sequences . . . . .	10
2.1.2 Global motion estimation . . . . .	13
2.1.3 Existing approaches on global motion estimation . . . . .	14
2.2 Improved global motion estimation through variable-size blocks . . . . .	16
2.2.1 The binary partition tree . . . . .	16
2.2.2 Least squares estimation . . . . .	17
2.2.3 Improved robust estimation through block size weighting . . . . .	19
2.2.4 Improved robust estimation through block size selection . . . . .	22
2.3 Evaluation of global motion estimation accuracy . . . . .	22
2.3.1 Test dataset . . . . .	22
2.3.2 Evaluation methodology . . . . .	22
2.3.3 Results . . . . .	23
2.4 Adaptive motion prediction through global motion . . . . .	34
2.4.1 Introduction and existing approaches . . . . .	34
2.4.2 Adaptive mode selection . . . . .	37
2.5 Evaluation of the AMS scheme . . . . .	38
2.5.1 Test dataset . . . . .	38
2.5.2 Methodology . . . . .	38
2.5.3 Results . . . . .	39
2.6 Chapter summary . . . . .	40
<b>3 Moving Object Segmentation</b>	<b>47</b>
3.1 Introduction . . . . .	48
3.1.1 Problem statement . . . . .	48
3.1.2 Related work . . . . .	48
3.1.3 Overview of the proposed approach . . . . .	51
3.2 Bidirectional error frame generation . . . . .	53
3.2.1 Global motion compensation . . . . .	53

3.2.2	Error fusion . . . . .	55
3.3	Thresholding of error maps using hysteresis . . . . .	56
3.3.1	Adaptive anisotropic diffusion filtering . . . . .	56
3.3.2	Weighted mean thresholding using spatial connectivity . . . . .	59
3.3.3	Weight selection . . . . .	61
3.4	Background classification consistency . . . . .	65
3.4.1	Sources of errors . . . . .	65
3.4.2	Temporal consistency . . . . .	65
3.5	Experimental evaluation . . . . .	66
3.5.1	Test dataset . . . . .	66
3.5.2	Evaluation methodology . . . . .	67
3.5.3	Results . . . . .	70
3.5.4	Application on H.264/AVC compressed video data . . . . .	85
3.6	Chapter Summary . . . . .	87
<b>4</b>	<b>Content-aware Video Quality Assessment</b>	<b>89</b>
4.1	Introduction . . . . .	90
4.1.1	Subjective video quality assessment . . . . .	91
4.1.2	Objective video quality assessment . . . . .	93
4.1.3	Advances in content-aware quality assessment . . . . .	96
4.1.4	Motivation of the proposed work. . . . .	101
4.2	Method M1: moving object-aware VQA Improvement . . . . .	103
4.2.1	Moving object segmentation . . . . .	104
4.2.2	Foreground and background pooling . . . . .	104
4.3	Method M2: motion saliency for VQA Improvement . . . . .	106
4.3.1	Motion saliency model . . . . .	106
4.3.2	Spatial pooling . . . . .	109
4.3.3	Temporal pooling . . . . .	112
4.4	Experimental evaluation . . . . .	115
4.4.1	Distortion indicators through image quality models . . . . .	116
4.4.2	Prediction performance indicators . . . . .	118
4.4.3	Quality prediction performance of M1 . . . . .	119
4.4.4	Quality prediction performance of spatial pooling in M2 . . . . .	121
4.4.5	Quality prediction performance of temporal pooling in M2 . . . . .	128
4.4.6	Comparison of method M1 and method M2 . . . . .	130
4.5	Chapter summary . . . . .	132
<b>5</b>	<b>Conclusion</b>	<b>133</b>
5.1	Summary of the thesis . . . . .	133
5.2	Discussion and outlook . . . . .	135



---

<b>A Description of datasets</b>	<b>139</b>
A.1 Video Dataset 1 . . . . .	139
A.2 Video Dataset 2 . . . . .	141
A.3 LIVE video database . . . . .	147
A.3.1 Sequences Description . . . . .	147
A.3.2 Distortion types . . . . .	149
<b>List of Figures</b>	<b>151</b>
<b>List of Tables</b>	<b>155</b>
<b>Publications</b>	<b>v</b>
<b>Bibliography</b>	<b>vii</b>



# Introduction

---

Analysis of motion in image sequences is the main focus of this thesis. Especially in the case of video content captured by a moving camera, the analysis and processing of low-level information towards high-level information extraction needs accurate modelling of existing motion. Towards this goal, we build a model to determine the motion resulting from the camera movement. Based on such a modelling we propose an object segmentation scheme and we further study the effect of object-knowledge on visual quality perceived by the viewer and introduce further approaches for improvement of objective video quality assessment.

In this introductory chapter, we will first explain the motivation of this work, and provide some background information. Furthermore, the contributions of this thesis are summarised and the organisation of the manuscript is introduced.

## 1.1 Motivation and objectives

Modern technologies and electronic devices together with increased networking capabilities have made it possible to experience a more advanced way of communication using digital media. People tend to produce increasingly more multimedia content, share it on social networks, store it and eventually utilize it. This results in an increasing amount of video material that can be practically usable if users have access to its actual content in a functional way. For instance, in case a user wishes to find a specific video it would be convenient to avoid going through his huge video collection, and instead be able to find the desired information in a semantically meaningful way.

Moreover, content knowledge is influential in a numerous applications such as object-based video coding [1, 2, 3], where depending on the content, each image region is encoded with different coding requirements, towards more efficient compression. Recently, the impact of content knowledge on the field of visual quality assessment has also gained significant attention [4, 5, 6, 7] by exploring the potential of exploiting attributes of the human visual system in the field of computer vision. This enables content to be redefined or more concretely defined for the specific application of designing more accurate automatic visual quality assessment algorithms, which are highly desired as will be discussed later on.<sup>1</sup>

Content-based video representation is thus regarded as an aspect of vital importance in contemporary multimedia systems. Depending on the application, the

---

<sup>1</sup>Extended related literature to the proposed approaches is reviewed at the respective sections of the main chapters of this thesis.

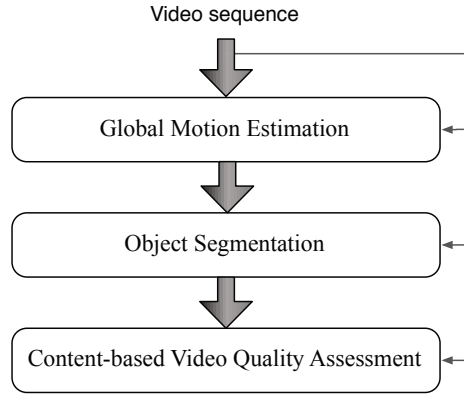


Figure 1.1: Overview of the proposed system.

definition of content, such as objects and object boundaries, may vary significantly, hindering the generic definition of the content-based video representation problem and hence its solution. Towards a functional solution to each scenario’s needs, existing approaches formulate the problem based on the grouping of (high-level) features that are extracted from (low-level) visual information. The challenge is thus to bridge the gap between human interpretation and low-level features, such as colour, motion and texture, that can be automatically derived from the visual content itself.

In the framework of this thesis we initially build a model to determine the motion resulting from the camera movement. Based on such a modelling we build the proposed object segmentation scheme and we subsequently propose approaches for improvement of objective video quality assessment by exploiting the effect of extracted motion features as well as object-knowledge on the visual quality perceived by the viewer. Figure 1.1 illustrates this work flow and subsequently each concept is introduced.

## Global motion estimation

In this thesis we particularly study motion, which relies on changes in the image intensities over time. Motion provides valuable information with respect to content-based video representation and the applications therefor. This is one of the main motivations of this thesis. In the case of video sequences captured by a moving camera there exist two kinds of motion; the motion induced by camera and the motion of existing objects in the scene. In a video analysis framework with moving camera, the effect of the moving camera on the visual scene has to be detected in order to further analyse the video data and achieve the desired outcome. Our aim is thus to formulate a model that describes accurately the motion between given frames of an image sequence in order to estimate the motion caused by the moving camera while recognising and discarding outliers.

Global motion estimation can be performed based on pixel correspondences [8, 9], block correspondences [2, 10] or, more generally, feature correspondences [11] at

arbitrary positions. In each case the identification of outliers is a critical step in this task, and therefore it has to be carefully decided. Global motion estimation approaches based on pixel correspondences may be regarded as the ones with the highest potential for achieving a high accuracy. Nevertheless, it has been shown that they are often outperformed by block based approaches [12]. This may be attributed mainly to the influence of outlier rejection which is a critical stage in global motion estimation, and determines strongly the effectiveness of the global motion estimation performance. In the case of global motion estimation based on block correspondences, block partitioning characteristics are often not taken into account. The underlying block partitioning however results in a block assignment that is determined by the image content itself. Therefore, global motion estimation could benefit from considering further aspects of block partitioning.

Furthermore, conventional motion prediction can benefit from the parametric representation of camera motion under specific circumstances studied in the framework of this thesis. Towards estimating camera motion, in this work we deal with three regression analysis approaches, namely the *M-Estimator*, the *Helmholtz trade-off estimator* and *RANSAC*. The M-Estimator is used in the following chapter 2, whereas the Helmholtz trade-off estimator and RANSAC are discussed in chapters 3 and 4, respectively. Comprehensive studies on robust regression theory can be found in [13, 14] and [15]. Global motion estimation may subsequently enable the accurate compensation of global motion which is a prerequisite for moving object segmentation that is a major topic in this thesis.

## Object segmentation

Research in the field of content retrieval is facing a wide range of challenges, driven by the need to describe and understand video content in an automatic fashion. Despite the increasing computer capabilities and the fact that the community has made substantial progress towards content understanding, it still remains a big challenge for computers to be able to understand content automatically. Even though it may be easy for a human to intuitively define an object region, it is quite difficult to formulate a generic and strict definition of the segmentation problem. On a similar context, quite early in the 5<sup>th</sup> century B.C., it was expressed by Parmenides that "... for it is the same thing that can be thought and that can be" [16]<sup>2</sup>, indicating that the existence of an object depends on the thoughts of the one who recognises what it is. This statement applies aptly in the search for the "correct" content-based decomposition of an image where the answer requires human perception. Incorporating this functionality into an artificial vision system is though still far from becoming reality.

In the case of video sequences, the task of object segmentation aims at a semantical decomposition of the scene to regions according to its content, where the term region denotes a set of spatially connected pixels. In this work, based on accurate parametric global motion estimation we address the *moving object segmentation*

<sup>2</sup> ... τὸ γὰρ αὐτὸ νοεῖν ἐστὶν τε καὶ εἶναι

task, which is an early and fundamental step that one encounters in many applications of image processing. Accurate global motion estimation has enabled the design of object segmentation algorithms [17, 11, 3] that strongly depend on the underlying global motion estimation. In fact, global motion estimation and object segmentation are often considered as interdependent information [18, 19], since the knowledge about object boundaries enables more accurate global motion estimation and on the other hand more accurate estimation of global motion enables accurate object boundary detection. Even though it would be very convenient, this information is typically not available a priori.

Challenges in the task of moving object segmentation are mainly associated with the automatic determination of underlying parameters that improve the accuracy of the segmentation results. This is considered to be a challenging task and existing approaches often focus on specific applications where it is possible to address the problem with empirical settings [18]. Robustness in the sense of broadening the applicability of the proposed approaches is always an important goal in this area, since several underlying assumptions have to be usually adopted. Furthermore, consistency, i.e. stability of the derived segmentation masks, is an additional requirement that is especially important.

### Content-based video quality assessment

An important aspect in modern communication systems, with respect to the delivery of multimedia services, is the quality of visual content. Contemporary video coding standards, such as the recent high efficiency video coding (HEVC) [20] and the most widely used H.264/AVC [21], enable the broad experience of multimedia data using mobile services. In the case visual data is transmitted or stored, it undergoes distortions mainly due to limited channel bandwidth and compression. However, content has to be delivered to a certain level of quality to the users. Therefore emerges the need to measure the delivered data quality, which is not achievable in a straightforward manner, since it relates to the way humans see and respectively perceive the level of visual quality. Currently there is indeed a great deal of interest in the research community in this topic and an effort to bring technology and neuroscience together. With respect to *seeing*, the human visual system is not functioning in a linear way, but as it is pointed out [4, 7] and shown in this thesis, it is rather strongly influenced by the visual content itself.

To illustrate this concept with an example, in the case of a viewer watching a football game in the television, it is rather impossible to notice a distortion (e.g. block artefact) that happens outside of the football field during the game. On the contrary a distortion on the field or on players, where the main focus is, will be probably more noticeable. In this line, there is a significant amount of research effort [4, 6] currently aiming to appropriately measure the experienced quality of the delivered content. Motivated by this idea, the last part of the thesis focuses on enhancing quality assessment methods. Specifically, the impact of motion-related features to the process of visual quality assessment based on global motion compensation is

extensively explored. Since the mechanism of seeing and perceiving visual quality is not fully understood among the research community, the goal is to improve the way algorithms assess automatically visual quality bringing them closer to the actually perceived quality by the user.

## 1.2 Contributions and organisation of the thesis

The structure of the thesis is delineated in the following, and the main contributions of this work as well as related publications are given in this section.

### Chapter 2

Dealing with sequences captured by a moving camera, the first step is to study the parametric modelling of motion between two-dimensional images. The first part of the thesis is thus dedicated to the calculation of the parametric model that describes camera motion. Hence, chapter 2 provides an overview and discussion of existing global motion estimation approaches and proceeds with the description of the proposed approach for enhanced robust global motion estimation. Furthermore, it is studied how this accurate representation can be beneficial towards motion prediction.

Specifically, existing approaches on global motion estimation based on compressed information deal with video frames that are partitioned in a predefined number of fixed size blocks or do not consider block partitioning characteristics. Block partitioning though, through error minimization, results in block assignment that is determined by the image content itself. Hence, further aspects of block partitioning and specifically block size variability are considered in order to propose an enhanced robust M-estimator approach for global motion estimation that improves the estimation accuracy. Furthermore, it is studied how the parametric representation of global motion can be utilized for accurate global motion prediction and how it can be used to improve conventional motion prediction.

Part of the work presented in chapter 2 and related work in this topic have been presented in:

- . *"Global motion estimation using variable block sizes and its application to object segmentation"*  
M. G. Arvanitidou, A. Glantz, A. Krutz, T. Sikora, M. Mrak, A. Kondoz  
in proceedings of the IEEE International Workshop on Image Analysis for Multimedia Interactive Services, London, 2009
- . *"Compressed domain global motion estimation using the Helmholtz Tradeoff Estimator"*  
M. Tok, A. Glantz, M. G. Arvanitidou, A. Krutz, T. Sikora  
in proceedings of the IEEE International Conference on Image Processing, Hong Kong, 2010

### Chapter 3

Having efficiently compensated motion derived from the camera movement, our goal is to determine the foreground object regions. Therefore the next contribution of this thesis is to investigate aspects of the segmentation process that lack robustness and propose an algorithm that incorporates new modules, dealing with certain deficiencies and improves the overall segmentation performance. In chapter 3, first existing object segmentation approaches are thus overviewed and discussed. A motion based segmentation approach is subsequently proposed yielding improvements in terms of efficiency and being less dependent on empirical parameter settings. Comprehensive evaluation demonstrates the validity of the proposed approach.

Specifically, an approach for bidirectional fusion of global motion compensated errors for inter-frame change detection is proposed. The contributions are in the following; spatial error localization is considered in the thresholding step for improving the segmentation accuracy. As this thresholding scheme introduces one more thresholding parameter, (two thresholds instead of one), an approach for the selection of appropriate parameters is utilized in order to define the weights for the weighted mean using hysteresis thresholding. This enables robust segmentation performance that avoids heuristics and training algorithms for parameter selection that are common approaches. Furthermore, a final post processing step using adaptive filtering according to foreground motion is proposed for mitigating temporal inconsistencies of the segmentation masks.

In addition, a preliminary version of this segmentation framework has been successfully applied in an audio processing framework [22] using audio similarity matrices, for the task of music structure segmentation. Audio similarity matrices [23] visualise the audio structure by its acoustic self-similarity in a two-dimensional representation of time. The image-oriented pre-processing of similarity matrices has proven to be beneficial for highlighting the conveyed musical information and reducing their complexity. This was achieved by handling particular image characteristics that resulted in meaningful spatial segments and thus enhanced the music segmentation.

The work presented in this chapter has in part been presented in:

- . *"Short-term motion-based object segmentation"*  
M. G. Arvanitidou, M. Tok, A. Krutz, T. Sikora  
in proceedings of the IEEE International Conference on Multimedia and Expo, Barcelona 2011
- . *"Motion-based object segmentation using hysteresis and bidirectional inter-frame change detection in sequences with moving camera"*  
M. G. Arvanitidou, M. Tok, A. Glantz, A. Krutz, T. Sikora  
Elsevier Signal Processing: Image Communication Journal, 28 (10), 1420 - 1434, 2013.
- . *"Audio Similarity Matrices Enhancement in an Image Processing Framework"*  
F. Kaiser, M. G. Arvanitidou, T. Sikora



in proceedings of the international workshop on Content-Based Multimedia Indexing, Madrid, 2011.

## Chapter 4

Object segmentation as mentioned above is often seen as a "ubiquitous problem". Humans have a complex way of understanding visual content, and it is even known that some aspects are influenced by individuality. The mechanism behind human perception of visual content is not completely understood by the research community, and the influence of low- and mid-level visual features thereupon is an attractive research topic. Inspired by such a motivation, in the final stage of the thesis, we have been interested in studying how the extracted motion features and the information of moving objects can have a positive impact on objective video quality assessment. Chapter 4 thus investigates visual quality assessment methodologies and describes the proposed approaches towards the improvement of objective algorithms for video quality assessment.

Specifically, the contributions are focused in three directions; a moving object-aware video quality assessment approach is examined employing the proposed moving object segmentation algorithm. Furthermore, a motion saliency model that is beneficial for the exploitation of motion features on spatial level is proposed. Furthermore, an approach for the consideration of global motion that leads to improvement in the temporal dimension is proposed. Apart from motion, the incorporation of other low-level features is additionally investigated and the performance is compared.

The work presented in this chapter has in part been presented in:

. *"Motion saliency for spatial pooling of objective video quality metrics"*

M. G. Arvanitidou, T. Sikora

in proceedings of the International workshop on Quality of Experience for Multimedia Content Sharing, Berlin, 2012

## Chapter 5

In chapter 5 conclusions are drawn, open issues are discussed and future directions in this field are also considered. Finally, appendices A.1, A.2 and A.3 provide a short description of the test datasets used throughout this thesis.



# Global Motion Analysis

---

## Contents

<b>2.1 Introduction</b>	<b>10</b>
2.1.1 Motion modelling in video sequences	10
2.1.2 Global motion estimation	13
2.1.3 Existing approaches on global motion estimation	14
<b>2.2 Improved global motion estimation through variable-size blocks</b>	<b>16</b>
2.2.1 The binary partition tree	16
2.2.2 Least squares estimation	17
2.2.3 Improved robust estimation through block size weighting	19
2.2.4 Improved robust estimation through block size selection	22
<b>2.3 Evaluation of global motion estimation accuracy</b>	<b>22</b>
2.3.1 Test dataset	22
2.3.2 Evaluation methodology	22
2.3.3 Results	23
<b>2.4 Adaptive motion prediction through global motion</b>	<b>34</b>
2.4.1 Introduction and existing approaches	34
2.4.2 Adaptive mode selection	37
<b>2.5 Evaluation of the AMS scheme</b>	<b>38</b>
2.5.1 Test dataset	38
2.5.2 Methodology	38
2.5.3 Results	39
<b>2.6 Chapter summary</b>	<b>40</b>

---

In this chapter we study the effect of block partitioning in the process of parametric global motion estimation. Adopting a parametric model representation, we show that the assignment of weights to the displacements, i.e. the motion vector field, improves the overall performance of global motion estimation. This approach involves two adjacent frames of a video sequence. Furthermore we show that this parametric description can be more accurate than conventional motion prediction.

## 2.1 Introduction

An image sequence, also known as video sequence, is a sequence of still images that creates the illusion of moving regions due to a property of the human visual system to perceive the rapid change of pictures under specific circumstances as motion.<sup>1</sup> In the general case of a video sequence, two kinds of motions may be present; motion induced by the movement of the acquisition device, i.e. camera and motion caused by the captured content, i.e. observed objects. The motion of the camera, which in this work is assumed to be the dominant motion, is widely known as *global motion* and is modelled by parametric transformations of two-dimensional images. The process of estimating the transformation parameters is called *global motion estimation* (GME).

Global motion leads to more efficient motion representation in image regions that can be described by homogenous motion. This can be particularly useful in image regions that are not on the main (contextual) focus and thus belong to the background. When the goal is to detect objects in a scene based on motion, the existence of global motion may mislead unsophisticated object detection algorithms. Therefore, it needs to be compensated and the remaining motion information (local motion) serves as basis for several video processing applications including object segmentation, object tracking and video coding [24, 25, 26]. The challenge is thus to distinguish global and local motion components contained in the captured motion information, which can be achieved by performing an accurate estimation of global motion.

Global motion can be estimated either in the pixel domain or in the block domain. Often, block-based approaches are based on fixed-size blocks while contemporary compression methods tend to use variable-size block schemes during motion estimation. In this chapter we propose an approach for global motion estimation based on motion vector fields that correspond to variable-size blocks. A block matching algorithm which is able to adapt block sizes according to the motion complexity within the frame is used. The resulting motion vectors are weighted according to their assigned spatial area and employed for global motion estimation. Furthermore, preliminary foreground-background masks are created based on the derived frame-by-frame motion compensated differences and by exploiting spatial conditions through anisotropic diffusion filtering.

### 2.1.1 Motion modelling in video sequences

As the word *geometry* (measurement of the earth) indicates, Euclidean geometry is commonly used to describe the three-dimensional world. Nevertheless, being a rather intuitive description of the three-dimensional space, Euclidean geometry adopts that parallel lines do not intersect, or intersect at "infinity". This is in fact an inconvenient

---

<sup>1</sup>This property, known as the *phi phenomenon*, is the optical illusion of perceiving continuous motion between separate objects viewed rapidly in succession. The phenomenon was defined by Max Wertheimer in the Gestalt psychology in 1912 and along with persistence of vision formed a part of the base of the theory of cinema.

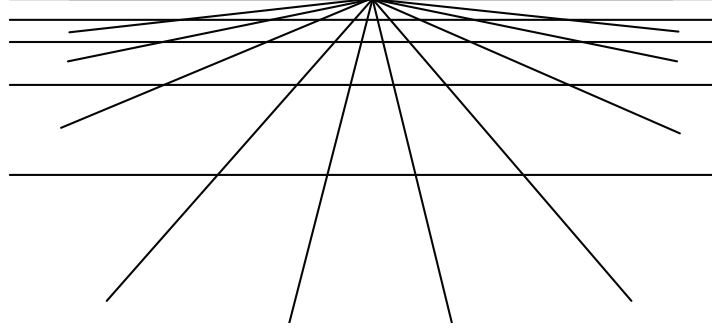


Figure 2.1: Perspective projection example where parallel lines intersect at an ideal point.

assumption for computer vision, where it is often that parallel lines to meet at a specific point. Therefore, *projective geometry* has been established enhancing the Euclidean plane by the addition of points where parallel lines meet, namely the *ideal points* [27], as illustrated in Figure 2.1. Projective geometry, facilitates thus the description of the projection of three-dimensional motion onto two-dimensional planes, and specifically its projection on planar images.

The general transformation between a pair of two-dimensional images in projective geometry, is represented by a homography  $\mathbf{H}$ . To describe motion over a long period sufficiently, high-order polynomial models are needed, such as the parabolic model, which is described by twelve parameters and is suitable to represent parabolic curvature for scene deformations [24]. For consecutive frames, the motion can be sufficiently described using the perspective transformation which has eight independent variables and the homography  $\mathbf{H}$  can be expressed as a  $3 \times 3$  matrix:

$$\mathbf{H} = \begin{pmatrix} m_0 & m_1 & m_2 \\ m_3 & m_4 & m_5 \\ m_6 & m_7 & 1 \end{pmatrix}. \quad (2.1)$$

The perspective transformation can describe translation, rotation, non-uniform scaling and shear between two frames along with projective transformations. The basic transformations that are described by the perspective motion model are illustrated in Figure 2.2. This model is often used to describe motion as a geometric transformation in the image plane since it is considered to be a good trade-off between complexity and accuracy [28]. Another less complex, commonly used model with six independent parameters that cannot describe perspective projections, is

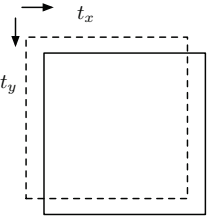
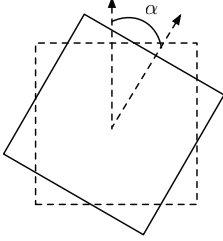
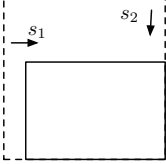
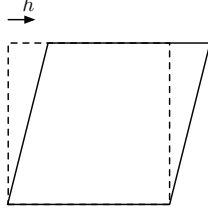
translation	rotation	scaling	shear
$\begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \cos(\alpha) & \sin(\alpha) & 0 \\ -\sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & h & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
			

Figure 2.2: Transformation matrices  $\mathbf{H}$  for translation, rotation, scaling and shear that transform a position  $\mathbf{p} = (x, y, 1)^T$  to a new position  $\mathbf{p}' = (x', y', 1)^T$  by  $\mathbf{p}' = \mathbf{H} \cdot \mathbf{p}$ .

Table 2.1: Transformations allowed in motion models.

Model	Euclidian	Similarity	Affine	Perspective
Independent parameters	2	4	6	8
Translation	•	•	•	•
Rotation	•	•	•	•
Uniform scaling		•	•	•
Nonuniform scaling			•	•
Shear			•	•
Perspective projection				•

referred to as the *affine* model and is formulated as:

$$\mathbf{H} = \begin{pmatrix} m_0 & m_1 & m_2 \\ m_4 & m_5 & m_6 \\ 0 & 0 & 1 \end{pmatrix}. \quad (2.2)$$

Reducing further the complexity, another popular model is the *similarity model* that involves four independent parameters. It is a combination of simple transformations (translation, rotation and uniform scaling) and can be formulated as:

$$\mathbf{H} = \begin{pmatrix} m_0 & m_1 & m_2 \\ -m_1 & m_0 & m_3 \\ 0 & 0 & 1 \end{pmatrix}. \quad (2.3)$$

The most common sub-classes of the perspective model are overviewed in Table 2.1, where the number of independent parameters and the corresponding basic transformations described by each model are also listed.

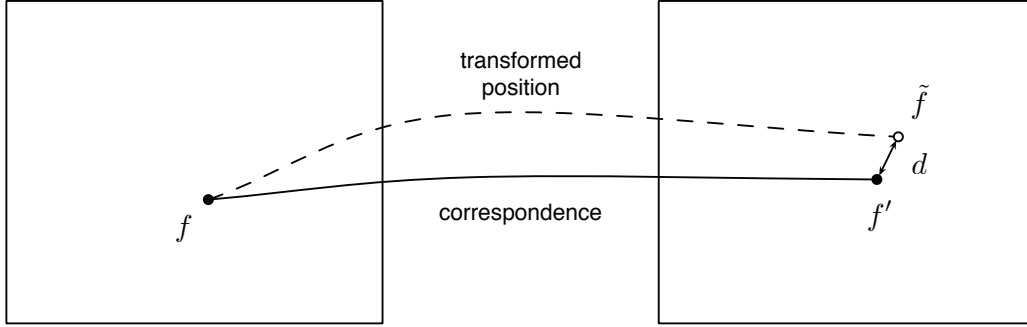


Figure 2.3: Feature matching based image registration. Feature correspondences are used to estimate the image transformation of a feature point  $f$  that minimizes the distance  $d$  between correspondence  $f'$  and transformed image point  $\tilde{f}$ .

### 2.1.2 Global motion estimation

In the case that several motions are present, the goal of global motion estimation is to derive the dominant one, which is assumed to be due to the respective camera movement. The term dominant implies the association with the occupying space rather than the associated magnitude. This is achieved by finding the transformation that best represents motion between a pair of images, which is also referred to as *image registration*.

With respect to the methodology, registration approaches can be generally classified [28] in two general categories: a) feature based and b) photometric consistency-based approaches. In each case the goal is to minimize a registration error metric, by finding the best parametric model which is a priori specified. Feature based approaches, whose general idea is shown in Figure 2.3, try to minimize the registration error between feature correspondences (that are previously tracked) and the transformed points. Photometric consistency-based approaches on the other hand aim at maximizing the correlation between a given image and the transformed one with respect to a parameter vector. The absolute or squared error between the image pair is subsequently used as metric for energy minimization. Approaches in the first category are based on spatial distance, whereas ones in the second category are based on colour or illumination values. In photometric consistency-based approaches in general all regions participate in the estimation and are thus computationally more demanding, in contrast to feature-based ones where only the strong correspondences between two images have impact on the result of the parameter computation. Therefore, in this work we are going to deal with feature based approaches, which are also an appropriate solution in the case that motion vectors are used.

In feature based approaches, in order to estimate the model that optimally describes the transformation between a pair of images, the following steps are performed: a) feature extraction b) establishment of correspondences based on a matching criterion, such as sum of absolute error (SAD) or sum of squared differences

(SSD) and finally c) estimation of the best fitting transformation. Features may be estimated based on pixel correspondences or on block correspondences, determined by the smallest area of the image that a feature correspondence can be assigned to. In the latter case blocks may have fixed- or variable-size.

### 2.1.3 Existing approaches on global motion estimation

In the literature, there exist a variety of global motion estimation approaches that are based on pixel level. In [8] the dominant motion between two successive frames is estimated as an affine model using the spatial intensity gradient, for shot change detection. In [9] the approach is based on image pyramid decomposition, which is shown to improve the performance of gradient descent. The global motion is modelled by a perspective motion model and estimated using a robust gradient-based technique. The approach is hierarchical and consists of three stages; after building a low pass image pyramid and estimating the initial translation model, in the last stage gradient descent is employed at each pyramid level and the affine motion model is finally obtained using an M-estimator in N iterative procedure to reduce the influence of the outliers. Similarly in [29] the previous technique is enhanced using a window approach and phase correlation for initialisation. In [26] a hierarchical approach has been proposed with an initial translation model estimation using Newton-Raphson error minimization. In the final stage, outliers that have been previously found to belong to the foreground object are rejected in order to increase robustness. Finally an affine model is obtained and exploited for object segmentation.

Considerable effort has been also given in the last years on exploiting - exclusively or in addition to pixel information - compressed domain data. As mentioned before, in video coding environments, motion related information may be extracted from the encoded stream. If this information is exploited, the computational cost of calculating the motion vectors, i.e. block matching, can be avoided and therefore methods based on block accuracy have an obvious advantage over pixel based methods, where the calculation of each pixel correspondence has to be performed. Image registration approaches based on pixel correspondences may be regarded as the ones with the highest potential for achieving high accuracy, however, it has been shown that they are often outperformed by block based approaches [12]. This may be attributed to the fact that motion vectors may (under circumstances) serve as accurate descriptors of motion, in addition to the fact that outlier rejection is a critical stage in global motion estimation, and determines strongly the performance of global motion estimation.

In video coding scenarios where the image sequence is encoded and decoded in order to be transmitted or stored, motion vectors which represent the displacement of predefined areas, namely blocks, between image pairs may be extracted from the encoded stream. Motion estimation techniques for the calculation of motion vectors can be very computational demanding, therefore efficient implementations with respect to computational complexity have been proposed, such as the enhanced



predictive zonal search algorithm (EPZS) [30], which is also employed in the state-of-the-art standard H.264/AVC. Often, fixed-size blocks e.g. blocks of  $16 \times 16$  pixels are employed, as in the MPEG-2 standard. Currently compression techniques, such as H264/AVC [21] and high efficiency video coding (HEVC) [20], use variable-size blocks instead of fixed size blocks during motion estimation and compensation. This allows for greater adaptivity to scene content as block sizes are adapted according to motion characteristics and complexity, which subsequently leads to higher compression efficiency. Regions with complex motion, which require multiple motion vectors in order to describe existing motion, are assigned small blocks, especially when high bit-rates are allowed. On the other hand, large blocks are used for regions that correspond to homogenous motion, often belonging to the background, or in case exceptionally low bit-rates are targeted.

Smolic *et al.* presented [31] a global motion estimation algorithm based on motion vectors using M-estimator for outlier rejection. Su *et al.* [10] used the Newton-Raphson method to estimate the motion model based on motion vectors and introduced an adaptive motion model selection. The adaptive motion model selection relies on adaptively choosing a parametric model, from two up to eight parameters, in order to save computational time. Another compressed domain approach is presented in [32] proposing to use discrete cosine transform (DCT) coefficients instead of motion vectors from the MPEG stream to compute subsampled representations of the initial images. Subsequently, feature extraction using parametric representations, based on the generalised Hough transform was proposed. The approach targeted at detecting predefined camera movements, namely pan, tilt, rotation and zoom, with application to video annotation. In [33], the authors compare several feature detection and matching algorithms and achieve a robust estimation of the motion model based on RANSAC. In [12] a two step-simplification scheme is used for robust regression approaches such as RANSAC and the Helmholtz tradeoff estimator towards complexity reduction. Approaches that fuse block and pixel information have been also reported. Chen *et al.* [34] proposed a combined compressed and pixel-domain approach, where Markov random field classification is used at first place for coarse segmentation followed by pixel information incorporation to extract colour and edge information for object boundary refinement.

Towards enhancing the robustness of global motion estimation, existing approaches try to exploit object segmentation information. Even though it would be very convenient, this information is usually not a priori available. Therefore, information regarding object boundaries, shall be assumed at the beginning of the global motion estimation process or it may be derived from global motion and used later towards enabling more accurate estimation. In fact, global motion estimation and object segmentation are considered as interdependent information, which has led to considering these two problems as a *chicken and egg* riddle [35, 11, 18, 19].

In line with this idea the authors in [36] adopted a block-based approach that assigned a motion vector to each block of the image, estimated a four-parameter global motion model and then performed an iterative elimination of foreground blocks towards refinement of the global motion model. In [37] linear regression and

a threshold decision were combined towards eliminating mismatching motion vectors belonging to the foreground object, according to a prior segmentation step in an iterative manner. Based on this, in the work presented in [38] discrete curve evolution was used in order to subdivide the motion trajectory in sub-trajectories with constant motion towards video shot detection. In [39] a pre-processing cascade filter for motion vectors has been applied in order to enhance global motion estimation and segmentation accuracy by iterating between these two interdependent procedures.

The techniques mentioned above that estimate global motion based on compressed video sequences deal with video frames that are partitioned in a predefined number of fixed-size blocks or they do not consider block partitioning characteristics in the global motion estimation process. Block partitioning though, through error minimization, results in block assignment that is determined by the image content itself. Hence global motion estimation could benefit from considering further aspects of block partitioning. In the following, we describe the proposed approach that exploits block size variability towards more accurate global motion estimation.

## 2.2 Improved global motion estimation through variable-size blocks

The proposed global motion estimation approach is based on the work of [31] where the influence of outliers is reduced in an iterative process based on the estimation error obtained using least-squares estimation. Further, we make the observation that smaller blocks are assigned by motion estimation to regions with high presence of edges, while larger blocks are assigned to homogenous regions. This may result in associating the block size with the classification of the block in the background or in the foreground region. Considering that this observation may be beneficial for global motion estimation, in the following we study the influence of block size variability, and based on this observation we propose an improved robust M-estimator approach.

The blocks are obtained using the binary partition tree which is described in the following and has demonstrated promising results in video coding. The binary partition tree model has been shown that it is suitable for application in 3D video coding [40], especially for depth-maps, since it enables excellent adaptability to the actual image content. Its main advantage is its capability to partition the frame along the actual motion boundaries.

### 2.2.1 The binary partition tree

A model that offers high flexibility of frame partitioning, and therefore better adaptation to the actual content, is the *binary partition tree* (BPT) [41]. This approach enables adaptive partitioning of video frames, originally motivated by rate-distortion optimization requirements in compression.

In schemes based on variable-size blocks, such as the BPT, the block sizes are

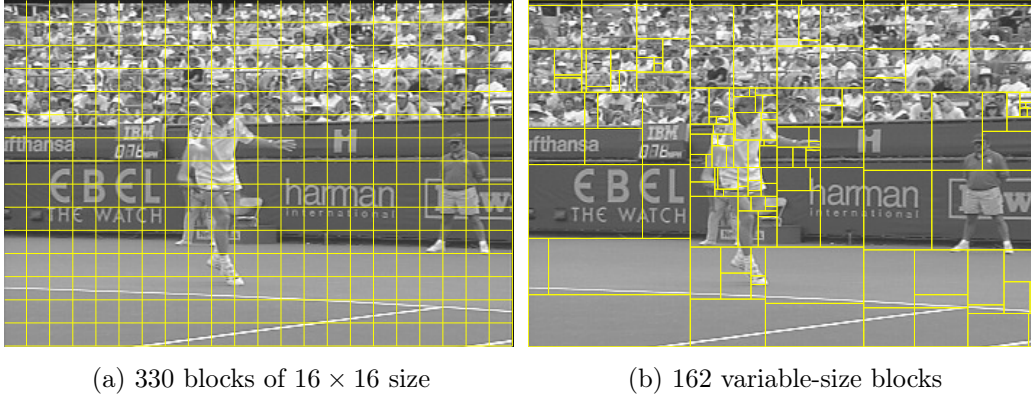


Figure 2.4: Block partitioning of frame 22 of the *Stefan* sequence, using (a) block matching algorithm and (b) the binary partition tree.

not predefined. Therefore, they vary in order to optimise the trade-off between the number of bits used to encode motion vectors and the residual (rate-distortion optimization requirements). The partitioning of a frame into blocks is described with a tree-structure and can be achieved using a two-step algorithm. First step is the growing of the tree by frame partitioning (top-down step). Second step is the pruning of the tree which finds the optimal partitioning with respect to given requirements (bottom-up step).

During the tree growing step, the entire picture is repeatedly split up to a target number of  $N$  blocks. Initially the whole frame is considered as one block. Optimal partitioning is achieved using motion estimation and its actual partitioning is described in the tree root and with its two new "branches" that represent two new blocks. Then the iterative procedure continues. At the bottom up step, the tree is pruned in order to find the optimal partitioning, towards complying with rate-distortion optimization requirements. Figure 2.4 illustrates an example, where as can be seen the use of variable-size blocks, according to the binary partition tree model allows for greater adaptivity to local scene content. The BPT model yields less blocks that vary in size and shape.

### 2.2.2 Least squares estimation

Given the motion vector field between two frames of a video sequence, the goal is to find the optimal parametric model that fits to it and which is considered to be associated with the global motion. To achieve this, we formulate the problem by fitting the parametric model using least squares estimation, minimizing the distance (error) between the sampled pixel values and the estimated ones.

The perspective model can be regarded as a suitable approximation of the global motion over a short period of time. The transformation  $T$  of a given point  $\mathbf{p} = (x, y)^T$  to a new position  $T(\mathbf{p}, \mathbf{H}) = (x', y')^T$ , based on a homography  $\mathbf{H}$ , is given as follows:

$$x' = \frac{m_0 + m_1x + m_2y}{1 + m_6x + m_7y} \quad (2.4)$$

$$y' = \frac{m_3 + m_4x + m_5y}{1 + m_6x + m_7y}. \quad (2.5)$$

For the  $n$ -th frame in an image sequence, the transformed position is estimated based on the  $(n-1)$ -th frame and the corresponding parametric model  $\mathbf{H}_n^{n-1}$ . The error at a certain pixel position  $j$  is thus measured as:

$$\varepsilon_j = \mathbf{p} - T(\mathbf{p}, \mathbf{H}_n^{n-1}). \quad (2.6)$$

$\varepsilon_j$  consists of the error in horizontal and vertical direction respectively, combined using the L1 norm. In the following, frame indices are omitted for brevity and  $\mathbf{H}$  denotes  $\mathbf{H}_n^{n-1}$  except otherwise mentioned.

For a frame consisting of  $L$  pixels, the optimal parametric model is given by the homography that minimizes the quadratic error,

$$\mathbf{H} = \arg \min_L \sum_j \varepsilon_j^2. \quad (2.7)$$

The transformation  $T$  of (2.4),(2.5) may be written as:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & x & y & 0 & 0 & 0 & -x \cdot x' & -y \cdot x' \\ 0 & 0 & 0 & 1 & x & y & -x \cdot y' & -y \cdot y' \end{pmatrix} \cdot \mathbf{H} \quad (2.8)$$

where  $\mathbf{H}$  is the rewritten homography model of (2.1):

$$\mathbf{H} = (m_0 \ m_1 \ m_2 \ m_3 \ m_4 \ m_5 \ m_6 \ m_7)^T \quad (2.9)$$

or equally

$$\begin{pmatrix} x' - x + x \\ y' - y + y \end{pmatrix} = \begin{pmatrix} MV_x + x \\ MV_y + y \end{pmatrix} = \begin{pmatrix} 1 & x & y & 0 & 0 & 0 & -x \cdot x' & -y \cdot x' \\ 0 & 0 & 0 & 1 & x & y & -x \cdot y' & -y \cdot y' \end{pmatrix} \mathbf{H}. \quad (2.10)$$

If  $(MV_{x_i}, MV_{y_i})$  denote the motion vector coordinates associated with the  $i$ -th block, out of  $N$  participating blocks, of a given frame and  $(x_i, y_i)$  and  $(x_i', y_i')$  are the center coordinates of the reference and the transformed block respectively, then the relation between the perspective motion parameters and the motion vectors is transformed to:

## 2.2. Improved global motion estimation through variable-size blocks 19

$$\begin{pmatrix} MV_{x_1} + x_1 \\ MV_{y_1} + y_1 \\ \vdots \\ MV_{x_N} + x_N \\ MV_{y_N} + y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & y_1 & 0 & 0 & 0 & -x_1 \cdot x_1' & -y_1 \cdot x_1' \\ 0 & 0 & 0 & 1 & x_1 & y_1 & -x_1 \cdot y_1' & -y_1 \cdot y_1' \\ & & & & & \vdots & & \\ 1 & x_N & y_N & 0 & 0 & 0 & -x_N \cdot x_N' & -y_N \cdot x_N' \\ 0 & 0 & 0 & 1 & x_N & y_N & -x_N \cdot y_N' & -y_N \cdot y_N' \end{pmatrix} \cdot \mathbf{H} \quad (2.11)$$

which using matrix notation is formulated as:

$$\mathbf{V} = \mathbf{D} \cdot \mathbf{H}. \quad (2.12)$$

The least squares solution of (2.12) with respect to  $\mathbf{H}$  is

$$\mathbf{H} = (\mathbf{D}^T \cdot \mathbf{D})^{-1} \cdot \mathbf{D}^T \cdot \mathbf{V} \quad (2.13)$$

The least squares approach relies on an explicit formulation of the desired parametric model, is however sensitive to outliers [14]. The presence of moving foreground objects whose motion differentiates from the global one as well as errors coming from other sources result in outliers that cause inaccurate global motion estimation and have to be discarded. Therefore, the incorporation of a robust outlier rejection approach is necessary. In the following, the proposed improved robust M-estimator approach that exploits size variability towards the elimination of the outliers' impact in global motion estimation is described.

### 2.2.3 Improved robust estimation through block size weighting

The least squares approach is straightforward and results in a first estimation of the fitted model. In case that the errors are independent and normally distributed with constant deviation, it is in fact a realistic solution [42]. Nevertheless, the existence of outliers cannot justify this assumption anymore.

Outliers may occur due to image noise (measurement errors), lighting changes, moving objects, model failures or mismatches due to homogenous surfaces, and have as consequence the global motion estimation to yield unreliable results. In order to eliminate the influence of outliers, namely of the correspondences that do not comply with the transformation model, a robust *M-estimator* is employed in [31]. We propose here an enhancement in terms of robustness for the M-estimator, by reducing the influence of outliers in a re-weighted iteration procedure based on the surface of the participating blocks. The procedure is described in the following.

The M-estimator, which has been introduced by Huber [13] reduces the influence of outliers in a re-weighted iteration procedure based on the estimation error

obtained using least squares estimation. It aims at finding the parametric model  $\mathbf{H}_{\text{opt}}$  that minimizes the residual function over the  $N$  pairs of points:

$$\mathbf{H}_{\text{opt}} = \arg \min_N \sum_i^N \rho(\varepsilon_i) \quad (2.14)$$

where  $\rho(\varepsilon)$  is a function of the residual between data samples and estimated values. Function  $\rho(\varepsilon)$  determines the robustness of the estimation, and in the case of least squares minimization it is  $\rho(\varepsilon) = \varepsilon^2$ . This results in larger values  $\varepsilon_i$  to have a heavier impact on the minimization. The idea of M-estimation is to eliminate the influence of extreme large error terms by forming an appropriate function  $\rho(\varepsilon)$ . Towards this direction, Tukey introduced a bisquare estimator, also known as *biweight* estimator, that enforces a radical elimination of outliers' impact. Thereby the function  $\rho(\varepsilon)$  is defined as:

$$\rho(\varepsilon_i) = \begin{cases} \frac{c^2}{6} \left[ 1 - \left[ 1 - \left[ \frac{\varepsilon_i}{c\mu_\varepsilon} \right]^2 \right]^3 \right] & \varepsilon_i < c\mu_\varepsilon \\ \frac{c^2}{6} & \varepsilon_i \geq c\mu_\varepsilon \end{cases} \quad (2.15)$$

where  $\varepsilon_i$  is the estimation error of the block in the  $j$ -th position,  $c$  is a tuning constant and  $\mu_\varepsilon$  is the mean estimation error over all blocks of the frame. The solution of equation (2.14), is given by differentiation of (2.15) with respect to the transformation parameter matrix and set to zero. Assigning  $\omega(\varepsilon_i) = \rho'(\varepsilon_i)/\varepsilon_i$  yields the weighting function:

$$\omega(j) = \begin{cases} \left[ 1 - \left[ \frac{\varepsilon_i}{c\mu_\varepsilon} \right]^2 \right]^2 & \varepsilon_i < c\mu_\varepsilon \\ 0 & \varepsilon_i \geq c\mu_\varepsilon \end{cases} \quad (2.16)$$

This weighting function is used to introduce a diagonal weighting matrix  $\mathbf{W}$  in equation (2.13) that subsequently becomes:

$$\mathbf{H} = (\mathbf{D}^T \cdot \mathbf{W} \cdot \mathbf{D})^{-1} \cdot \mathbf{D}^T \cdot \mathbf{W} \cdot \mathbf{V} \quad (2.17)$$

where

$$\mathbf{W} = \text{diag}(\omega(1) \ \omega(1) \ \omega(2) \ \omega(2) \ \dots \ \omega(N) \ \omega(N)) \quad (2.18)$$

contains the weights that correspond to each motion vector pair. Each component  $\omega(n)$  appears twice because it affects two rows of  $\mathbf{D}$  within the computation of the transformation model. This procedure is conducted iteratively until the weights converge and  $\mathbf{W}$  is calculated for each iteration  $k$ . The influence of outliers, i.e. the weight  $\omega(i)$  for the motion vector corresponding to the  $i$ -th block at each iteration  $k$  is decreased with the estimated error as determined by equation (2.16).

## 2.2. Improved global motion estimation through variable-size blocks 21

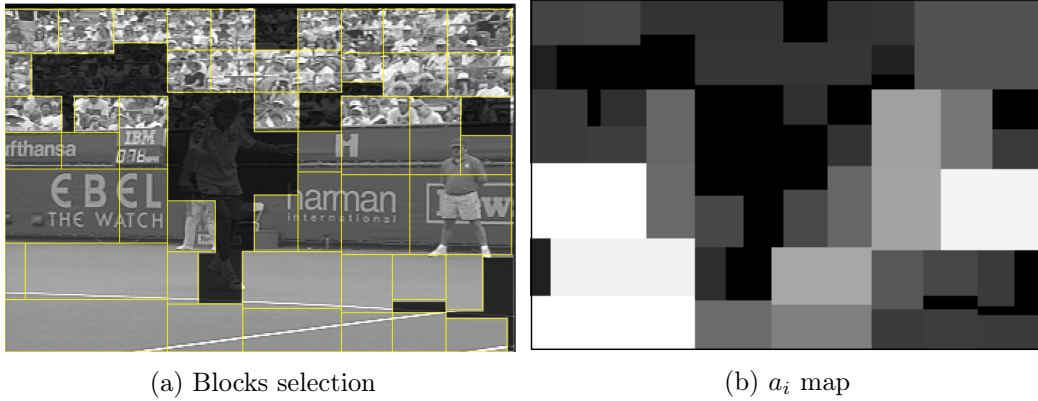


Figure 2.5: Blocks of frame 22 of the *Stefan* sequence depicted in Figure 2.4(b). (a) Selected blocks after discarding the black shaded ones, (b) assigned weights  $a_i$  depicted as grayscale intensity values. The greater the intensity (brighter tone) the higher the assigned weight  $a_i$ .

As shown in Figure 2.4(b), the binary partition tree model yields variable-size blocks; large ones that correspond to homogeneous areas (e.g. tennis court) and tend to belong to the background and smaller ones that correspond to regions with smaller details (e.g. tennis player) that tend to belong to moving foreground objects. The proposed approach, named *VSBasel*, reduces the influence of outliers according to the estimation error by taking advantage of this observation. This procedure is performed per frame, considering the frame's partitioning for the weight assignment. A similar idea has been shown in [43] but in this case the weight values were fixed and no adaptive weighting per frame was taken into account.

Let  $M$  be the total number of blocks that a frame is divided into. We consider only  $N$  blocks whose surface  $S_i$  is  $S_i > \mu_s$  out of  $M$  blocks, where:

$$\mu_s = \frac{1}{M} \sum_i^M S_i. \quad (2.19)$$

Subsequently we assign a weight  $a_i$  to each motion vector in position  $i$  out of the  $N$  participating ones:

$$a_i = \frac{S_i}{\mu_s}. \quad (2.20)$$

In this way the participation of each motion vector in the least squares solution is adapted according to the size of the block partition in relation to the existing partitions of the examined frame.

Thus equation (2.17) becomes:

$$\mathbf{H} = (\mathbf{D}^T \cdot \mathbf{A} \cdot \mathbf{W} \cdot \mathbf{D})^{-1} \cdot \mathbf{D}^T \cdot \mathbf{A} \cdot \mathbf{W} \cdot \mathbf{V} \quad (2.21)$$

where  $\mathbf{A}$  is the weight matrix that incorporates the associated block size in the

calculation, excluding basically blocks that correspond to the moving foreground objects.

### 2.2.4 Improved robust estimation through block size selection

A subcase of VSBasel is when  $N$  blocks whose surface  $S_i$  is  $S_i > \mu_S$  out of  $M$  blocks are considered, which have uniform influence on the global motion estimation. This approach is named *VSBsel*, and in this case  $a_i = 1$  for each motion vector in position  $i$  out of the  $N$  participating ones.

An example of the weight assignment per block is illustrated in Figure 2.5 referring to the example of Figure 2.4(b). Figure 2.5 depicts the selected blocks (VSBsel),  $N = 47$  out of  $M = 162$ , as well as the corresponding assigned weights in grayscale (VSBasel). It can be observed that blocks that have a higher impact on global motion estimation are mostly located in the background region and especially in the region of the tennis court, while regions that have a slight or zero impact are mostly located on the foreground region.

## 2.3 Evaluation of global motion estimation accuracy

### 2.3.1 Test dataset

For evaluation of the global motion estimation accuracy, we use the *Stefan*, *Mountain* and *Biathlon* sequences which are described in appendix A.2. The camera motion in these test sequences is described by a combination of pan, tilt and zoom, while the *Stefan* sequence is especially complex both in terms of colour and motion. It consists of low-frequency regions such as the tennis court as well as high-frequency ones such as the audience. As our goal is to evaluate the estimation in the background region of an image sequence, it is reasonable to exclude moving foreground objects for the evaluation. For the above mentioned image sequences we employ thus manually created ground-truth (examples are given in A.2), and perform evaluation only in the background as explained below.

### 2.3.2 Evaluation methodology

The following algorithms are evaluated in this section:

- **FSB** Fixed-size blocks, the algorithm proposed in [31] where fixed-size  $8 \times 8$  blocks, created by typical block matching, are used. All blocks (i.e. the corresponding motion vectors) participate in the estimation of global motion using M-estimator.
- **VSB** Variable-size blocks, the algorithm proposed in [31] where variable-size blocks are employed that have been created by the binary partition tree and all blocks contribute to global motion estimation.



- **VSBsel** Variable-size blocks selection refers to the proposed approach (subsection 2.2.4) that has been presented in [44]. Out of the  $M$  blocks of a frame, only  $N < M$  blocks, whose surface is  $S_i > \mu_S$  according to equation (2.19), are uniformly considered. Thus, small blocks that possibly correspond to foreground objects are excluded. It is therefore a binary decision of using or discarding motion vectors and no weighting is involved.
- **VSBasel** Variable-size blocks adaptive selection refers to the proposed approach, described in subsection 2.2.3, where a selection of  $N < M$  blocks as previously is considered that are additionally assigned a weight  $a_i$  according to equation (2.20) in the improved robust M-estimator approach. Thereby the participation of each motion vector in the least squares solution is adapted according to the size of the block partition in relation to the existing partitions of the examined frame.

It is noted that in the least squares solution proposed in [31] (FSB algorithm), a four-parameter motion model was used. Here, we employ the more accurate eight-parameter perspective motion model into the FSB algorithm as well as for the VSB, VSBsel and VSBasel algorithms in order to have a fair basis for comparison.

In order to evaluate the motion estimation performance, we use the following approach: given a frame  $I_n$  and the reference frame  $I_{n-1}$  of an image sequence, we employ the estimated parametric model between them to obtain an estimation of  $I_n$ , namely  $\tilde{I}_n^{n-1}$ . Bilinear interpolation and half-pixel accuracy are subsequently used for image warping. As the goal is to estimate global i.e. background motion, the evaluation is performed comparing the background regions of  $I_n$  and  $\tilde{I}_n^{n-1}$ . This is achieved using *background peak signal to noise ratio (bPSNR)* that is calculated as:

$$bPSNR = 10 \cdot \log_{10} \frac{\Lambda^2}{bMSE} \quad (2.22)$$

and

$$bMSE = \frac{1}{XY} \sum_{x=1}^X \sum_{y=1}^Y [I_n(x, y) - \tilde{I}_n^{n-1}(x, y)]^2 \quad (2.23)$$

where  $\Lambda$  is the maximum intensity pixel value (typically 255) and every pixel pair  $(x, y)$  belongs to the background region  $\mathcal{R}_{b_n}$  of frame  $n$ , as defined by the corresponding ground-truth mask.

### 2.3.3 Results

Table 2.2 presents the evaluation of the accuracy of global motion estimation in terms of bPSNR. The proposed VSBasel algorithm outperforms the reference algorithm FSB and it also outperforms the VSB and VSBsel algorithms. In the case of the VSB algorithm, the uniform participation of all motion vectors derived from the BPT model causes a decrease in bPSNR compared to the case of fixed-size blocks (FSB algorithm). This is attributed to the fact that the BPT assigns motion vectors

to very small regions, especially on the borders and on the foreground regions that participate equally in the GME. By discarding such areas, illustrated as dark shaded blocks in the example of Figure 2.5(a), the GME performance improves, resulting in the VSBsel algorithm to bring higher performance in terms of bPSNR. Assigning weights that are adapted according to each frame's block partitioning, i.e. in the VSBasel algorithm, improves further the performance of GME. Figures 2.6 - 2.8 illustrate the bPSNR per frame for the examined algorithms, for the *Stefan*, *Biathlon* and *Mountain* sequences respectively.

**Amount of participating blocks** An interesting point to be studied is the relation between the amount of participating blocks in the parametric global motion estimation and the corresponding improvement. As explained in section 2.2.1 the binary partition tree enables free selection of block sizes, and its main advantage is the capability to partition each frame along actual motion boundaries. The number of produced blocks is thus not predefined and may significantly vary according to the image detail. Figures 2.9 - 2.11 show the improvement in terms of bPSNR of VSBasel over VSB together with the percentage of participating blocks over frames. The percentage of selected blocks refers to the number of selected blocks  $N$  out of  $M$  available blocks per frame in the global motion estimation process. Studying the relation between the relative amount of blocks in each frame that are selected to participate in the estimation process, we observe that the increasing percentage of employed blocks tends to have a negative impact on the quality of global motion estimation. Comparing the bPSNR values over time (frames) and the percentage of selected blocks over time, we observe negative correlation, specifically:  $-0,444$ ,  $-0,167$ ,  $-0,572$  for the *Stefan*, *Biathlon* and *Mountain* sequences respectively. The negative correlation indicates that the fewer blocks selected, the better the estimation.

**Size of participating blocks** Next, we examine the relative block size of the selected blocks employed for GME in each algorithm under examination. Table 2.2 reports additionally, for each algorithm and each test sequence, the rescaled relative block size (RRBS) that refers to the ratio of the average size of the employed blocks over the average size of all available blocks in each frame (scaled here by a factor of two, to enable comparison). In the case of the VSBasel algorithm, we observe positive correlation with the corresponding bPSNR improvement, specifically:  $0,571$ ,  $0,110$ ,  $0,502$  for the *Stefan*, *Biathlon* and *Mountain* sequences respectively. Figures 2.12 - 2.14 present bPSNR and RRBS in more detail for each test sequence.

These two points allow us to observe that higher percentage of participating blocks per frame results in lower accuracy of GME, while higher relative block size results in higher accuracy of GME. It may be thus concluded that in cases where the BPT model results in few large blocks, more accurate global motion estimation can be achieved compared to cases where many small blocks are involved in global motion estimation. This is consistent with the initial intuition that large blocks are

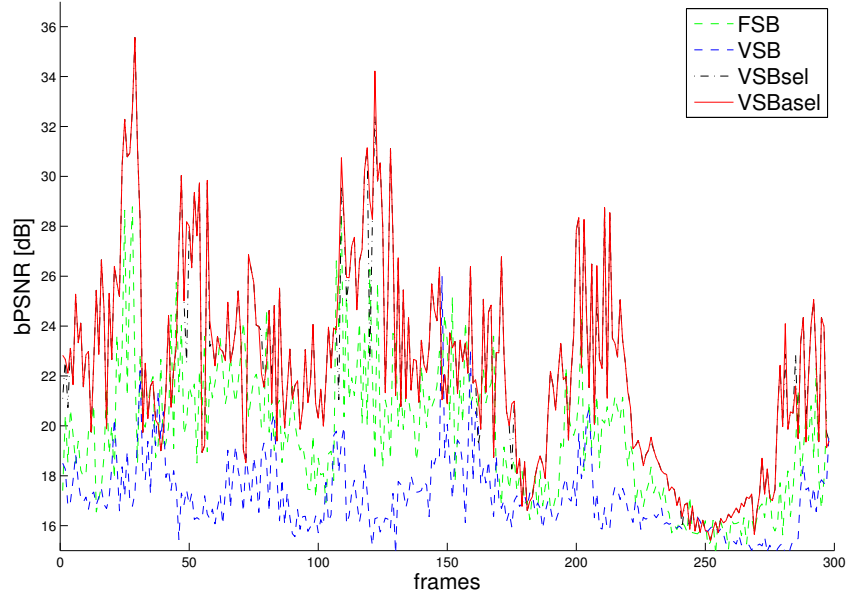


Figure 2.6: Background PSNR in dB for the *Stefan* sequence for the proposed approach (VSBasel) and reference ones (Algorithms FSB, VSB and VSBsel).

mainly observed in the background region, whereas smaller ones are mainly observed to the foreground region.

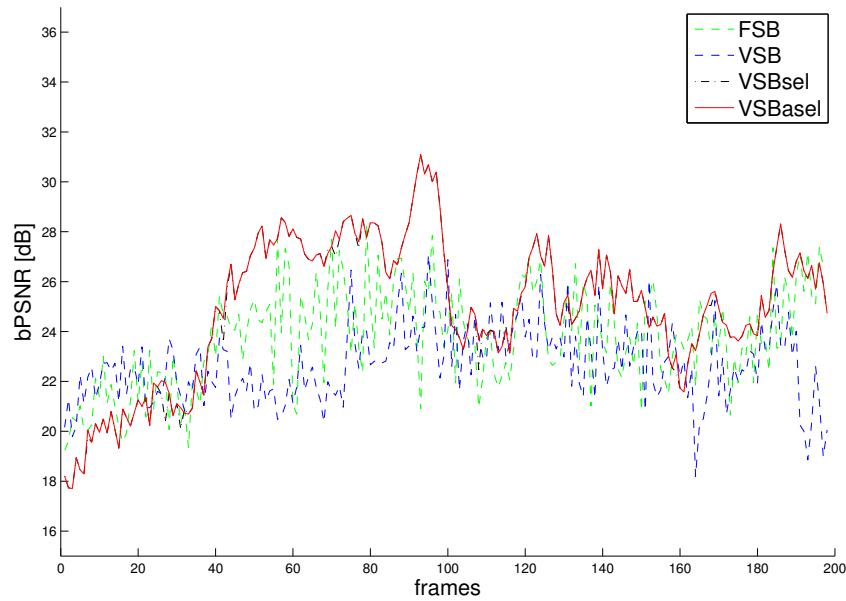


Figure 2.7: Background PSNR in dB for the *Biathlon* sequence for the proposed approach (VSBasel) and reference ones (Algorithms FSB, VSB and VSBsel).

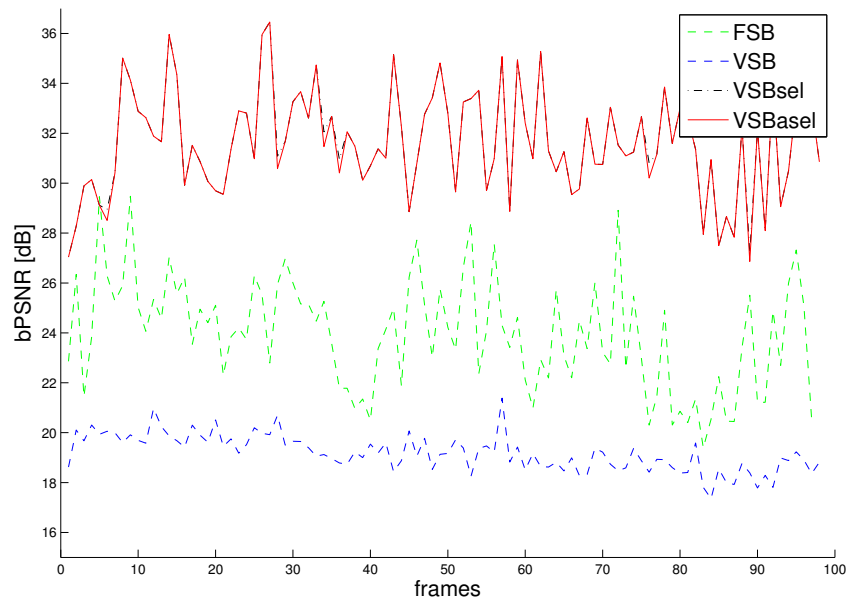


Figure 2.8: Background PSNR in dB for the *Mountain* sequence for the proposed approach (VSBasel) and reference ones (Algorithms FSB, VSB and VSBsel).

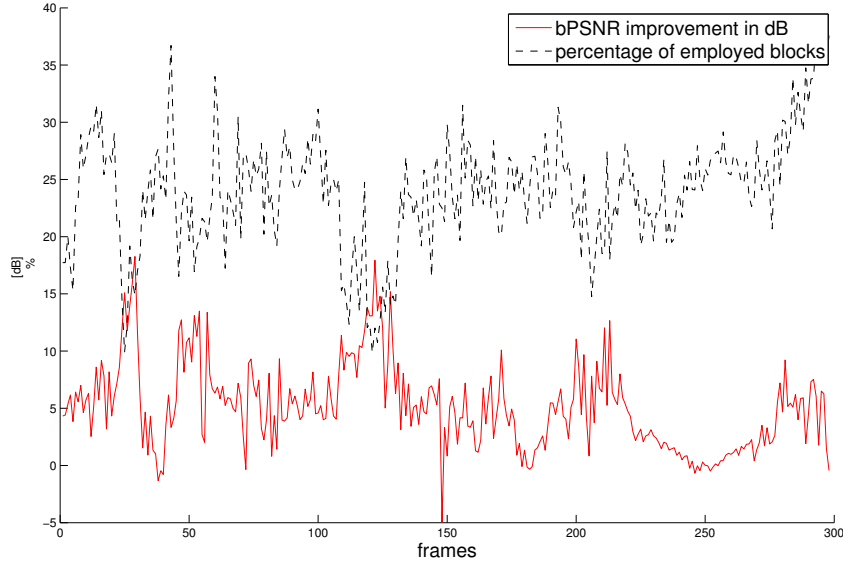


Figure 2.9: Background PSNR improvement for the *Stefan* sequence using the proposed VSBasel compared to the reference VSB algorithm. The solid red line indicates the background PSNR improvement, whereas the dotted black line shows the percentage of employed blocks  $N/M$  per frame.

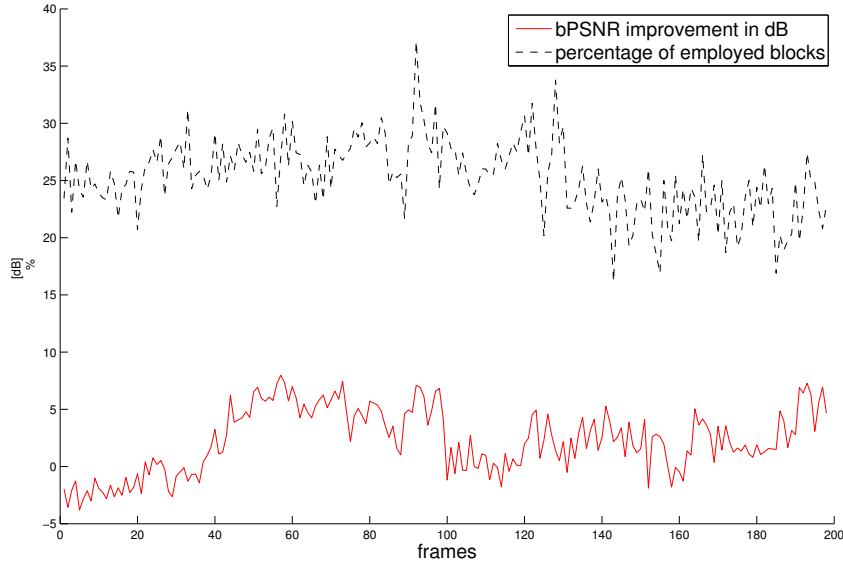


Figure 2.10: Background PSNR improvement for the *Biathlon* sequence using the proposed VSBasel compared to the reference VSB algorithm. The solid red line indicates the background PSNR improvement, whereas the dotted black line shows the percentage of employed blocks  $N/M$  per frame.

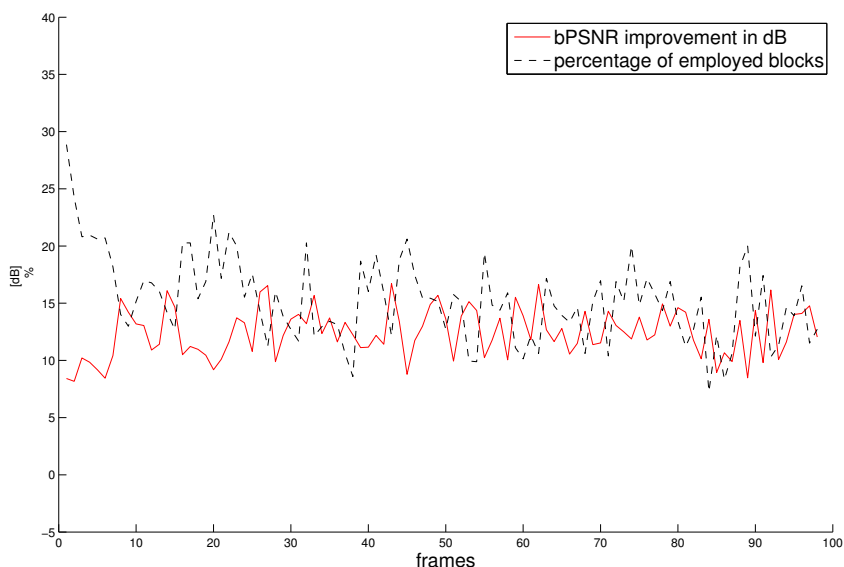


Figure 2.11: Background PSNR improvement for the *Mountain* sequence using the proposed VSBasel compared to the reference VSB algorithm. The solid red line indicates the background PSNR improvement, whereas the dotted black line shows the percentage of employed blocks  $N/M$  per frame.

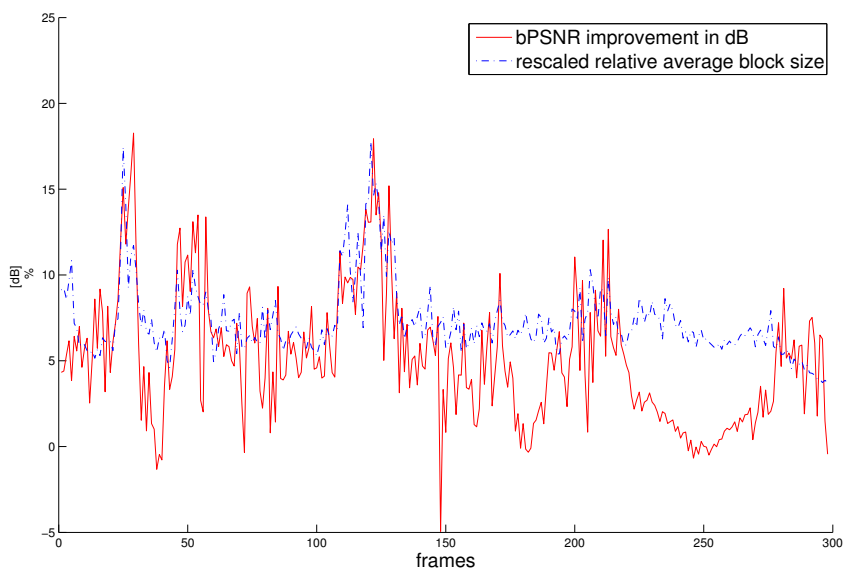


Figure 2.12: The solid red line indicates the background PSNR improvement for the *Stefan* sequence using the proposed VSBasel compared to the reference VSB algorithm, whereas the dotted blue line shows the relative average size of employed blocks (scaled by factor two) per frame.

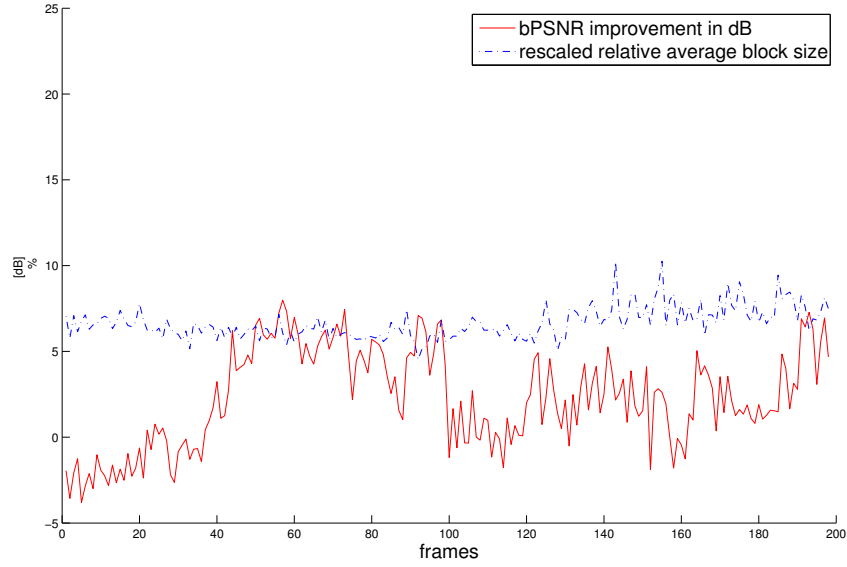


Figure 2.13: The solid red line indicates the background PSNR improvement for the *Biathlon* sequence using the proposed VSBasel compared to the reference VSB algorithm, whereas the dotted blue line shows the relative average size of employed blocks (scaled by factor two) per frame.

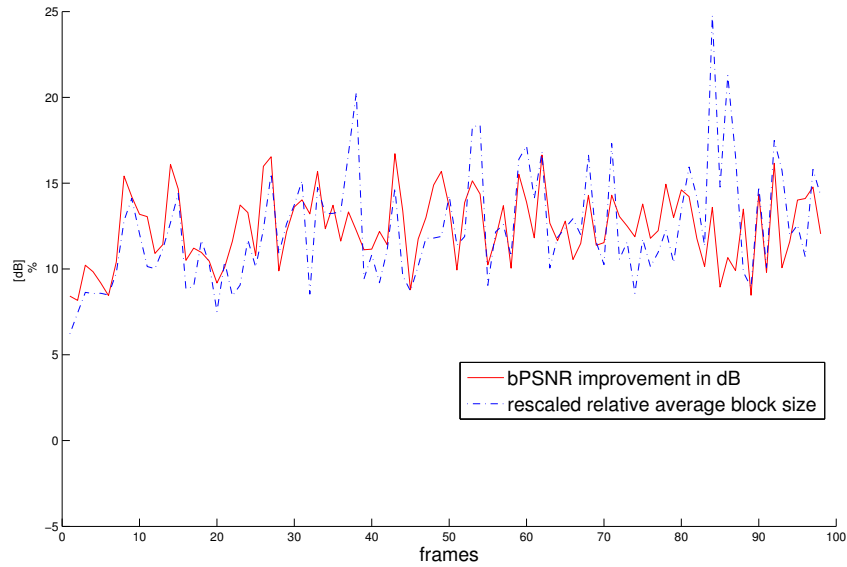


Figure 2.14: The solid red line indicates the background PSNR improvement for the *Mountain* sequence using the proposed VSBasel compared to the reference VSB algorithm, whereas the dotted blue line shows the relative average size of employed blocks (scaled by factor two) per frame.

Table 2.2: Mean background PSNR (in dB) and average block size (in pixels) of the participating blocks (RRBS) comparing reference and proposed algorithms. FSB, VSB are the reference algorithms, whereas VSBsel, VSBasel are the proposed ones. The two best performances are highlighted with boldface.

Sequence	Algorithm	RRBS	bPSNR [dB]
<i>Biathlon</i>	FSB	64	25.380
	VSB	513.14	23.041
	VSBsel	1710.79	<b>27.685</b>
	VSBasel	1710.79	<b>27.770</b>
<i>Mountain</i>	FSB	64	23.894
	VSB	380.19	19.149
	VSBsel	2321.37	<b>31.434</b>
	VSBasel	2321.37	<b>31.403</b>
<i>Stefan</i>	FSB	64	19.836
	VSB	394.04	17.188
	VSBsel	1483.09	<b>22.761</b>
	VSBasel	1483.09	<b>22.864</b>
average	FSB	64	23.037
	VSB	429.12	19.793
	VSBsel	1838.42	<b>27.293</b>
	VSBasel	1838.42	<b>27.346</b>



### Subjective evaluation based on segmentation performance

Once the parametric model  $\mathbf{H}$  is computed for every pair of adjacent frames, the subtraction of the estimated frame  $\tilde{I}_n^{n-1}$  from the current frame  $I_n$  yields the global motion compensated frame  $D_n = I_n - \tilde{I}_n^{n-1}$ . Subsequently, we apply an object segmentation algorithm based on the global motion compensated frames  $D_n$  that are derived from the examined algorithms. Towards this goal, we employ the segmentation algorithm described in [45], which is summarized in Table 2.3.

Accurate compensation of global motion allows for precise object segmentation in image sequences with moving camera. We can achieve thus an indirect indication of the performance of the proposed method, as the resulting segmentation masks reflect the accuracy of the global motion estimation for comparison.

The proposed VSBasel algorithm is compared with the case of fixed-size  $8 \times 8$  blocks used for GME (FSB algorithm) and VSB (uniform participation of variable-size blocks using the BPT). Figures 2.15 - 2.16 illustrate object segmentation results for subjective evaluation. As it can be observed, the GME accuracy deterioration in the case of VSB compared to the FSB algorithm, is reflected in the degradation of the segmentation performance observed in the visual examples. This is especially evident in the background region, where many small groups of pixels are falsely marked as foreground. As it can be seen, the VSBasel algorithm outperforms VSB and FSB, by producing more accurate segmentation results, where false detections are eliminated.

Table 2.3: Segmentation approach [45] used for preliminary evaluation.

Objective	
Foreground and background mask creation.	
Algorithm	
(i)	Anisotropic lowpass filtering of the global motion compensated frame $D_n$ using diffusion.
(ii)	Intensity rescaling in $[0, 1]$ and image binarization using threshold $\tau = \text{mean}(D) + \tau_c \cdot (\text{max}(D) - \mu)$ , where $\tau_c$ is a tuning constant and $\mu$ is the mean intensity of the rescaled $D$ .
(iii)	Small objects removal and hole filling using morphological operators.

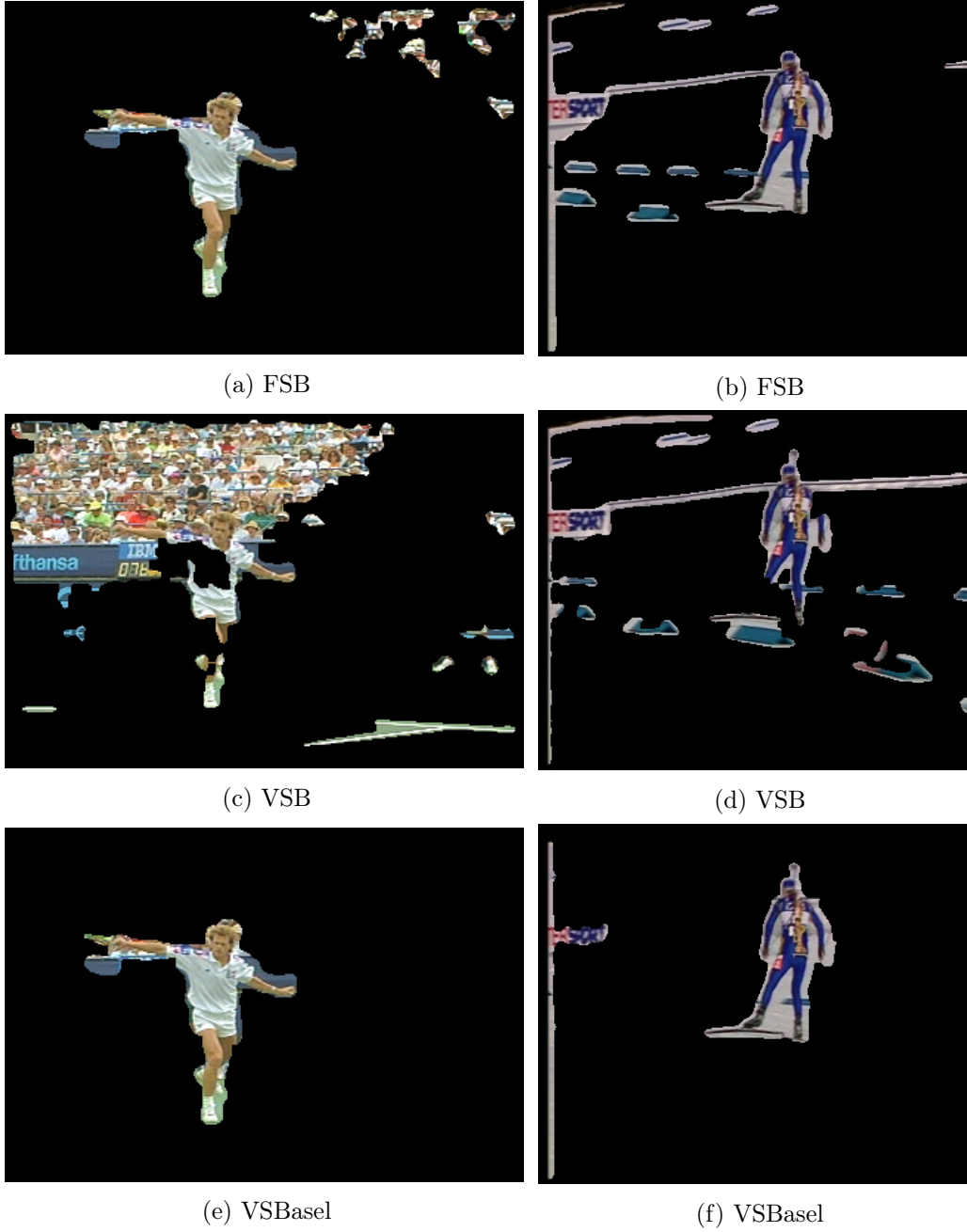


Figure 2.15: Segmentation results based (a)-(b) on  $8 \times 8$  block size (FSB), (c)-(d) on uniform participation of BPT blocks (VSB) (e)-(f) on proposed VSBasel algorithm for the *Stefan* and *Biathlon* test sequences, frames 27 and 87 respectively.



(a) FSB



(b) VSB



(c) VSBasel

Figure 2.16: Segmentation results based (a)-(b) on  $8 \times 8$  block size (FSB), (c)-(d) on uniform participation of BPT blocks (VSB) (e)-(f) on proposed VSBasel algorithm for the *Mountain* test sequence, frame 15.

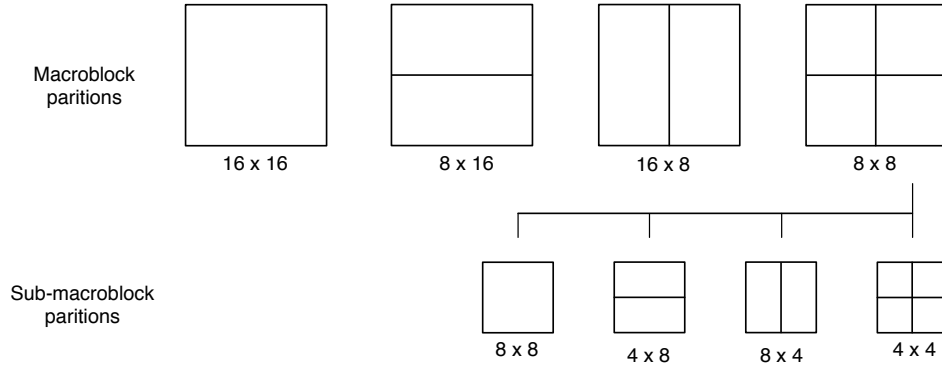


Figure 2.17: Possible partition of macroblocks for motion compensation in H.264/AVC. Top: partition of macroblocks, bottom: partition of  $8 \times 8$  blocks.

## 2.4 Adaptive motion prediction through global motion

Given that the parametric representation of global motion can provide an accurate prediction of the background region of an image, one question that arises is whether this prediction can be more accurate compared to motion vector prediction. In the following, we deal with this question and we further study the use of global motion information for improving conventional motion prediction. Global motion estimation is performed based exclusively on motion information available at the encoder, according to the proposed approach described in the previous section. The coding environment is the state-of-the-art H.264/AVC and several test sequences are used for experimental evaluation.

A brief overview of the motion compensated prediction, related issues regarding H.264/AVC and existing approaches are provided in the following sections. Our examined approach that serves as proof of concept by incorporating global motion prediction through an adaptive prediction mode is subsequently described. Finally, experimental evaluation on several test sequences follows and section 2.6 concludes the chapter.

### 2.4.1 Introduction and existing approaches

Image sequences are characterised by strong temporal correlation. Therefore *motion compensated prediction* (MCP), that enables the exploitation of the redundancy between frames, is a key strategy towards efficient video coding. Predicting a region of the current frame from a matching region of a reference one, which is displaced by a displacement motion vector (MV) is known as *motion prediction* (here also denoted as *motion vector prediction*). The search for the best MV, namely motion estimation, is basically performed using block matching algorithms, as already mentioned, where the idea is to locate the best match of a block of the current frame in the reference frame according to a predefined criterion.

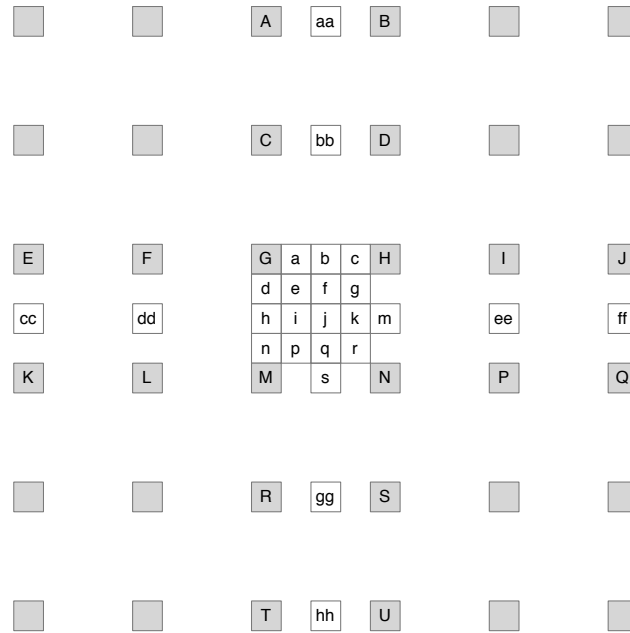


Figure 2.18: Filtering for quarter pixel accurate motion compensation. Upper-case letters indicate values at integer positions and lower-case ones indicate values at half- and quarter pixel positions.

**Motion vectors in H.264/AVC** H.264/AVC, is the coding standard proposed by the ITU-T video coding experts group and the ISO/IEC moving picture experts group (MPEG). H.264/AVC that is well established and is currently considered as the most widely used standard, employs variable-size blocks [21, 46]. It supports a certain flexibility in the selection of block sizes, in contrast to its preceding standards, with  $4 \times 4$  being the minimum block size. The macroblock partitions in H.264/AVC are not arbitrarily chosen, partitions with block sizes of  $16 \times 16$ ,  $16 \times 8$ ,  $8 \times 16$  and  $8 \times 8$  pixels are rather supported [21]. The corresponding  $8 \times 8$  partition is possible to be further partitioned into partitions of  $8 \times 4$ ,  $4 \times 8$  or  $4 \times 4$  pixels. The specific partitioning of macroblocks into motion-compensated blocks of varying size is known as *tree structured motion compensation* [47] and is illustrated in Figure 2.17.

**Motion compensated prediction** Motion vector prediction within H.264/AVC considers spatial and temporal correlations between blocks. The accuracy of motion compensation can be in units of one quarter of the distance between pixels. If the motion vector points to an integer-pixel position, the predicted value derives from the corresponding pixel in the reference frame.

In the case that a motion vector points to a non integer-pixel position, the predicted pixel values are obtained as follows; at half-sample positions a 6-tap FIR filter is employed horizontally and vertically. Predicted values at quarter sample positions are generated by averaging pixel values at integer- and half-pixel positions. Figure

2.18 illustrates symbolically pixels to be interpolated ( $a-k$  and  $n-r$ ). Values at half pixel positions  $b$  and  $h$  are derived by calculating intermediate values  $b_1$  and  $h_1$ , respectively by applying the 6-tap filter as follows [21]:

$$b_1 = (E - 5F + 20G + 20H - 5I + J) \quad (2.24)$$

$$h_1 = (A - 5C + 20G + 20M - 5R + T). \quad (2.25)$$

Half pixel positions values  $j$  are obtained using the intermediate values  $j_1$  as:

$$j_1 = cc - 5dd + 20h_1 + 20m_1 - 5ee + ff \quad (2.26)$$

where  $cc$ ,  $dd$ ,  $ee$ ,  $m_1$  and  $ff$  denote the intermediate values which are obtained similarly to  $h_1$ . Quarter pixel positions values  $a$ ,  $c$ ,  $d$ ,  $n$ ,  $f$ ,  $i$ ,  $k$  and  $q$  are derived by averaging the two nearest pixel values at integer and half pixel positions, whereas the quarter pixel positions values  $e$ ,  $g$ ,  $p$  and  $r$  are derived by averaging the two nearest pixel values at half sample positions in the diagonal direction.

### Motivation and existing approaches

In block matching two key aspects are often considered as shortcomings. The first is the underlying assumption that block correspondences are described by uniform motion and thus each block is represented by a single motion vector. The second is that the translational motion model is adopted to represent block correspondences. Therefore, it fails to capture scaling, rotation and other perspective deformations. To alleviate the second shortcoming, the use of higher order motion models in place of the translational model has been a field of research study over the last years. Nevertheless, standard video codecs do not incorporate sophisticated approaches using such models, mainly due to difficulties regarding the coding of the estimated motion parameter set, beside the extra complications of the estimation itself.

The motivation behind the proposed approach is that global motion estimation enables high accuracy prediction (i.e. *global motion prediction*) employing higher order model transformations. In this way, alterations in a specific area of the frame that have been changed uniformly with respect to the reference frame can be reconstructed using a homography resulting in better prediction of the background region. In case there are objects in a scene whose motion is differentiated and independent of the background one, a frame can not be accurately predicted using a single homography. The idea is thus to exploit global motion prediction in the motion prediction process towards improving it. This is expected to be especially beneficial in the background region and in cases there are no moving objects present.

In [48] global motion using the perspective model was considered on macroblock basis and a decision was made, based on rate-distortion criteria, to determine whether global motion prediction is preferred over conventional motion prediction. Besides, it has been shown that sequences containing significant global motion and no independently moving objects, can be coded exclusively using high order

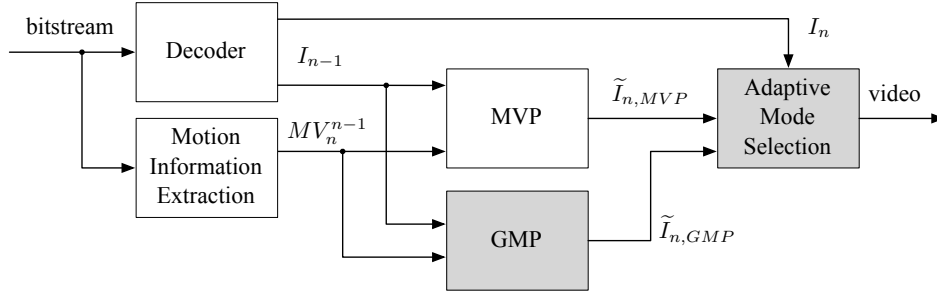


Figure 2.19: Block diagram of the examined scheme using adaptive mode selection.

motion models and the coding efficiency can be increased. In line with this idea Yu *et al.* [49] showed that video sequences that are well described by a single affine transformation per image pair, can benefit from using such a representation towards more efficient coding. The authors in [50] introduced the affine transformation to generate reference pictures in order to handle multiple independently moving objects while avoiding the assignment of affine motion parameters to each image segment, thus avoiding the image partition itself. In [51] affine motion prediction has been used, in a post processing approach within the coding loop, combining the benefits of affine motion prediction and the advantages of the conventional use of motion vectors. In [52] a parametric global motion model between current and reference frames has been used to introduce a new mode, namely the parametric skip mode, that was incorporated into the design of H.264/AVC.

### 2.4.2 Adaptive mode selection

The examined scheme is illustrated in Figure 2.19. For the current frame  $I_n$ , the two predicted versions of it, namely  $\tilde{I}_{n,MVP}$  and  $\tilde{I}_{n,GMP}$ , are compared in the adaptive mode selection (AMS) step and a block wise decision is made for the best prediction mode, based on the lowest mean square error. For the prediction (the predicted versions of  $I_n$ ) the reference frame  $I_{n-1}$  and the corresponding motion vector field  $MV_n^{n-1}$ , extracted from the H.264/AVC encoder, are employed using motion vector prediction (MVP) and global motion prediction (GMP) resulting in  $\tilde{I}_{n,MVP}$  and  $\tilde{I}_{n,GMP}$  respectively. Global motion estimation is performed, according to the proposed approach, described in section 2.2.3.

For a given frame  $I_n$  (in the following the frame index is omitted) and a block in position  $i$ , the lowest mean square error between the two prediction modes is given as:

$$e_i = \min \{ e_{i,MVP}, e_{i,GMP} \} \quad (2.27)$$

where

$$e_{i,MVP} = E[(I_i - \tilde{I}_{i,MVP})^2] \quad (2.28)$$

and

$$e_{i,GMP} = E[(I_i - \tilde{I}_{i,GMP})^2]. \quad (2.29)$$

The block size that the decision is made upon can be the macroblock size up to the smallest block size that H.264/AVC defines.

In cases with high amount of intra predicted macroblocks in a frame, i.e. macroblocks where the spatial redundancy is exploited and that are coded without reference to other frames, the remaining macroblocks that are used for the homography estimation may mislead the calculation and produce falsely calculated homographies. Therefore, in such cases we should expect almost no improvement in prediction using the AMS scheme, something that is also confirmed in the experimental evaluation section that follows.

In cases with significant global motion it is possible to achieve precise prediction of the background region. Moreover, given that the foreground objects are moving independently in relation to the background, it is expected that blocks on the foreground region will be better predicted using MVP in contrast to blocks on the background region that are expected to be better predicted using GMP.

## 2.5 Evaluation of the AMS scheme

### 2.5.1 Test dataset

The test sequences *Allstars*, *Biathlon*, *Birds*, *Foreman*, *Horse*, *Monaco* and *Stefan*, are employed for experimental evaluation of the accuracy of the AMS scheme. These sequences contain a diversity of content characteristics that are described in detail in appendices A.1 and A.2. With regard to the foreground size, *Foreman* and *Horse* contain large foreground objects in relation to the size of the background, whereas *Biathlon* and *Stefan* contain medium size objects, *Allstars* and *Birds* include small objects and finally *Monaco* has no foreground objects. Concerning the relation between local (foreground) and global (camera) motion direction, in *Birds*, *Horse* and *Biathlon* sequences the camera is following almost constantly the foreground object, that remains at the center area of the frame, whereas in *Allstars* and *Stefan* the local and the global motion are not obviously related with each other since the camera does not always keep pace with the foreground object(s).

### 2.5.2 Methodology

The prediction accuracy in terms of average PSNR of each test sequence using the adaptive mode selection (AMS) and the motion vector prediction (MVP) are compared. The adaptive mode selection is examined for the cases that the decision step (i.e. the block size that the decision is made) is  $ds = 4$  and  $ds = 16$ .  $\Delta P_4$  and  $\Delta P_{16}$  denote the PSNR improvement brought by AMS compared to MVP (in dB) for  $ds = 4$  and  $ds = 16$  respectively, and in each case the best results are highlighted with boldface. Table 2.4 reports the results.

We examine cases where the motion vector fields are obtained from encoding the test sequences using H.264/AVC, reference software KTA [53], with varying quanti-



zation parameters (QP), namely  $QP \in \{4, 16, 28, 38, 48\}$ . The following settings are used: IPPP... GOP structure, EPZS motion estimation with  $32 \times 32$  search range,  $4 \times 4$  smallest block size and quarter-pel precision. Figure 2.20 presents two examples of the same sequence encoded with  $QP = 4$  and  $QP = 40$  showing the macroblock assignments. Intra (I) blocks as well as skip (S) ones, which are a special case of the prediction that exploits temporal redundancy (inter prediction), are not used in the described global motion estimation approach. In case there is not sufficient amount of motion vectors due to intra or skip coding, all the available motion vectors are used and no weighting according to the block size is performed. Figures 2.22 and 2.23 show the PSNR curves for MVP and AMS for the examined quantization parameters.

### 2.5.3 Results

As observed in Table 2.4, the most accurate motion vector prediction for each test sequence is achieved for  $QP = 4$ . By increasing the quantisation parameter, a tendency for deterioration of motion vector prediction accuracy is observed. In contrast, this is not happening with global motion prediction. In the case of low quantization parameters, there are many intra coded macroblocks and less motion vectors are available for global motion estimation. This results in low accuracy global motion estimation for very low quantization parameters e.g.  $QP = 4$ . The reason may also rely on the fact that by increasing  $QP$  the deblocking filtering is gradually causing more blurring of (motion) details which results in a greater quantity of larger macroblocks that tend to be indicators of global motion, thus enabling more accurate global motion estimation. This is especially obvious in sequences with homogenous regions such as *Birds*, *Stefan* and *Biathlon*. For larger quantization parameters there is a higher amount of available motion vectors enabling more accurate global motion estimation. However, for very large quantisation parameters (i.e. beyond  $QP = 38$ ) the global motion estimation deteriorates as the amount of skipped macroblocks may decrease and the amount of available and reliable motion vectors may gradually decrease.

The AMS introduces always improvement in the prediction accuracy. For the majority of the sequences, the highest prediction improvement takes place for  $QP = 38$  and for  $QP = 28$ . Figures 2.24 and 2.25 present an overview of the improvement using the AMS compared to MVP in terms of PSNR. The adaptive mode selection is set in such a way that a block is decided to be predicted using global motion estimation compensation if the global motion prediction is better or equally good in comparison with the conventional motion compensated prediction.

Figure 2.21 depicts examples showing the selected prediction mode. The blocks where global motion prediction is preferred are marked with white. As expected, GMP is selected mainly in the background regions of the images, while in the foreground regions the conventional MVP is preferred. According to Table 2.4 the highest improvement is reported in the *Horse* sequence.

Regarding the block size  $ds$ , that the AMS decision is made, we observe that

$ds = 4$  introduces higher improvement on average compared to the case where  $ds = 16$ . This is explained mainly by the flexibility associated with smaller size decisions which can be adjusted in more details to the frame's content.

To conclude, we have shown that the adaptive mode selection introduces certain improvements compared to motion vector prediction in terms of prediction accuracy. This is observed mainly on the background region and higher precision for the mode selection (smaller  $ds$ ) enables more accurate prediction.

## 2.6 Chapter summary

We have presented an improved robust global motion estimation approach that takes into account the variable-size motion vector field. Typically, smaller blocks are assigned by motion estimation to regions with high presence of edges, while larger blocks are assigned to homogenous regions. Based on this observation we have exploited aspects of the block assignment, specifically the block size, towards improving global motion estimation. By studying the case of the binary partition tree, we show improvements in the performance of global motion estimation in terms of accurate background prediction by making appropriate selection and weighting the influence of the participating motion vectors in global motion estimation. Improvement is also shown in comparison to the case of fixed-size blocks. Preliminary object segmentation results reflect also the benefits of the proposed approach.

We have also reported possibilities for improving conventional motion prediction using a parametric global motion model. The parametric representation of global motion using high order motion models (here the perspective model) that describe a combination of several motions (e.g. scaling, shear, projective transformation) than merely the translational one, can be beneficial for motion prediction. The work and particularly approaches for estimation of parametric global models will serve as basis for the next chapters in this thesis.

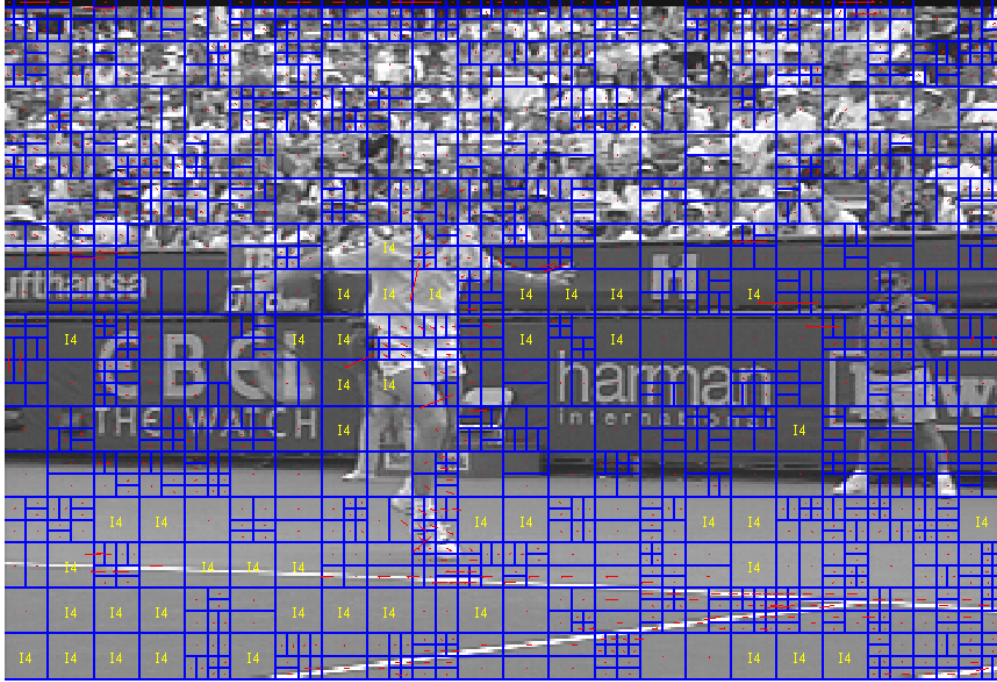
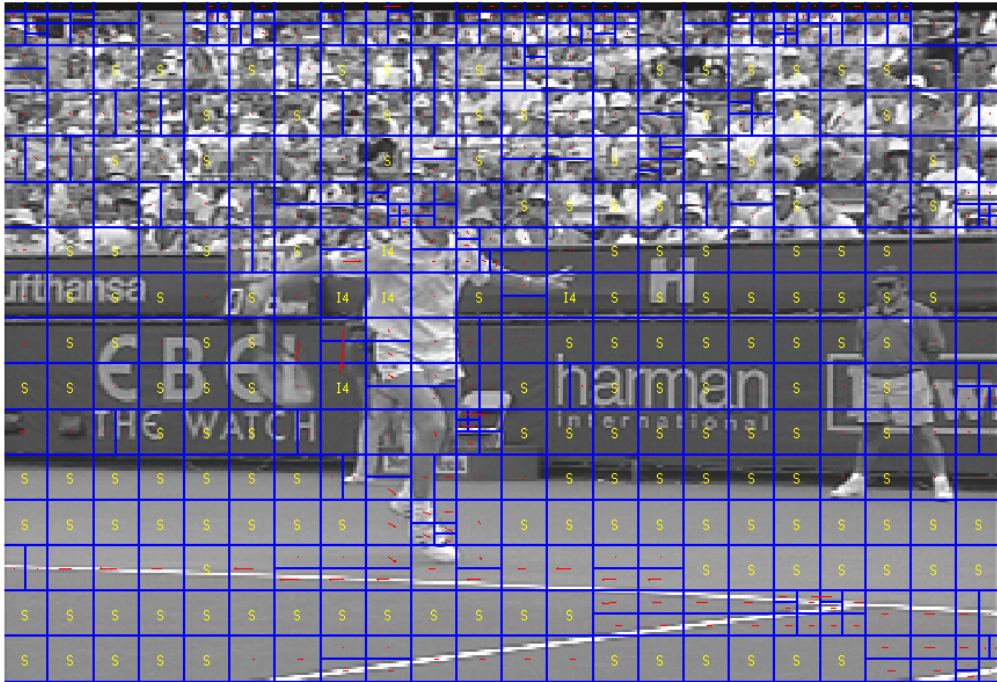
(a)  $QP = 4$ (b)  $QP = 40$ 

Figure 2.20: Block partitioning of frame 23 of *Stefan* sequence using H.264/AVC, reference software KTA [53], for (a)  $QP = 4$  and for (b)  $QP = 40$ . *S* stands for *skip* macroblocks, *I4* for *intra* macroblocks, while *inter* coded macroblocks are depicted with the corresponding overlaying motion vector (in red).

Table 2.4: PSNR (in dB) for motion vector prediction (MVP), global motion prediction (GMP), adaptive mode selection (AMS) and the corresponding improvements for  $bs = 16$  and  $bs = 4$ :  $\Delta P_{16} = PSNR_{AMS_{16}} - PSNR_{MVP}$  and  $\Delta P_4 = PSNR_{AMS_4} - PSNR_{MVP}$ .

Sequence	QP	MVP	GMP	$AMS_{16}$	$\Delta P_{16}$	$AMS_4$	$\Delta P_4$
<i>Allstars</i>	4	38.082	25.476	38.119	0.037	38.23	0.145
	16	38.081	25.475	38.119	0.038	38.23	0.146
	28	33.115	24.914	33.258	0.143	33.39	<b>0.280</b>
	38	29.102	24.652	29.280	<b>0.179</b>	29.38	<b>0.280</b>
	48	25.976	22.615	25.981	0.005	25.99	0.011
<i>Biathlon</i>	4	35.392	18.150	35.393	0.000	35.533	0.140
	16	35.709	18.770	35.720	0.011	35.886	0.176
	28	31.515	18.841	31.617	<b>0.102</b>	31.913	<b>0.398</b>
	38	28.205	18.948	28.294	0.089	28.506	0.301
	48	24.893	18.555	24.933	0.041	25.003	0.110
<i>Birds</i>	4	40.611	22.535	40.611	0.001	40.74	0.134
	16	38.746	23.494	38.834	0.088	39.02	0.271
	28	29.769	23.738	29.949	<b>0.180</b>	30.30	0.534
	38	28.169	23.527	28.341	0.172	28.83	<b>0.664</b>
	48	26.973	24.491	27.020	0.047	27.04	0.072
<i>Foreman</i>	4	37.502	20.902	37.505	0.003	37.59	0.089
	16	36.861	21.183	36.991	0.130	37.11	0.251
	28	31.325	20.992	31.632	<b>0.307</b>	31.82	<b>0.496</b>
	38	27.235	20.754	27.446	0.211	27.60	0.361
	48	23.691	19.543	23.781	0.089	23.85	0.160
<i>Horse</i>	4	29.596	14.583	29.597	0.001	29.83	0.231
	16	29.533	14.581	29.538	0.004	29.77	0.236
	28	25.647	14.697	26.317	0.670	26.79	1.140
	38	22.363	15.092	23.256	<b>0.893</b>	23.71	<b>1.347</b>
	48	19.924	15.937	20.572	0.649	20.86	0.941
<i>Monaco</i>	4	37.469	23.268	37.474	0.005	37.53	0.064
	16	37.469	23.612	37.469	0.000	37.50	0.026
	28	30.290	24.361	30.367	0.077	30.41	0.117
	38	26.045	23.926	26.162	<b>0.117</b>	26.21	<b>0.162</b>
	48	22.489	20.299	22.495	0.006	22.50	0.010
<i>Stefan</i>	4	28.199	15.606	28.204	0.005	28.389	0.190
	16	28.496	15.970	28.508	0.012	28.709	0.213
	28	26.066	16.074	26.233	0.167	26.507	0.442
	38	21.665	16.283	21.998	<b>0.333</b>	22.337	<b>0.672</b>
	48	18.426	16.671	18.551	0.124	18.706	0.279

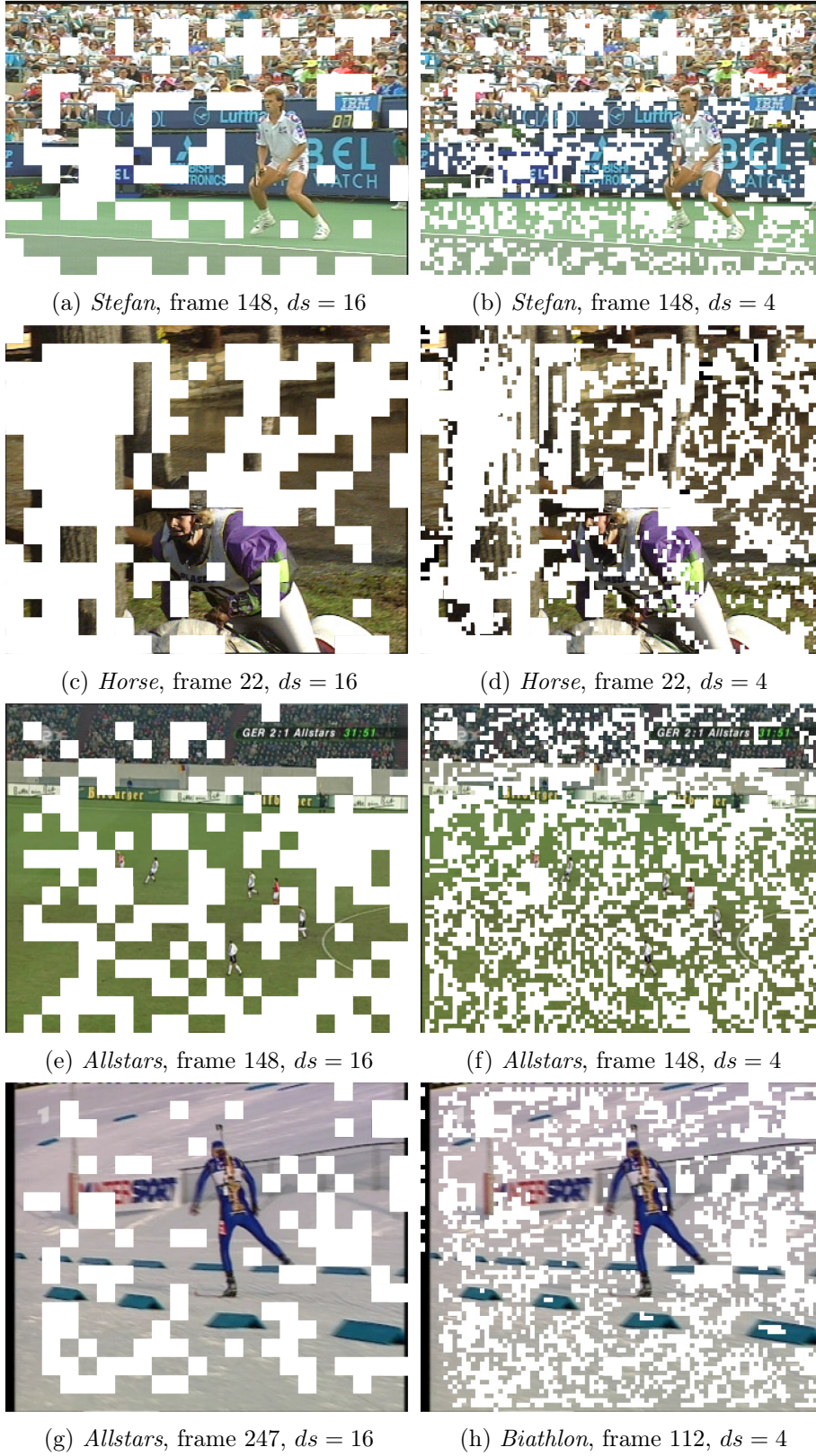


Figure 2.21: Example frames showing block mode allocation in the case of  $QP = 38$  for several test sequences. White corresponds to blocks where global motion prediction (GMP) is preferred, whereas uncovered blocks (image content) are predicted using motion vector prediction (MVP).

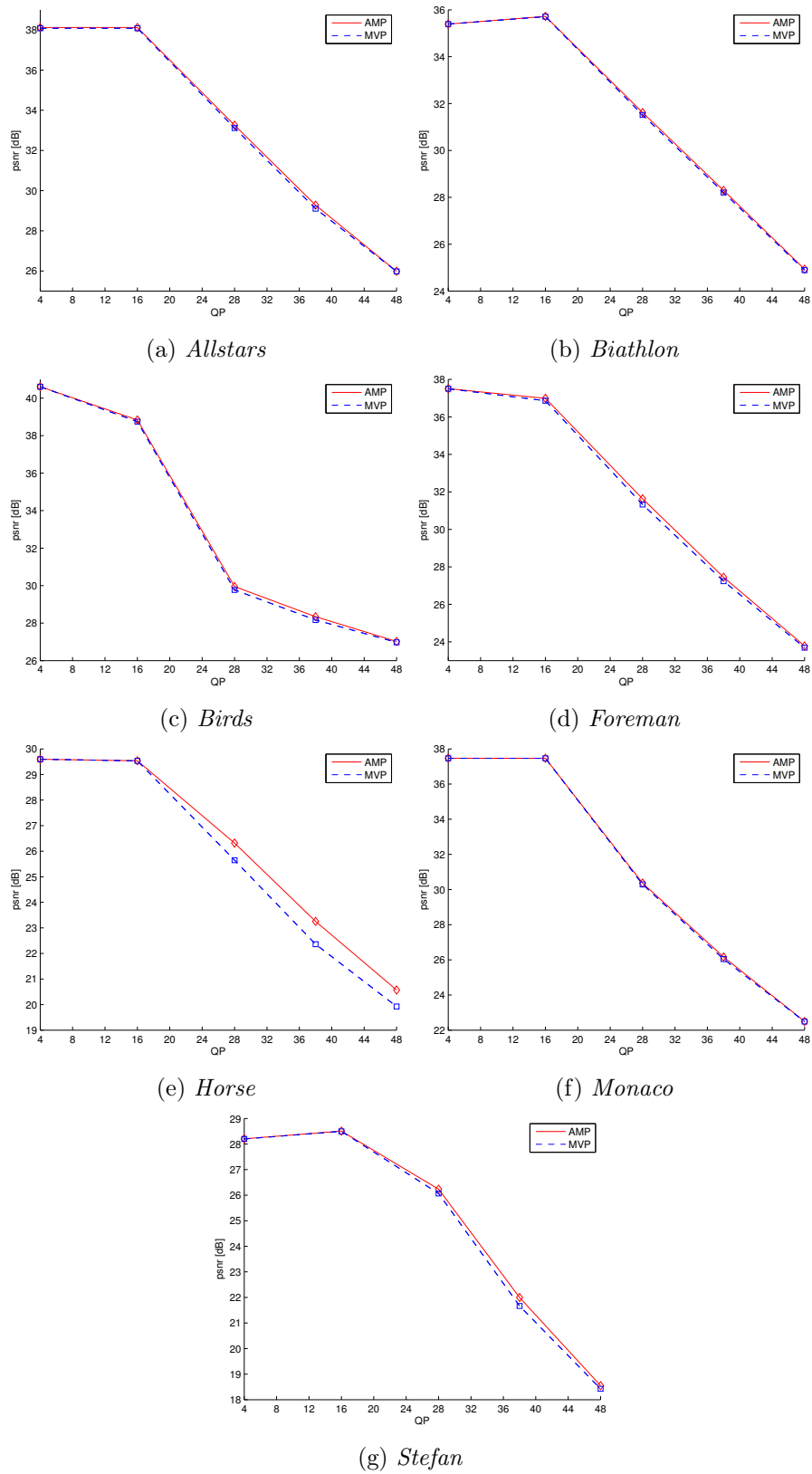


Figure 2.22: PSNR using adaptive mode selection (AMS) with  $ds = 16$  compared to motion vector prediction (MVP) for each test sequence, for varying QP.

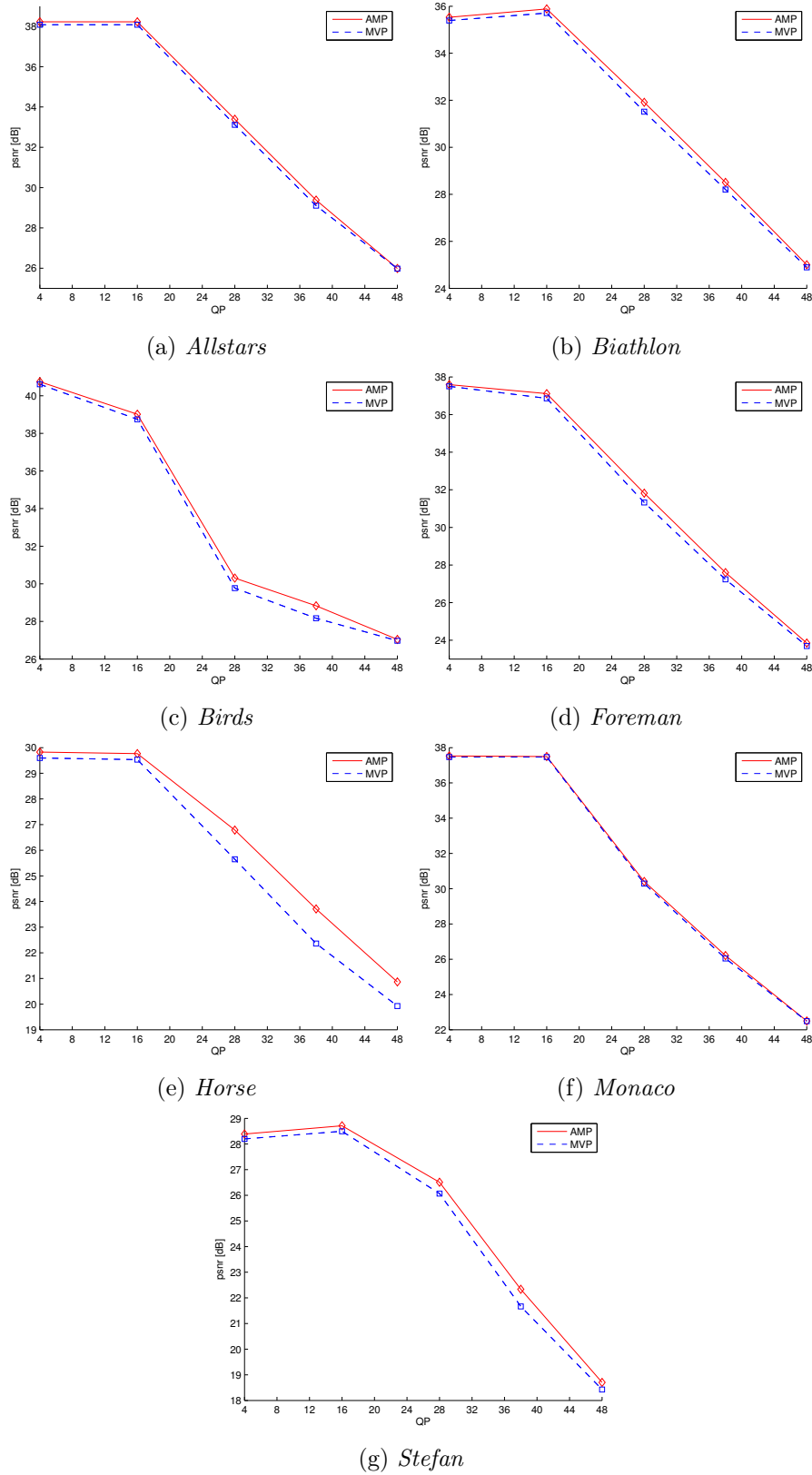


Figure 2.23: PSNR using adaptive mode selection (AMS) with  $ds = 4$  compared to motion vector prediction (MVP) for each test sequence, for varying QP.



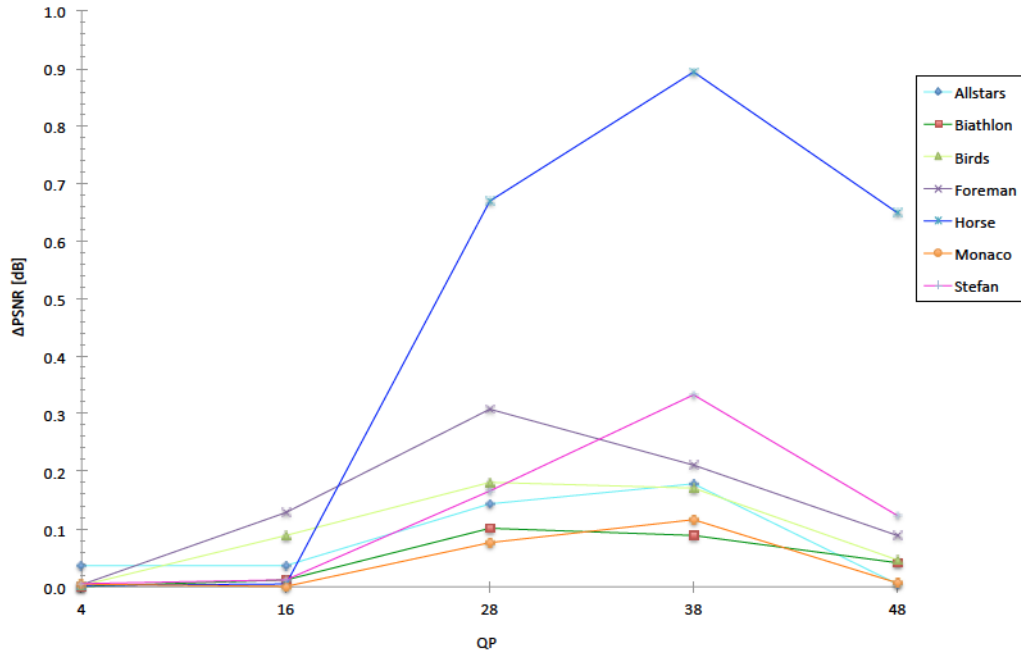


Figure 2.24: PSNR improvement using adaptive mode selection (AMS) over motion vector prediction (MVP), using  $ds = 16$  for all test sequences for varying  $QP$ .

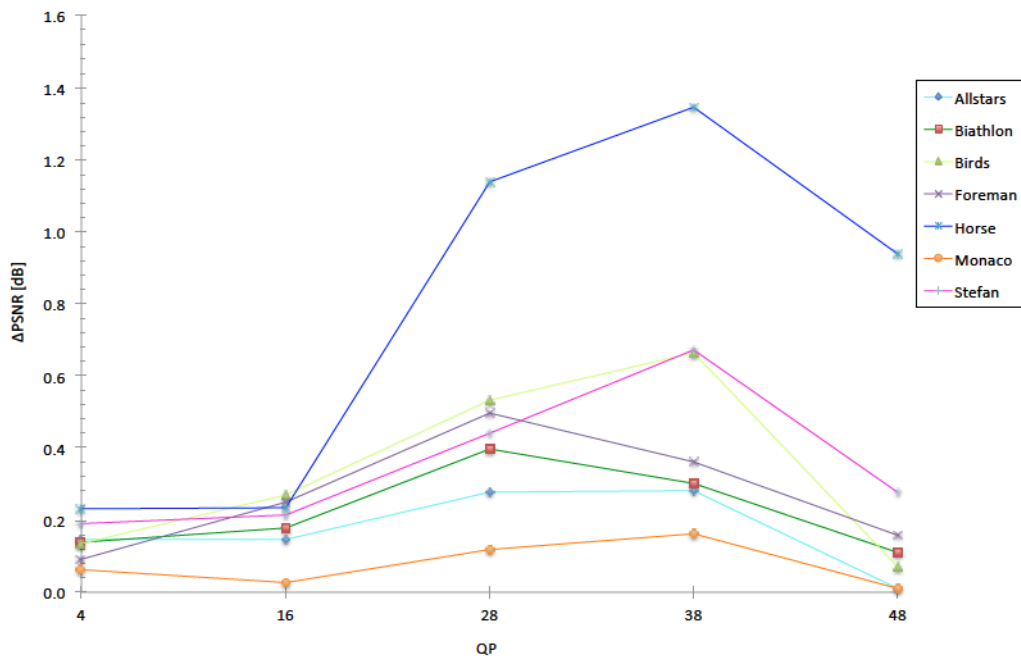


Figure 2.25: PSNR improvement using adaptive mode selection (AMS) over motion vector prediction (MVP), using  $ds = 4$  for all test sequences for varying  $QP$ .



# Moving Object Segmentation

---

## Contents

<b>3.1</b>	<b>Introduction</b>	<b>48</b>
3.1.1	Problem statement	48
3.1.2	Related work	48
3.1.3	Overview of the proposed approach	51
<b>3.2</b>	<b>Bidirectional error frame generation</b>	<b>53</b>
3.2.1	Global motion compensation	53
3.2.2	Error fusion	55
<b>3.3</b>	<b>Thresholding of error maps using hysteresis</b>	<b>56</b>
3.3.1	Adaptive anisotropic diffusion filtering	56
3.3.2	Weighted mean thresholding using spatial connectivity	59
3.3.3	Weight selection	61
<b>3.4</b>	<b>Background classification consistency</b>	<b>65</b>
3.4.1	Sources of errors	65
3.4.2	Temporal consistency	65
<b>3.5</b>	<b>Experimental evaluation</b>	<b>66</b>
3.5.1	Test dataset	66
3.5.2	Evaluation methodology	67
3.5.3	Results	70
3.5.4	Application on H.264/AVC compressed video data	85
<b>3.6</b>	<b>Chapter Summary</b>	<b>87</b>

---

In this chapter a motion-based object segmentation algorithm for video sequences captured by a moving camera, employing bidirectional inter-frame change detection is presented. A three-frame approach is adopted using a simple and effective error fusion scheme and spatial error localization is considered in the thresholding step. We derive appropriate weights for the weighted mean thresholding algorithm that enables robust moving object segmentation. Furthermore, a post-processing step for improving the temporal consistency of the segmentation masks is incorporated and thus we achieve improved performance compared to previously proposed methods. The experimental evaluation and comparison with reference methods demonstrates the validity of the proposed method.

### 3.1 Introduction

Approaches towards object segmentation in video sequences in the literature are based on a wide range of features, such as motion, colour and texture. Motion was recognised quite early as a valuable feature for segmentation and it seems that it was first expressed when the *grouping law of common fate* [54] was formulated, suggesting that motion provides important information for perceptual grouping. Nowadays, studies on the human visual system consider motion among the salient characteristics perceived by humans, and thus it has deservedly attracted much attention in the image processing community for addressing object segmentation tasks.

An effective and straightforward approach for dealing with the object segmentation task [55] is change detection. Given a set of video frames of the same scene, the set of pixels that are significantly different between frames is considered the change detection mask. The change detection mask may be associated with a combination of underlying factors, including appearance or disappearance of objects, motion of objects relative to the background, or shape changes of objects. In this chapter we deal with moving object segmentation in sequences with camera motion, and we focus on change detection and the automatic determination of optimal parameter selection for the involved thresholding step.<sup>1</sup>

#### 3.1.1 Problem statement

As outlined, a strict and generic definition of object segmentation does not really exist, and each class of applications may have its own specific description depending on the requirements. The aim of the work presented in this chapter is to extract the foreground moving objects from a video sequence. We define as *foreground* the groups of pixels that exhibit distinctive motion which is substantially different in relation to the remaining part of the image or to the neighbouring pixels. Foreground is mostly connected with meaningful, in terms of content, objects such as people, animals, objects. In the following we refer to the set of pixels of an image that correspond to the largest part of the image as *background*. Background is mostly considered as the part of an image that is of less importance compared to the foreground for the viewer, for instance the green field in a soccer game, when the players and the ball are considered to belong to the foreground.

#### 3.1.2 Related work

In the literature there is a significant amount of work, concentrating on moving object segmentation. There are numerous classifications of these approaches which vary significantly. For instance, Tekalp's classification of motion segmentation [58] is based on the classification of the employed motion estimation: (a) direct methods (change detection), (b) optical flow segmentation approaches and (c) simultaneous

---

<sup>1</sup>Aspects described in this chapter have been discussed in [56],[57].

estimation and segmentation ones. Zhang et al. [18] classify moving object segmentation approaches in two groups: (a) motion-based and (b) spatio-temporal. Further, among motion-based segmentation techniques, they identify two subgroups: (a) 2D approaches and (b) 3D approaches, based on the dimension of motion models employed in the segmentation. Within the second category, there are structure from motion methods that mostly deal with rigid object motion in 3D scene, and parametric methods which deal with piecewise rigid motion in 2D scenes.

Based on the classification suggested by Zappella et al. [59] we classify existing approaches into several main categories, based on the main principle of the underlying algorithm: (a) statistical framework, (b) optical flow, (c) layer representation (d) factorization and (e) change detection. In the following we provide a short discussion on the most representative works in each of the above categories. It is noted that this classification may not be strict, in the sense that some of the approaches might be classified in more than one category.

**Statistical framework** Approaches in this category rely on statistical procedures, and motion segmentation is seen as a classification problem. Approaches such as the maximum a posteriori probability which are based on the Bayes rule [60, 61] or expectation maximization principles are among the most popular in this category. In [62], graph labelling is employed towards video object plane extraction, corresponding to each moving object. The necessary initial segmentation step is provided through implementation of a watershed algorithm and labelling is modelled as Markov random field (MRF). Chen *et al.* [34] proposed a motion segmentation approach where the priors are obtained through an initial coarse quantisation of the motion vectors into several classes and a maximum a posteriori estimate of the MRF label process, followed by boundary refinement. The algorithms in this category usually require sophisticated priors capable of imposing spatial coherence or demanding knowledge that may not be a priori available. On the other hand their advantage is that they can deal with multiple objects, occlusions and cases where the objects stop moving for a short time (temporal stopping).

**Optical flow** Optical flow is based on derivatives of the image function and provides a dense field of correspondences between images. Under the assumption of continuous flow, segmentation approaches rely on discontinuities in the flow field [63]. Bugeau and Perez [64] combine motion information, spatial continuity and photometric information to deal with the insufficiencies of optical flow. Due to the heavy computational cost, many approaches [65] have studied solutions to overcome this issue.

**Layer representation** Approaches in this category [66] [67] are based on the idea of dividing the image into layers with uniform motion, and further associating them with depth and transparency levels. The depth information enables exploitation of these techniques to stereo vision applications, while the transparency level indicates

the behaviour of the layers in case of overlapping. Initially such a technique was proposed by Wang et al. [68]. Approaches of this category can handle well occlusion problems, due to the depth information that is involved. The main drawbacks are the level of complexity and the dependence on multiple parameters.

**Factorization** Tomasi and Kanade introduced [69] a factorization technique to recover shape and motion from multi view cameras. The key issue in their work was to avoid the computation of depth as an intermediate step, thus triggering research on factorization methods that became attractive due to their simplicity and efficiency [70] [71]. The idea in this category is to factorize the trajectory matrix, that contains the position of the features tracked throughout a number of frames, into two matrices that correspond to motion and structure. The trajectory matrix can be decomposed with approaches such as singular value decomposition, or nonnegative matrix factorization. Approaches of this category can provide the three-dimensional structure of the moving object and the motion of the camera, and are often limited in providing this information for a single object. Moreover, the segmentation is based on features that are assumed to belong to the objects and in this way the object boundaries are hardly detected.

**Change detection** Various approaches are found in the literature dealing with change detection [72] [73] [74]. A typical method is background subtraction, where the involved steps are the calculation of a background model, subtraction of each frame from it and processing of the resulting information. Several background models have been introduced to deal with various related aspects, such as small motion activity [75], dynamic backgrounds and illumination variations [76], vanishing foreground objects [77] and related benchmarks have been also published [78]. These approaches usually rely on a training step to learn the reference background model and often take into account temporal relations between frames implicitly.

*Inter-frame change detection* algorithms fall also into this category. They employ the difference between temporal neighbouring video frames to perform object segmentation, whereas no background modelling is involved. Several related algorithms have been proposed that focus on inter-frame change detection employing one adjacent frame. Kim *et al.* [79] derive an edge map from the difference between two successive frames and after removing edge points which belong to the previous frame, the remaining edge map is used to extract the video object plane. The algorithm involves two thresholds, that are set heuristically and also requires manual definition of a background edge map. In the segmentation method proposed in [1] the change detection mask is obtained using the difference between two successive frames and a local thresholding relaxation technique is employed to enforce spatial continuity. In order to increase temporal stability, a buffer is incorporated such that the previous change detection masks participate in the final segmentation decision step. In the case of sequences captured by a moving camera, Qi *et al.* [26] presented a global motion estimation approach that is using one adjacent frame towards video

object segmentation. This global motion estimation approach is employed to perform object segmentation, which is also used iteratively to predict and reject outliers for global motion estimation in the following frames.

Consideration of only one adjacent frame for inter-frame change detection yields partial foreground detection, since only the edges of the corresponding motion direction are detected. The *double change* detection approach - based on three successive frames - has thus been adopted to overcome this issue. Kameda *et al.* [80] proposed to use error frames from both directions. They end up with two binary masks and fuse them using the intersect operation. In [81] a two-stage segmentation approach is adopted in order to perform subsequently motion estimation and segment labelling. Under the assumption that the number of objects is known a clustering approach is employed to classify the motion models, and two neighbouring frames are considered in order to deal with occlusions. Shih *et al.* [82] employ three adjacent frames in a similar manner and additionally perform motion compensation followed by optical flow estimation to address cases with non-stationary background. Huang *et al.* [83] employ three successive frames for change detection in the wavelet domain [84] and obtain the moving object edge map after applying the intersect operation between the edge maps of significant difference pixel of each pair in each direction. Liu *et al.* [85] employ a similar technique to [83] using three successive frames but they use fuzzy C-means clustering instead of frame difference to classify motion features. The change detection masks are obtained in the wavelet domain after applying the intersect operation to the binary masks of each directions.

### 3.1.3 Overview of the proposed approach

In this chapter, we focus on inter-frame change detection algorithms and specifically under the presence of camera motion. We propose a three-frame segmentation approach that employs a bidirectional fusion scheme of the global motion compensated error. We demonstrate that our error fusion scheme outperforms the intersection fusion scheme that has been commonly adopted, as seen in the previous subsection.

At first step, global motion is compensated between temporally adjacent video frames and between their corresponding motion vector fields. The compensated frames are employed for generating global motion compensated error maps and the compensated motion vector fields are employed in the post-processing step for improving temporal consistency. After low-pass filtering of the error maps, hysteresis thresholding follows, exploiting spatial connectivity of global motion compensated errors. In this step, we avoid setting the thresholding parameters heuristically, which is commonly met in the literature. Instead, we study the problem of optimal weight selection for hysteresis thresholding of error images using the previously proposed weighted mean thresholding approach. Furthermore, we propose a novel adaptive scheme for mitigating the negative effect of temporal inconsistencies while avoiding the incorporation of a buffer. In this way, a large number of previous masks is not necessary to be processed in the final segmentation decision step. As shown in the experimental evaluation section, background detection accuracy is increased while

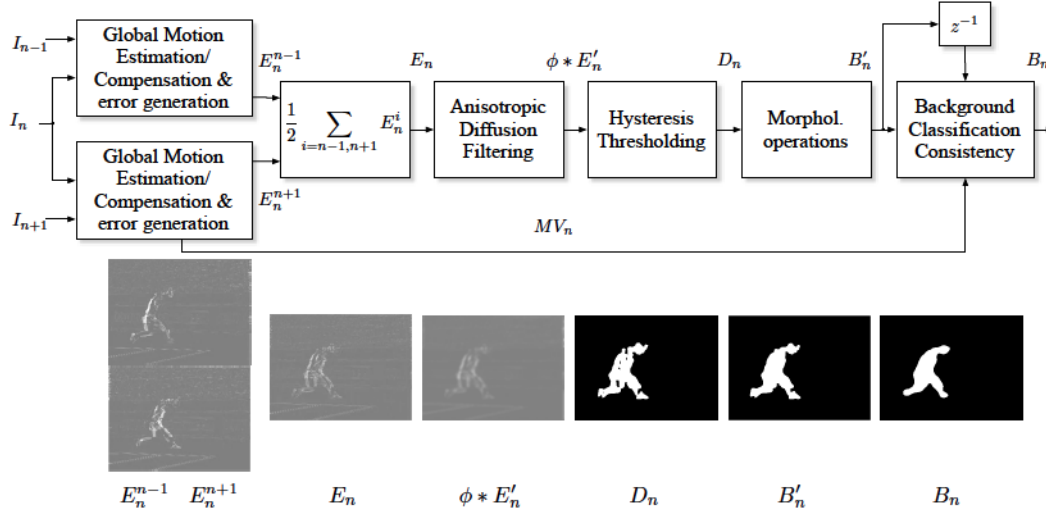


Figure 3.1: Proposed segmentation system overview and examples.

the foreground detection accuracy is maintained high through filtering of the preliminary binary masks, which is adapted according to the motion of the foreground.

The parametric description of the camera motion has been introduced in chapter 2. Ideally, when the camera motion has been compensated the foreground motion can be easily distinguished, but in practice there are many issues, such as the existence of noise, that have to be addressed. To deal with noise and also the existence of foreground objects, we employ a robust regression approach where the key issue is the high breakdown point, i.e. the ability to treat sufficiently cases with up to 80% outlier presence, namely the *Helmholz tradeoff estimator* which is presented in [86].

A description of the proposed algorithm and examples are illustrated in Figure 3.1. For the  $n$ -th frame of a video sequence we employ two adjacent frames, one for each temporal direction. The luminance component contains the most important information for the scope of motion segmentation. Since our approach deals with sequences with moving camera, the parametric model that describes global (i.e. mainly induced by camera) motion between two given video frames in each temporal direction (global motion estimation) has to be estimated first. Subsequently their global motion is compensated (global motion compensation) and eventually the error maps  $E_n^{n-1}$  and  $E_n^{n+1}$  are obtained. Global motion is also compensated between the corresponding motion vector fields and the resulting information is employed in the post-processing step for improving temporal consistency. The resulting error energy maps are fused using averaging, resulting in  $E_n$ . While error locations indicate moving objects' boundaries in real scenes, exploiting directly the error frames for extracting boundary information of moving objects would suffer from great deal of noise even in the ideal case of perfectly compensated global motion. This is due to the fact that random noise created in one frame is different from the one created in successive frames [79], and thus results in slight changes of the error locations (i.e. potentially moving objects) in successive frames. Therefore, the error frame



$E_n$  is filtered, and subsequently a thresholding segmentation scheme, encompassing spatial localization of the error energy, is applied. In the obtained preliminary binary image  $B'_n$  every pixel is labeled as either foreground or background. Finally, the background classification consistency (BCC) step reinforces spatiotemporal consistency, resulting in the final segmentation mask  $B_n$ .

The assumptions under which the proposed algorithm performs well, as well as the strong points and limitations are listed below:

- The approach is designed mainly for the case of moving camera. In the case that the camera stops moving or in the case that new objects appear in the scene, the algorithm performs well as long as there is apparent motion differentiation between foreground and background.
- The camera viewpoint is assumed to be fixed, for a valid representation of background motion by the parametric motion model involved in global motion estimation.
- Multiple foreground objects are detected. There is no limitation in the number of objects in the scene that can be detected. Nevertheless, when objects are very close to each other they tend to be classified as one combined object.
- Regarding foreground object size, due to morphological processing the objects are not detected if they are smaller than approximately 10% of the image frame. Additionally, if the object is larger than approximately 80% of the frame the global motion estimation reflects inaccurately real camera motion and thus segmentation performance decreases dramatically.
- There is no background modeling involved and the algorithm does not need any training stage for parameter setting.
- No a priori information is assumed on the shape and texture of objects. In cases of low textured, low coloured sequences the algorithm works well as long as there is apparent motion differentiation between foreground and background.

## 3.2 Bidirectional error frame generation

### 3.2.1 Global motion compensation

The employed global motion estimation algorithm has been presented by Tok *et al.* in [86] and is briefly described here. It is based on the Helmholtz principle and is overviewed in Figure 3.2. The algorithm derives background motion models from a set of local translational motion models such as motion vectors of encoded video streams. An example for such motion vector fields is shown in Figures 2.20(a) and 2.20(b). Misestimated motion vectors and ones belonging to foreground objects are removed by applying the Helmholtz tradeoff estimator (HTE) that can estimate parametric models from motion vector sets that have up to  $\varepsilon = 80\%$  of outliers. In this section, frame indices are omitted for brevity.

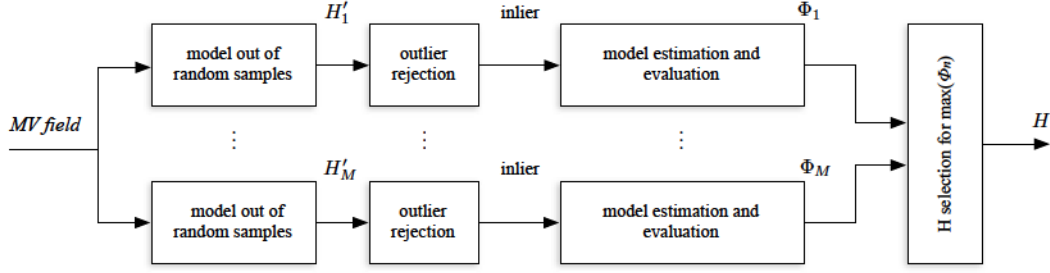


Figure 3.2: Global motion estimation algorithm using the Helmholtz tradeoff estimator and two motion models.

As described in section 2.1.1, the goal of the global motion estimation algorithm is to find the parametric motion model that is described by a  $3 \times 3$  transformation matrix  $\mathbf{H}$  that transforms a given position  $\mathbf{p} = (x, y, 1)^T$  to a new position  $\mathbf{p}' = (x', y', 1)^T$  by  $\mathbf{p}' = \mathbf{H} \cdot \mathbf{p}$ . Rewriting equations (2.4) and (2.5) in homogenous coordinates yields:

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} m_0 & m_1 & m_2 \\ m_3 & m_4 & m_5 \\ m_6 & m_7 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}. \quad (3.1)$$

For a pair of video frames, the HTE global motion estimation approach generates preliminary motion models  $\mathbf{H}'$  from randomly selected motion vectors and evaluates the fit of such a model to the set of all  $K$  vectors. This step is repeated  $M$  times. In each iteration step  $\nu \in \{1, \dots, M\}$ , two randomly, based on uniform distribution, selected vectors are taken from the motion vector field to derive a preliminary four parameter model:

$$\mathbf{H}'_{\nu} = \begin{pmatrix} m'_{0,\nu} & m'_{1,\nu} & m'_{2,\nu} \\ -m'_{1,\nu} & m'_{0,\nu} & m'_{3,\nu} \\ 0 & 0 & 1 \end{pmatrix} \quad (3.2)$$

to roughly describe the translational, rotational and zoom deformation (Figure 2.2) between two frames induced by camera motion. For each vector of the whole set a fitting error related to the model  $\mathbf{H}'_{\nu}$  is calculated. Following [14], the  $(1 - \varepsilon)^{th}$  percentile  $\lambda_{\nu}$  is then taken to estimate an error standard deviation:

$$\sigma_{\nu} = 1.4826 \cdot \left( 1 + \frac{5}{K - p} \right) \cdot \lambda_{\nu}, \quad (3.3)$$

where  $p$  is the amount of observations (motion vector components,  $\mathbf{MV}_X$  and  $\mathbf{MV}_Y$ ) needed to describe the model  $\mathbf{H}'_{\nu}$  in equation (3.2).

A new subset  $\Theta_{\nu}$  of all vectors that fit the motion defined by  $\mathbf{H}'_{\nu}$  with an error smaller than  $\frac{5}{2}\sigma_{\nu}$  is defined. This subset is rated by its standard deviation  $\sigma_{\Theta,\nu}$  and size  $I_{\Theta,\nu}$ :

$$\Phi_{\nu} = \frac{I_{\Theta,\nu}}{\sigma_{\Theta,\nu}}. \quad (3.4)$$



Finally the subset  $\Theta_\nu$  with the highest rating  $\Phi_\nu$  is taken to derive a perspective eight parameter model as in equation (2.1) using least squares regression. This model, as already discussed, can describe more complex deformations between two video frames, such as translation, rotation, zoom and perspective deformation.

The probability  $P$  for selecting two vectors to derive a preliminary model  $H'_\nu$  with  $p = 4$  parameters and an expected outlier percentage of  $\varepsilon$  is:

$$P = 1 - (1 - (1 - \varepsilon)^p)^M. \quad (3.5)$$

Thus, the iteration count  $M$  can be estimated as:

$$M = \frac{\log(1 - P)}{\log(1 - (1 - \varepsilon)^p)}. \quad (3.6)$$

In this work,  $P$  has been set to 99.5% and  $\varepsilon$  has been set to 70% to ensure accurate estimation of the background motion.

The transformed pixel positions of equation (3.1) usually are not integer positions, and therefore their corresponding values have to be interpolated. Since nearest neighbour interpolation wouldn't provide accurate image registration, we apply the more sophisticate and accurate third degree bicubic spline interpolation to obtain values also at sub-pixel locations.

### 3.2.2 Error fusion

For the  $n$ -th frame of the video sequence, let  $\tilde{I}_n^{n-1}$  and  $\tilde{I}_n^{n+1}$  be the estimations of  $I_n$  based on the corresponding eight-parameter global motion models as in equation (2.1) between  $I_{n-1}$  and  $I_{n+1}$  respectively. Based on these, as depicted in Figure 3.1, the global motion compensated error frames for the two temporal directions are given by:

$$E_n^{n-1} = |I_n - \tilde{I}_n^{n-1}| \quad (3.7)$$

and

$$E_n^{n+1} = |I_n - \tilde{I}_n^{n+1}|. \quad (3.8)$$

As discussed in section 3.1.2, many inter-frame change detection algorithms in the literature focus on motion information of one temporal direction i.e.  $I_{n-1}$  or  $I_{n+1}$ . In this way, only edges of one motion direction are included in the foreground region. To overcome this issue, Kameda *et al.* proposed to use error frames from the preceding and succeeding frames. In [80], they perform thresholding on the global motion compensated errors in each direction  $E_n^{n-1}$  and  $E_n^{n+1}$  separately and then obtain a "double-difference image" by a logical intersect operation between the resulting binary masks  $B_n^{n-1}$  and  $B_n^{n+1}$ . This concept is also adopted by [82] and [83]. The intersect operation ensures that foreground misclassifications are drastically reduced (resulting in high accuracy of background detection, as shown in section 3.5) in the obtained  $B_n$  mask, but at the same time a significant amount of foreground regions are misclassified (resulting in low accuracy of foreground detection).

This shortcoming affects the overall segmentation quality in a bad manner as we show experimentally in section 3.5. In this work, we overcome this issue by including information from both directions in an accumulative manner, and thus avoid the intersect operation. The global motion compensated errors in each direction are combined as:

$$E_n = \frac{E_n^{n-1} + E_n^{n+1}}{2} \quad (3.9)$$

and the thresholding segmentation algorithm is then applied on  $E_n$ . By fusing the information of these two error frames, a more complete foreground detection is achieved, which should be reflected in higher *recall* rates in the evaluation. This is due to "approaching" each frame *bidirectionally* as illustrated in Figure 3.3. Additionally, accurate global motion estimation enables the elimination of high error energy in the background region and consequently high *precision* rates are achieved. Precision and recall metrics are discussed in section 3.5.

It is noted here that the incorporation of the chroma components in the global motion error fusion step, as illustrated in the example in Figure 3.4, does not bring substantial improvement, and thus only the luminance component that contains the most meaningful motion information is taken into account.

### 3.3 Thresholding of error maps using hysteresis

The advantages of segmentation algorithms based on inter-frame change detection are that they are straightforward to implement and enable automatic detection of new appearing objects. Their drawbacks include noise (small misclassified regions) and irregular object boundaries [79]. Thus, the error maps should be filtered prior to thresholding and morphological operations such as opening and closing might be incorporated after thresholding to alleviate noise.

#### 3.3.1 Adaptive anisotropic diffusion filtering

In the ideal case where the motion of the camera has been compensated perfectly, the global motion compensated error images contain noise that comes from various sources such as lighting changes, and small movements in the static background. One solution to this problem is to use Gaussian filtering to eliminate high frequencies noise. Nevertheless, this solution would introduce a spatially invariant blurring of the global motion compensated image where regions that contain noise would be blurred in the same way as edges that (ideally) contain motion boundaries. For this reason, we have chosen to use anisotropic diffusion which is first proposed in [87] for denoising. Indeed, anisotropic diffusion offers a non-linear and space-variant filtering of the error frame, that while having a low pass character preserves or enhances the edges of the image. In this way it serves the reduction of high frequency noise due to

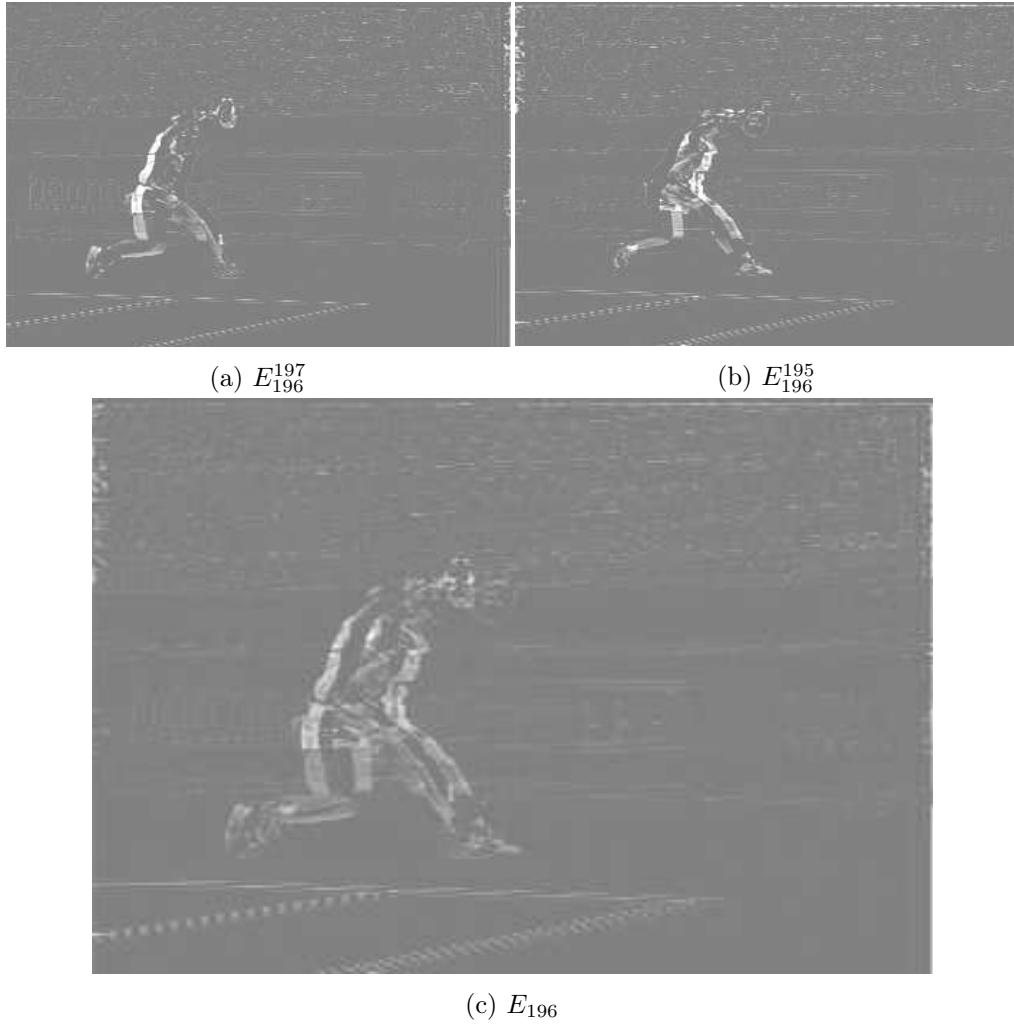


Figure 3.3: Example global motion compensated error frames of the *Stefan* sequence. In (a) and (b) the error energy is located mostly in the left and right side of the foreground object respectively, while in (c) error location indicates better the moving object location.

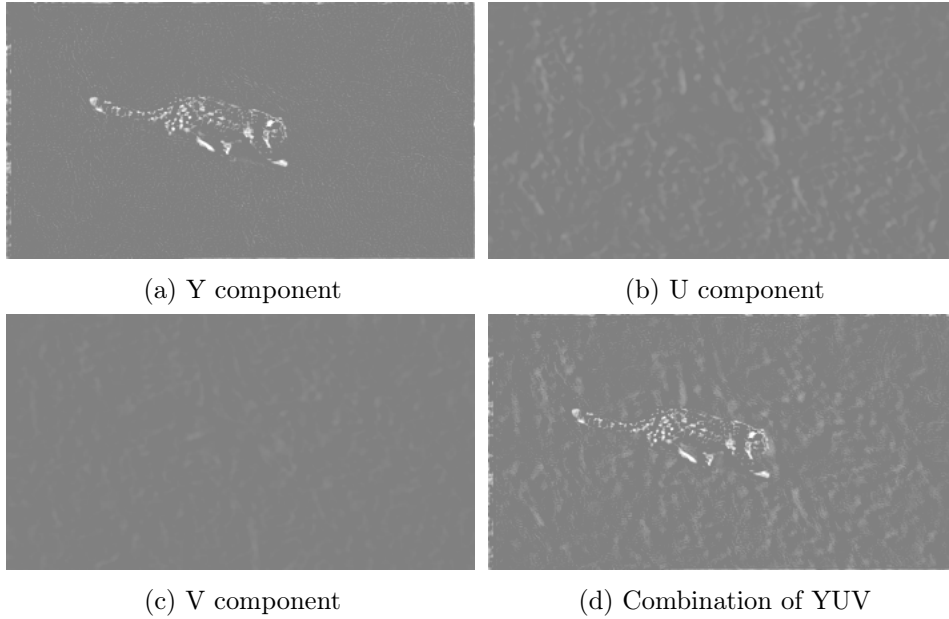


Figure 3.4: Example of global motion compensated error frames for luminance and chrominance components as well as combination of them for the *Mountain sequence*, frame 3.

misestimations in the background while enhancing edges. In the following we give a brief introduction of diffusion.

Diffusion can intuitively be described as the physical process that equilibrates concentration differences without creating or destroying mass. For a signal  $E$ , this process can be expressed as:

$$\frac{\partial}{\partial t} E(x, y, t) = \text{div}(D \cdot \nabla E(x, y, t)) \quad (3.10)$$

where  $D$  is the diffusion function,  $x, y$  are the spatial coordinates and  $t$  denotes time. Diffusion has deservedly attracted the attention of the image processing community since it works as a denoising filter while preserving or even enhancing important image features, and especially edges. A good overview on this topic can be found in [88]. A constant value for the conduction function e.g.  $D = 1$  leads to Gaussian blurring. In the case that the diffusion function  $D$  is spatially constant over the image, then the diffusion is *isotropic* and if  $D$  depends on the location then the diffusion is considered *anisotropic*. Moreover, if the diffusion function  $D$  depends only on the initial image, we deal with *linear* diffusion, otherwise if it depends on the evolving versions of the initial image, the diffusion is considered *nonlinear*.

In case of isotropic diffusion filtering, the diffusion direction is constant, and only its strength can be adjusted. Perona and Malik [87] formulated a nonlinear diffusion approach that reduces the diffusivity at locations that have larger likelihood to be edges for avoiding blurring and localization problems of linear diffusion. They suggested the diffusion function  $D$  to be a function of the magnitude of the gradient

of the brightness function:

$$D = g(\|\nabla E(x, y)\|). \quad (3.11)$$

Specifically,  $g(\cdot)$  shall be a monotonically decreasing function, with  $g(0) = 1$  and  $\|\nabla E(x, y)\| \rightarrow 0 \Rightarrow g(\|\nabla E(x, y)\|) \rightarrow 1$ .

In the case of error images, such as the ones created after global motion compensation, it is important to assign less diffusion (blurring) in regions that the image is changing fast (i.e. edges) and more diffusion otherwise. In this work, being  $E$  in equation (3.10) the global motion compensated error image, and  $t$  the scale-space parameter used to enumerate the scale iterations in the discrete case, we employ the following diffusion function:

$$g(\|\nabla E(x, y)\|) = \frac{1}{1 + \frac{1}{\left(\frac{\|\nabla E(x, y)\|}{\kappa}\right)^2}} \quad (3.12)$$

where  $\kappa$  is a constant that controls diffusion. This specific function is selected because it privileges wide regions over smaller ones. Thus, (3.10) becomes:

$$\begin{aligned} \frac{\partial}{\partial t} E(x, y, t) &= \text{div} (g(\|\nabla E(x, y)\|)) \cdot \nabla E(x, y, t) \\ &= \text{div} \left( \frac{1}{1 + \frac{1}{\left(\frac{\|\nabla E(x, y)\|}{\kappa}\right)^2}} \right) \cdot \nabla E(x, y, t). \end{aligned} \quad (3.13)$$

The selection of  $\kappa$  should be based on the noise level of the error image, which is at this point unknown. Therefore, we use an estimation of it and adapt it according to the statistical distribution of the global motion compensated error image. More specifically, the local contrast  $\kappa$  is set to the 80% value of the integral histogram of the global motion compensated error image.

### 3.3.2 Weighted mean thresholding using spatial connectivity

Subsequently, thresholding is applied to the filtered global motion compensated error frame. As discussed in section 3.1.2 thresholding is a widely used technique for change detection. The weighted mean thresholding approach, proposed in [45], is given by:

$$T(w) = w \cdot \max(E'_n) + (1 - w) \cdot \mu \quad (3.14)$$

where  $w$  is a constant and  $\mu$  is the mean of the normalized filtered error frame  $E'_n$  ( $E_n$  is normalized by its maximum). This thresholding is adapted according to the intensity histogram of every frame, but does not take into account the error localization. In the global motion compensated error frame, e.g. as depicted in Figure 3.3(c), there are significant error values in the foreground area and errors resulting from misestimations in the background area. To eliminate these misestimations, we enhance the weighted mean thresholding approach, as follows.

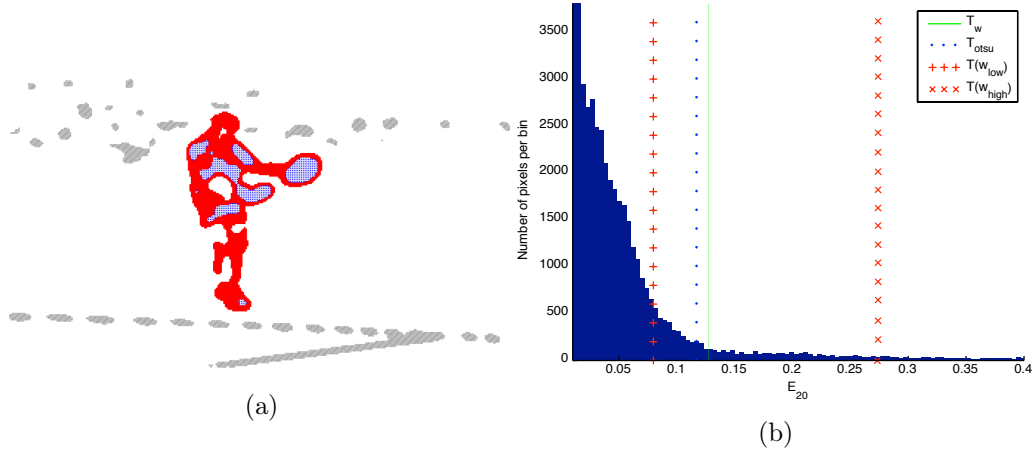


Figure 3.5: Thresholding example for frame 20 of the *Stefan* sequence. (a) Segmentation initial classes using hysteresis thresholding. Pixels with  $E'_{20}(x, y) > T(w_{high})$  are depicted in solid red (class  $F_0$ ). Pixels with  $E'_{20}(x, y) > T(w_{low})$  that are connected with class  $F_0$  are depicted in dotted blue and with dashed grey are the discarded pixels for which  $E'_{20}(x, y) > T(w_{low})$  and are not connected with the ones in class  $F_0$ . (b) The weighted mean ( $T_w$ ), Otsu ( $T_{otsu}$ ) and hysteresis weighted mean thresholds ( $T_{w_{low}}, T_{w_{high}}$ ) are depicted on the intensity histogram of the normalized error.

At first stage, pixels assigned with high error energy are labeled as foreground ( $F_0$  region). Subsequently, pixels with lower error energy, that are spatially connected with  $F_0$ , are favoured against the ones not connected with  $F_0$ , even when the latter have high error energy. Thus, we employ two *hysteresis* thresholds [89]. We begin by applying a low threshold  $T(w_{low})$  using (3.14). This results in high amount of falsely detected foreground pixels, but we can be fairly sure that most regions of the foreground are correctly classified. We then apply a higher threshold  $T(w_{high})$  only on regions that are connected with the binary result from  $T(w_{low})$ . Once this process is complete we have a binary mask where each pixel is marked as either foreground or background. An example is illustrated in Figure 3.5(a).

Eventually, the obtained segmentation mask  $B'_n$  is given by:

$$B'_n = k(D_{n,(w_{low}, w_{high})}) \quad (3.15)$$

where

$$D_{n,(w_{low}, w_{high})} = \theta(\phi * E'_n). \quad (3.16)$$

$E'_n$  is the normalized filtered error frame,  $\phi$  denotes anisotropic diffusion filtering,  $\theta$  weighted mean thresholding using hysteresis and  $k$  morphological filtering. Frame indices are omitted for brevity in the following.

### 3.3.3 Weight selection

One main issue that affects the robustness of the weighted mean algorithm is the appropriate selection of the weight parameter  $w$  involved in (3.14). In [45], it was used  $w = 0.1$  heuristically. By adopting hysteresis thresholding for sake of increasing accuracy, we have one more degree of freedom, due to the fact that we have to search for two optimal thresholding parameters i.e. the corresponding weights  $w_{\text{low}}$  and  $w_{\text{high}}$ .

Finding the optimal generic solution for hysteresis thresholding is considered to be a challenging issue [90, 91], mainly due to the strong dependency of the optimal solution on the input image. A survey on this topic is presented in [92]. The method of Yitzhaky and Peli [91] is to the best of our knowledge the method that selects the optimal pair of hysteresis thresholds from a set of possible values. It is not a parametric approach and it eliminates manual determination to the parameter set selection. The algorithm performs statistical analysis on detection results produced by different parameters, to create an estimated ground truth (EGT) and finds the optimal pair of parameters for edge detection on images. We employ this algorithm to find the suitable weights for weighting mean thresholding using hysteresis on the global motion compensated error maps. As suggested in [91], the obtained optimal parameter set is appropriate for similar images, thus we find the optimal weight set of the first frame of a video sequence, and employ this for the rest of the frames. The range of parameters to be tested in this work is from 0.005 to 0.4 in steps of 0.05, which appears to be reasonable since it covers a wide range of detection results from noisy to sparse. Given a set with  $\nu$  elements and the possible combinations of  $\kappa$  elements are:  $\nu! / (\kappa! \cdot (\nu - \kappa)!)$ . Consequently here, for  $\nu = 8$  and  $\kappa = 2$  there are 28 possible combinations. The procedure is described in the following and is overviewed in Table 3.1. Given a set of  $L$  possible weight combinations:

$$\mathbf{W} = \{W_j = (w_{\text{low}}, w_{\text{high}})_j | w_{\text{low}}, w_{\text{high}} \in [0, 1] \text{ and } w_{\text{low}} < w_{\text{high}}\} \quad (3.17)$$

where  $j = 1, \dots, L$ , use the segmentation masks  $\mathbf{D} = \{D_1, D_2, \dots, D_L\}$  derived using (3.16), that correspond to these combinations, to construct the estimated ground truth: A pixel location which is identified as foreground in all segmentation masks, will be assigned the highest level in the EGT, while a location identified as foreground only in one segmentation mask will be assigned the lowest level. Thus, the EGT is constructed having values within  $[1, L]$ . An EGT example is shown in Figure 3.6. The EGT is then thresholded with each threshold level  $i$  in the set  $I = \{1, \dots, L\}$  forming the potential ground truth ( $PGT_i$ ) for the corresponding level  $i$ . Subsequently, each  $PGT_i$  mask is compared to each  $D_j$  segmentation mask, where  $j = 1, \dots, L$  corresponds to each weight combination  $(w_{\text{low}}, w_{\text{high}}) \in \mathbf{W}$  and generate four probabilities for each individual match:



Figure 3.6: Estimated ground truth the first processed frame of the *Biathlon* sequence,  $L = 28$ .

$$\begin{aligned}
 \overline{TP}_{PGT_i} &= \frac{1}{N} \sum_{j=1}^N TP_{PGT_i, D_j} \\
 \overline{TN}_{PGT_i} &= \frac{1}{N} \sum_{j=1}^N TN_{PGT_i, D_j} \\
 \overline{FP}_{PGT_i} &= \frac{1}{N} \sum_{j=1}^N FP_{PGT_i, D_j} \\
 \overline{FN}_{PGT_i} &= \frac{1}{N} \sum_{j=1}^N FN_{PGT_i, D_j}.
 \end{aligned} \tag{3.18}$$

Now, if each  $PGT_i$  is regarded as ground truth, the above statistical terms are defined as: *true positives* (TP): correctly classified as foreground pixels, *true negatives* (TN): correctly classified as background pixels, *false positives* (FP, also known as *Type I error*): falsely classified as foreground pixels and *false negatives* (FN, or *Type II error*): falsely classified as background pixels.

The best  $PGT_i$  mask is the one that yields the best match according to the *Chi-square test* metric. The Chi-square test of the optimal weight set [91] is:

$$\bar{\chi}_{PGT_i}^2 = \frac{\overline{sn}_{PGT_i} - Q_{PGT_i}}{1 - Q_{PGT_i}} \cdot \frac{\overline{sp}_{PGT_i} - (1 - Q_{PGT_i})}{Q_{PGT_i}} \tag{3.19}$$

where

$$Q_{PGT_i} = \overline{TP}_{PGT_i} + \overline{FP}_{PGT_i} \tag{3.20}$$

$$\overline{sn}_{PGT_i} = \frac{\overline{TP}_{PGT_i}}{P} \tag{3.21}$$



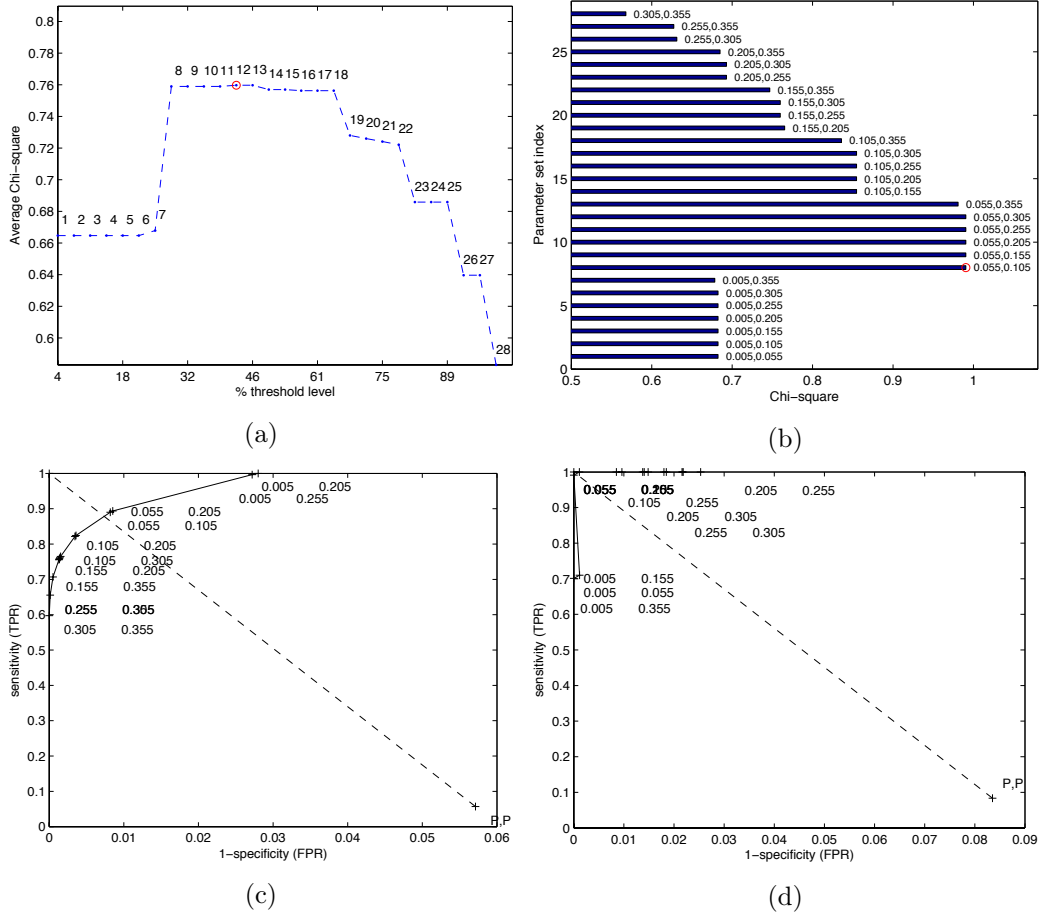


Figure 3.7: Chi-square test for finding the optimal weight pair for weighted mean thresholding for *Biathlon* sequence. 3.7(a) average chi-square ( $\bar{\chi}_{PGT_i}^2$ ) for every threshold level shows a maximum at level  $k = 12$ . 3.7(b) chi-square ( $\chi^2(G_j)$ ) between the different detections and the *EGT* shows a maximum for the weight set (0.055, 0.105).

Table 3.1: Chi-square test for optimal weight selection for hysteresis thresholding.

<u>Objective</u>	
Find the weight pair, from a set of weights $\mathbf{W}$ that perform optimal thresholding of a given greyscale image $I$ .	
<u>Algorithm</u>	
(i)	Threshold image $I$ $L$ times using the set of weights $\mathbf{W} = \{W_j\}$ , $j = 1, \dots, L$ and calculate $D_j$ .
(ii)	Calculate the Estimated Ground Truth, $EGT$ using the masks from (i).
(iii)	Threshold $EGT$ using threshold $i = 1, \dots, L$ and produce the corresponding $PGT_i$ .
(iv)	Compare $PGT_i$ and $D_j$ to find the optimal $PGT_{i=k}$ using $\chi^2$ test and calculate $PGT_k$ .
(v)	Find $\zeta$ for the optimal segmentation mask $D_j$ using $\chi^2$ test and calculate $D\zeta$ .

$$\overline{sp}_{PGT_i} = \frac{\overline{FP}_{PGT_i}}{1 - P}. \quad (3.22)$$

$\overline{sn}_{PGT_i}$  is the *sensitivity* or true positive rate (TPR), and  $\overline{sp}_{PGT_i}$  is the *specificity* which is equivalent to 1-FPR (where FPR is the false positive rate). *prevalence*  $P$  is the average relative number of positive detections. A higher  $\overline{\chi}_{PGT_i}^2$  indicates a better parameter set selection. Figure 3.7 demonstrates an example of the values of the Chi-square measure for different weight levels. The best match between  $PGT_i$  and the EGT is given for  $k = \arg\max_i \overline{\chi}_{PGT_i}^2$ , thus obtaining the optimal potential ground truth  $PGT_k$ . Based on this, the following Chi-square is calculated:

$$\chi^2(D_j) = \frac{sn_{PGT_k, D_j} - Q_{PGT_k, D_j}}{1 - Q_{PGT_k, D_j}}. \quad (3.23)$$

$$\frac{sp_{PGT_k, D_j} - (1 - Q_{PGT_k, D_j})}{Q_{PGT_k, D_j}}$$

where:

$$Q_{PGT_k, D_j} = TP_{PGT_k, D_j} + FP_{PGT_k, D_j} \quad (3.24)$$

$$sn_{PGT_k, D_j} = \frac{TP_{PGT_k, D_j}}{TP_{PGT_k, D_j} + FN_{PGT_k, D_j}} \quad (3.25)$$

$$sp_{PGT_k, D_j} = \frac{FP_{PGT_k, D_j}}{FP_{PGT_k, D_j} + TN_{PGT_k, D_j}} \quad (3.26)$$

and finally the segmentation mask  $D_\zeta$  for  $\zeta = \operatorname{argmax}_j \chi^2(D_j)$  yields the optimal segmentation mask.

### 3.4 Background classification consistency

The obtained segmentation mask ( $B'_n$ ) usually suffers from misclassifications, i.e. falsely classified foreground pixels or falsely classified background pixels, caused by various sources. In this section, we identify the circumstances under which such misclassifications occur and then propose a strategy to address them.

#### 3.4.1 Sources of errors

We consider as misclassifications, the falsely classified pixels. More specifically, falsely classified foreground pixels are, which are called false positives and falsely classified background pixels, called false negatives. In the following we identify the most important cases where such misclassifications occur:

- In cases where the sequence contains background noise (e.g. spectators' movement in sports sequences) high false positives are observed.
- When motion vectors are not describing real motion (e.g. when generated to optimize the rate-distortion trade-off) both types of misclassifications occur, namely false positives and false negatives.
- In cases that the perspective motion model is unable to describe accurately the undergoing camera motion, false positives and false negatives can be caused.
- If the motion of foreground objects (or part of them) matches the dominant motion of the video frame, then their relative velocity (between foreground and background) is almost zero and consequently false negatives are observed.
- Very high foreground velocity occurs, i.e. large displacement between adjacent frames can cause false negatives. This effect is known as *ghosting effect* and characterizes situations where the object seems to appear twice [93]. It is present in cases of inter frame change detection due to the lack of background modeling.

Additionally, one effect that may deteriorate the segmentation result is the temporal coherence of the estimated sequence of segmentation masks. Non-smooth changes between consecutive frames might cause negative side-effects, such as *flickering*.

#### 3.4.2 Temporal consistency

The hysteresis scheme can handle some of the above mentioned error cases to certain extend (as can be seen in figure 3.5(a)), due to the fact that it favours object

boundaries' connectivity. In order to deal with the above described misclassifications and temporal inconsistencies, we propose the following strategy which we name *background classification consistency (BCC)*. First, the obtained preliminary binary masks  $B'_{n-1}$  and  $B'_n$ , are filtered with a two-dimensional isotropic Gaussian lowpass filter with standard deviation that is adapted to every frame according to the average magnitude of the motion vectors of the current frame, which correspond to the foreground region of the previous frame. Next, the (grayscale) mask that is the Hadamard product (pairwise multiplication) of the filtered versions of the preliminary masks is binarized using Otsu thresholding [94] to produce the final segmentation mask  $B_n$ . The multiplication of the filtered preliminary masks serves the elimination of temporal inconsistencies that are observed, when every binary mask is produced independently of its adjacent ones. Error propagation is not an issue here, since  $B'_{n-1}$  and  $B'_n$  are created independently up to this point.

In more detail, filtering serves in creating a spatial attenuation of the object boundaries so that when the filtered masks are combined, and depending on the foreground object's velocity, the new parts of the foreground in  $B'_n$  that do not exist in  $B'_{n-1}$  are maintained. Especially in cases of fast moving objects, filtering helps towards a more complete object detection in the final mask. Filtering is adapted as described in the following:

$\mathbf{H}_n$  is the estimated eight-parameter model for the  $n$ -th frame of the video sequence, as in equation (2.1). The corresponding global motion compensated vector field is calculated as:

$$\mathbf{MV}^{\text{GMC}}(x, y, n) = \mathbf{MV}(x, y, n) - \mathbf{MV}(x, y; \mathbf{H}_n) \quad (3.27)$$

where  $\mathbf{MV}(x, y, n)$  is the motion vector field and  $\mathbf{MV}(x, y; \mathbf{H}_n)$  is the motion vector field that represents the estimated global motion.  $\mathbf{MV}^{\text{GMC}}(x, y, n)$  and  $B'_{n-1}$  are used to calculate an adaptive isotropic Gaussian filter. Let  $\Omega$  be the region that  $B'_{n-1}$  defines and corresponds to  $N$  motion vectors. The preliminary binary mask  $B'_n$  is then convolved with Gaussian filter with kernel size  $(\phi_n \times \phi_n)$ , where  $\phi_n = \lceil 4 \cdot \sigma_n + 1 \rceil$  and

$$\sigma_n = \frac{1}{N} \sum_{i=1}^N \sqrt{(\mathbf{MV}_{Xi}^{\text{GMC}})^2 + (\mathbf{MV}_{Yi}^{\text{GMC}})^2} \quad (3.28)$$

standard deviation.  $i \in \Omega$  and  $\mathbf{MV}_{Xi}^{\text{GMC}}, \mathbf{MV}_{Yi}^{\text{GMC}}$  are the motion vector components for  $X$  and  $Y$  direction respectively at frame  $n$ .

## 3.5 Experimental evaluation

### 3.5.1 Test dataset

The test sequences considered for experimental evaluation are *Allstars*, *Biathlon*, *Mountain*, *Race*, *Stefan*, *BBC fish* and *Horse*. They are characterized by a variety in content, camera motion, number and movement of objects. In order to objectively

evaluate the performance of the proposed algorithm we employ manually created moving objects ground-truth segmentation sequences. The test dataset is detailed described in appendix A.2.

It is noted here to the best of the author's knowledge there exist no publicly available database containing video sequences with strongly moving camera and segmentation ground-truth. The majority of the publicly available existing video segmentation benchmarks, contain video sequences acquired by a camera, with either no or minor background motion such as dynamic changes (i.e. lighting, contrast, shadow alterations) or moving camera attached on moving vehicles, which does not match the underling assumption of fixed camera viewpoint in this work due to the moving viewpoint. Since the proposed work focuses on sequences with moving camera we use a self-created video dataset for the experimental evaluation, as relevant works in the bibliography do. Therefore, we have used several well known sequences to evaluate our algorithm using manually created segmentation ground-truths, which are publicly available.

In order to quantify the spatial-temporal variation of image sequences, ITU has defined the following measures [95]. For a given image sequence, the *spatial perceptual information* (SI) indicates the level of spatial information. For its computation, the edge-detecting *Sobel* filter is applied on each frame  $I_n$  and the standard deviation  $\sigma$  of the filtered version of the frames is computed. The SI indicator is then defined as the maximum standard deviation of all frames:

$$SI = \max \{ \sigma[Sobel(I_n)] \} \quad (3.29)$$

The *temporal perceptual information* (TI) indicates the temporal detail in a sequence and is based on the motion difference feature  $MF_n$  of successive frames in time, that is defined as the difference between pixel values at the same location in space:

$$MF_n(i, j) = I_n(i, j) - I_{n-1}(i, j) \quad (3.30)$$

where  $(i, j)$  are the pixel coordinates of the n-th frame. The standard deviation  $\sigma$  is following computed for each frame difference, and TI is given as the maximum standard deviation over all frames as:

$$TI = \max \{ \sigma[MF_n(i, j)] \} \quad (3.31)$$

Higher values of SI correspond to higher spatial information content, whereas high TI values correspond to more motion in adjacent frames. Figure 3.8 shows the test sequences on the SI-TI plane. As it can be seen they span a large region on the SI-TI plane indicating the variation of spatial and temporal characteristics.

### 3.5.2 Evaluation methodology

To evaluate the efficiency of the segmentation results, the produced segmentation masks are compared to the manually created ground truth masks. To that end the following quantities, already introduced in section 3.3.3, are defined as follows:

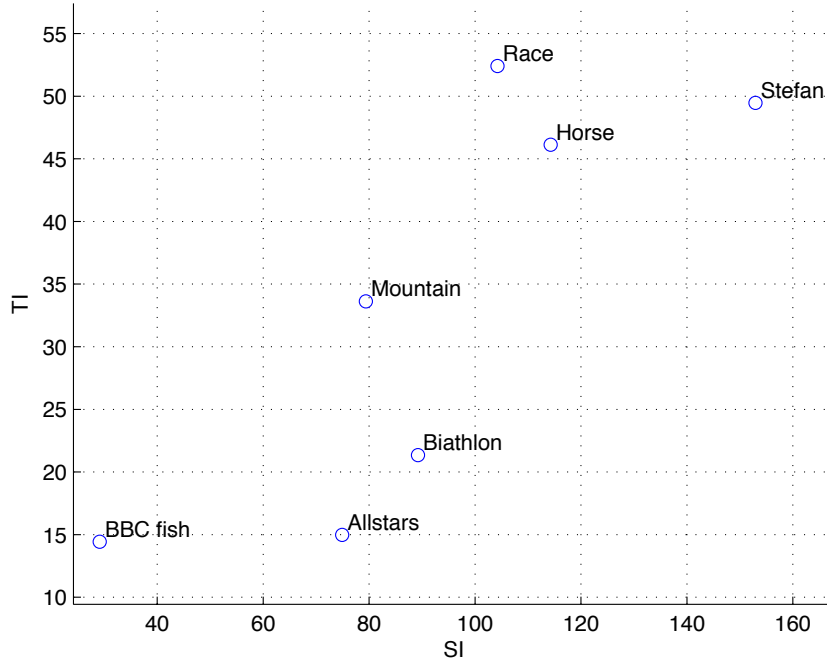


Figure 3.8: Test dataset on the spatial - temporal perceptual information (SI - TI) plane.

- *True positives (TP)*: correctly classified as foreground pixels.
- *True negatives (TN)*: correctly classified as background pixels.
- *False positives (FP)*: falsely classified as foreground pixels.
- *False negatives (FN)*: falsely classified as background pixels.

Based on these, the segmentation efficiency is measured in terms of precision ( $P$ ), recall ( $R$ ) and F-measure ( $F$ ), that are defined as:

$$P = \frac{TP}{TP + FP} \quad (3.32)$$

$$R = \frac{TP}{TP + FN} \quad (3.33)$$

$$F = 2 \left( \frac{P \cdot R}{P + R} \right). \quad (3.34)$$

Precision indicates how exact the segmentation is, meaning how accurately the background is estimated, whereas recall shows how complete the foreground segmentation is. Balancing between these two contradictory quantities, precision and recall,

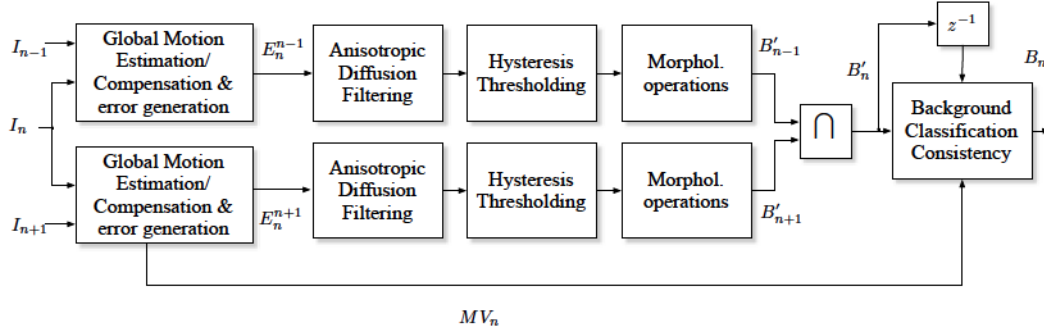


Figure 3.9: Reference algorithm 2 system overview.

comprises the main challenge that algorithms dealing with the task of object segmentation must address. F-measure is the harmonic mean of precision and recall and is widely used as an objective overall indication of the segmentation quality.

Example frames produced using the reference and the proposed approaches are presented in the following for subjective evaluation of the segmentation results. The number of correctly detected objects is also considered as a metric for evaluating the efficiency of each algorithm. Furthermore, computational complexity and runtime are compared.

**Algorithm scenarios** In order to compare the proposed algorithm with other global motion compensated error fusion approaches for object segmentation in sequences with moving camera, we compare the following approaches:

- **Algorithm 1** the approach proposed in [26] which uses one adjacent frame for the detection of object segmentation mask that is also used to predict and reject outliers for global motion estimation. The authors here proposed also a global motion estimation approach where in each loop step outliers are detected according to the segmentation map and are rejected. Code provided by the authors has been used for the comparison.
- **Algorithm 2** the approach proposed in [80] that employs two frames and the intersection fusion scheme as described in section 3.2.2 and overviewed in Figure 3.9.
- **Algorithm 3** the proposed approach that is detailed described in section 3.1.3 and overviewed in Figure 3.1.

The global motion estimation algorithm described in section 3.2.1 is used for the error fusion scheme of Algorithm 2, and the segmentation of global motion error frames as described in sections 3.3.2 - 3.4 are used in each case in order to have a fair comparison of segmentation performance.

### 3.5.3 Results

**Segmentation performance** Algorithm 1 produces segmentation masks that suffer from background misclassifications as well as incomplete foreground detection, especially in one side of the foreground object, due to fact that one motion direction is used for global motion compensation. This results in low precision, but fairly good recall rates. Algorithm 2 presents improved background detection accuracy, since the intersect operation ensures that most of background misclassifications are avoided, but the segmentation masks suffer from incomplete foreground detection, as described in section 3.2.2. This is reflected by high precision but very low recall rates. Algorithm 3 enables complete foreground detection due to the proposed error fusion scheme and at the same time produces 33.1% on average more accurate background detection (in terms of recall) compared Algorithm 2 on the whole test dataset.

Figure 3.12 illustrates examples of the above cases and Table 3.2 shows precision, recall and F-measure for the algorithm scenarios described above. Algorithm 2 performs on average 32.4% better than Algorithm 1 in terms of precision, but suffers from 37.3% lower recall rates. The best performance in terms of precision is achieved by Algorithm 2 and the best one in terms of recall is achieved by Algorithm 1. Nevertheless, precision and recall are two contradictory quantities; often increment of each one of them means decrement of the other one. Thus, by achieving good but not the best precision and recall rates, but still above at least 59%, our proposed algorithm outperforms the reference algorithms and clearly improves the results in terms of F-Measure. Figure 3.10 reports a comparative overview of the percentage of frames in each test sequence that have quality above 75% in terms of F-measure and Figure 3.11 presents a concise overview of the performance in terms of F-measure on the whole dataset, and a comparison between proposed and reference algorithms. Figures 3.19 - 3.21 illustrate examples of the test dataset, as well as F-measure curves, using the reference and the proposed algorithms.

**Background classification consistency** By incorporating BCC, the segmentation masks are temporally more consistent, false positives are eliminated, while foreground object detection is more complete. The *ghosting effect*, which appears when foreground objects are moving fast, is also eliminated due to the filter adaptation according to the foreground object's velocity, and the object boundaries are smoothed over time as can be seen e.g. in Figure 3.12. Nevertheless, in the case of *Allstars*, one of the football players is repeatedly moving and stoping in front of a static object, and he is in some cases falsely regarded to belong to the background together with that static object. This results in a slight (1.6%) degradation in terms of F-measure, as can be seen in Table 3.3 which presents the performance improvement in terms of F-measure by incorporating BCC, compared to the case where no background refinement step is involved, which has been presented in [56].



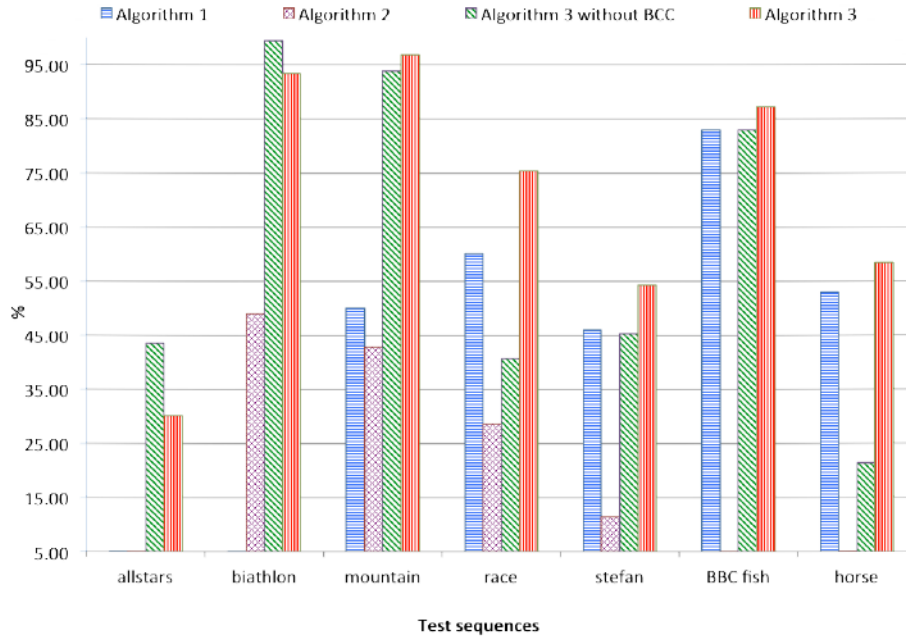
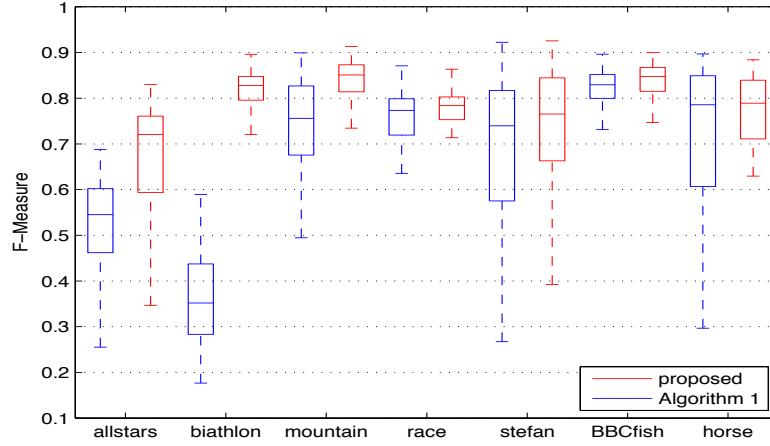


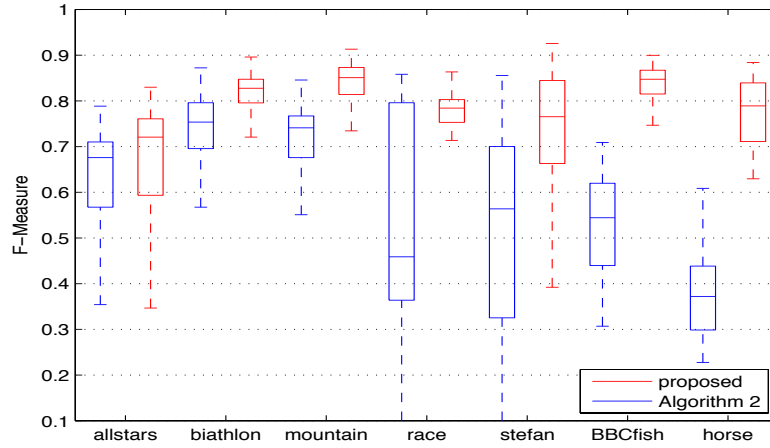
Figure 3.10: Percentage of frames with quality above 75% in terms of F-measure. Comparison of reference and proposed algorithms.

Table 3.2: Test sequences and results of experimental evaluation in terms of average precision (P), recall (R) and F-measure (F) of reference and proposed algorithms. The best precision, recall and F-measure results are shown in bold.

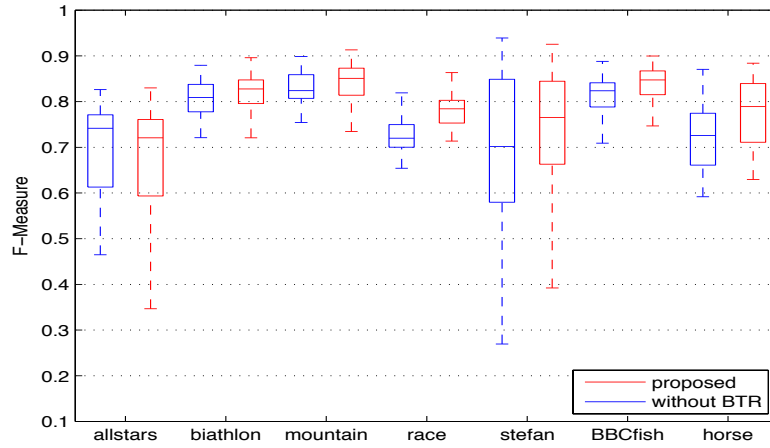
Sequence	Algorithm 1 [26]			Algorithm 2 [80]			Algorithm 3 Proposed		
	P	R	F	P	R	F	P	R	F
<i>Allstars</i>	0.44	<b>0.69</b>	0.52	<b>0.84</b>	0.49	0.61	0.77	0.59	<b>0.65</b>
<i>Biathlon</i>	0.24	<b>0.83</b>	0.36	<b>0.92</b>	0.63	0.74	0.78	0.87	<b>0.82</b>
<i>Mountain</i>	0.60	<b>0.95</b>	0.73	<b>0.93</b>	0.59	0.72	0.84	0.85	<b>0.84</b>
<i>Race</i>	0.69	<b>0.84</b>	0.75	<b>0.89</b>	0.41	0.53	0.74	0.83	<b>0.78</b>
<i>Stefan</i>	0.65	<b>0.80</b>	0.69	<b>0.86</b>	0.41	0.52	0.71	0.79	<b>0.73</b>
<i>BBC fish</i>	0.75	<b>0.87</b>	0.80	<b>0.89</b>	0.38	0.53	0.82	0.83	<b>0.81</b>
<i>Horse</i>	0.65	<b>0.78</b>	0.70	<b>0.96</b>	0.24	0.38	0.88	0.70	<b>0.78</b>
average	57.4%	<b>82.2%</b>	64.9%	<b>89.7%</b>	45.0%	57.3%	79.0%	78.1%	<b>77.4%</b>



(a)



(b)



(c)

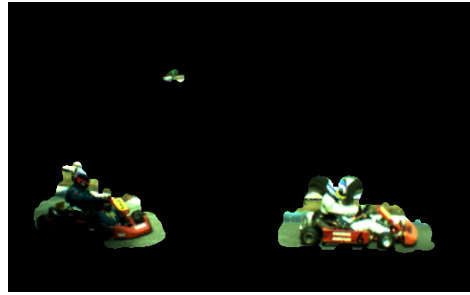
Figure 3.11: Comparison of F-measure distributions on the test dataset between the proposed and reference algorithms. Each box indicates the median (central mark) and the 25<sup>th</sup> and 75<sup>th</sup> percentiles (edges of the box), whereas the whiskers extend to the most extreme values. The extreme values correspond to:  $[Q_2 - 1.57(Q_3 - Q_1)/\sqrt{n}, Q_2 + 1.57(Q_3 - Q_1)/\sqrt{n}]$ , where  $Q_2$  is the median,  $Q_1$  and  $Q_3$  are the 25<sup>th</sup> and 75<sup>th</sup> percentiles respectively, and  $n$  is the number of values.

Table 3.3: Contribution of the background classification consistency to the performance improvement in terms of average precision (P), recall (R) and F-measure (F).

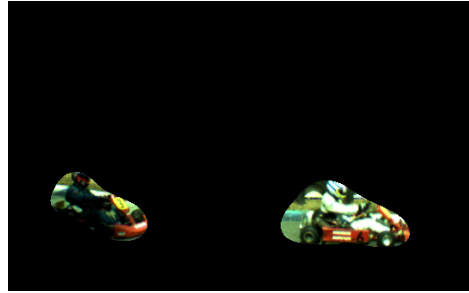
Sequence	Without BCC [56]			With BCC [57]				$\Delta F$
	P	R	F	P	R	F		
<i>Allstars</i>	0.71	0.66	0.67	0.77	0.59	0.65		−1.6%
<i>Biathlon</i>	0.71	0.94	0.80	0.78	0.87	0.82		+1.6%
<i>Mountain</i>	0.77	0.90	0.83	0.84	0.85	0.84		+1.2%
<i>Race</i>	0.63	0.87	0.73	0.74	0.83	0.78		+5.2%
<i>Stefan</i>	0.61	0.83	0.69	0.71	0.79	0.73		+4.2%
<i>BBC fish</i>	0.74	0.89	0.80	0.82	0.83	0.81		+2.1%
<i>Horse</i>	0.81	0.65	0.72	0.88	0.70	0.78		+5.7%



(a) Original frame



(b) Algorithm 3 without BCC



(c) Algorithm 3 with BCC

Figure 3.12: Original frame, segmentation example of proposed algorithm without and with BCC for *Race* sequence, frame 27.

**Thresholding** Table 3.5 presents the evaluation of segmentation results that are produced using three thresholding schemes generated by Algorithm 3 for all the test sequences. The thresholding schemes compared are: i) the well known Otsu thresholding [94], which maximizes the ratio of inter/intra-class variance, ii) the weighted mean thresholding (*WM*) [45] and iii) the hysteresis weighted mean (*HWM*) as described in section 3.3.2 (an example is illustrated in Figure 3.5b). In every case, hysteresis mean thresholding outperforms the other two thresholding algorithms in terms of segmentation efficiency.

**Anisotropic diffusion filtering** Regarding anisotropic diffusion filtering, we examine the contribution of parameter  $\kappa$  on the segmentation performance. This parameter controls the diffusion strength in a way that low  $\kappa$  values allow for small gradients to block diffusion across edges whereas large values reduce the influence of gradients on diffusion. To study the effect of  $\kappa$  we group the error values in ten equally sized intervals, that are replaced by their mean value that is representative for each interval. Figure 3.13 shows the average performance of each test sequence in terms of F-measure, for the range of  $\kappa$  between the 1-st and 10-th interval of the global motion compensated error. In our experiments we have used the value of the 3-rd interval as  $\kappa$ . It is observed that in cases of *Race* and *Stefan* sequences, a smaller  $\kappa$  would be more appropriate. This can be explained by the fact that these sequences have high TI index (Figure 3.8), so larger values do not give the optimum performance. In fact, the selection of  $\kappa$  according to the global motion compensated error is a direct way of taking into consideration into the filtering the temporal information of the foreground, which is adapted according to the motion (and size) of foreground objects.

**Number of correctly detected objects** Additionally, the number of correctly detected objects is considered as a quality measure. As described in Appendix A.2, the test sequences contain foreground objects with various sizes that may move independently. As shown in Figure 3.14, the proposed algorithm detects foreground objects with good accuracy. In the case of sequences with multiple objects presence, at least 79.13% of the foreground objects are detected with the proposed algorithm, 94.44% are detected with Algorithm 1 and 77.83% with Algorithm 2, whereas in sequences with single object presence (*Biathlon*, *Mountain*, *BBC fish* and *Horse*) the object is always correctly detected. As it is observed, Algorithm 1 shows higher detection rates than the proposed algorithm (Algorithm 3), which is also in agreement with the higher recall rates in Table 3.2. However, these high detection rates are followed by high false foreground detection rates, which makes the performance of the proposed algorithm in general better compared to Algorithm 1. In more details, the average number of correctly detected of objects in *Allstars* is: 88.41%, 69.24% and 79.74%, in the case of algorithms 1, 2 and 3 (proposed) respectively, whereas in *Race* these rates are: 81.27%, 73.70% and 78.53% and finally in *Stefan*: 95.17%, 93.50% and 94.67% respectively.

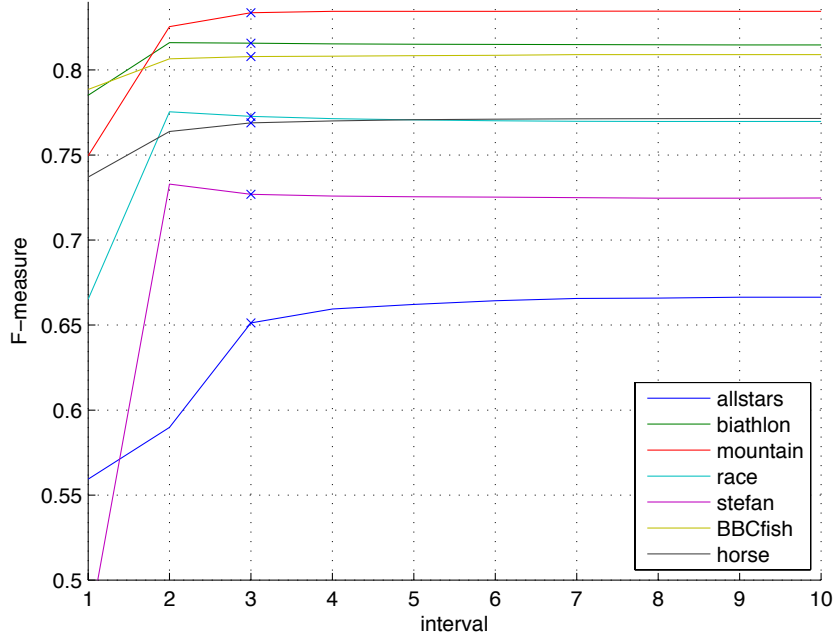


Figure 3.13: Dependence of performance in terms of F-measure on parameter  $\kappa$  in anisotropic diffusion filtering. The selected value was equal to the mean of the 3-rd interval of the global motion compensated error.

**Computational complexity** Regarding computational complexity, each part of the proposed algorithm is examined separately. For a frame with  $m \times n$  pixels, bearing in mind that the number of iterations (motion vector outlier rejection  $M$ , anisotropic diffusion filtering, set of weights  $W$  etc.) is fixed, and the involved parameters (motion model, gaussian kernel, etc.) have fixed size, the computational complexity of each part of the algorithm is presented in Table 3.4.

The  $n \cdot m \log(m \cdot n)$  term in the global motion estimation and compensation term derives from the Helmholtz tradeoff estimator approach, where the fitting errors between all motion vectors and the preliminary motion model are calculated,

Table 3.4: Computational complexity of each step of the proposed algorithm.

Step	Complexity
GME/GMC	$O(n \cdot m \cdot \log(n \cdot m))$
Error generation	$O(n \cdot m)$
Filtering	$O(n \cdot m)$
Weight selection	$O(n \cdot m)$
Thresholding	$O(n \cdot m)$
Morphological processing	$O(n \cdot m)$
BCC	$O(n \cdot m)$

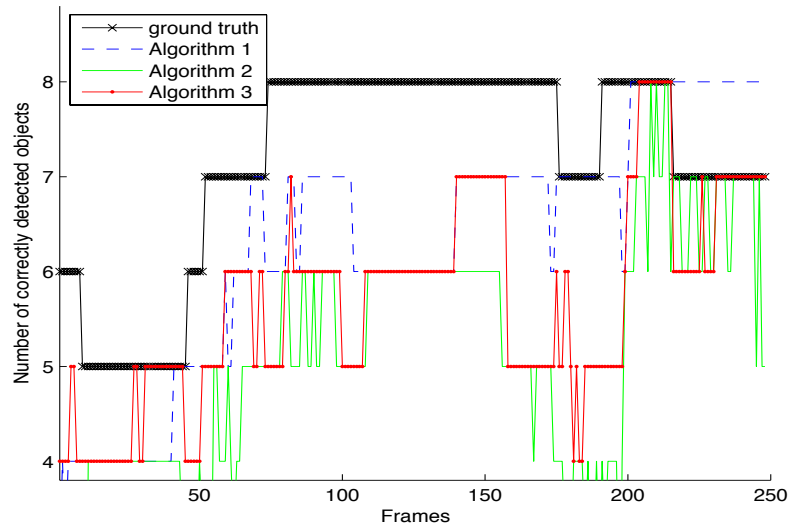
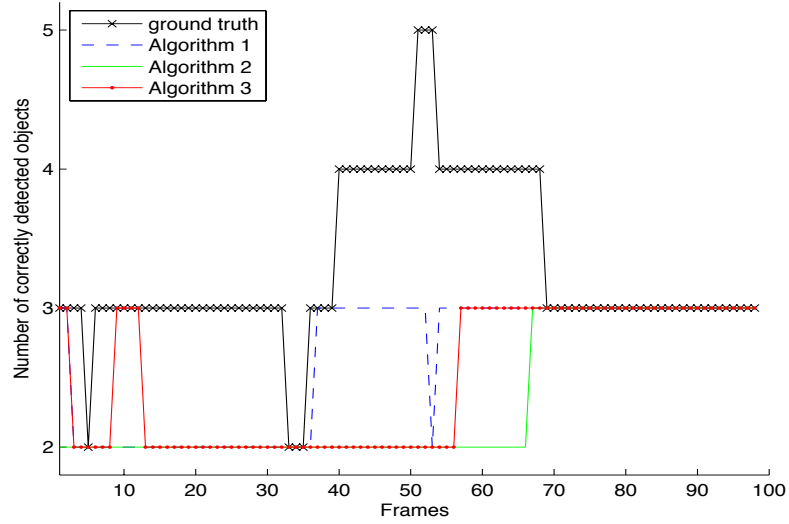
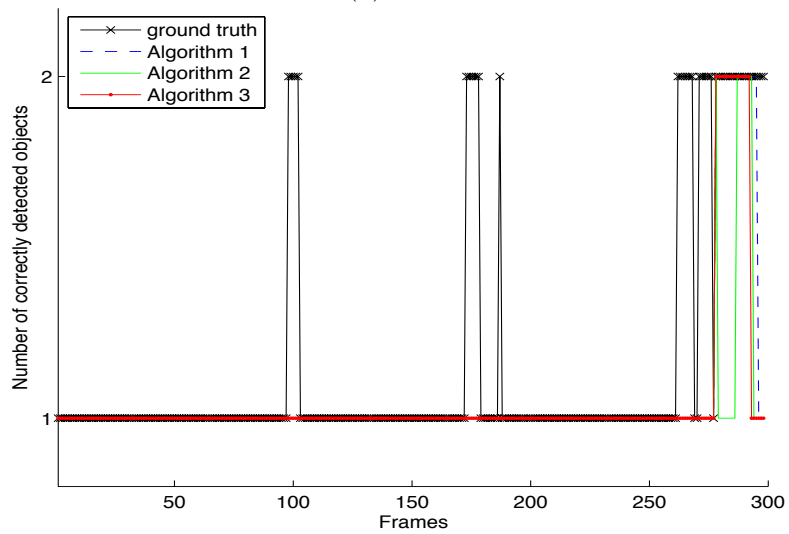
(a) *Allstars*(b) *Race*(c) *Stefan*

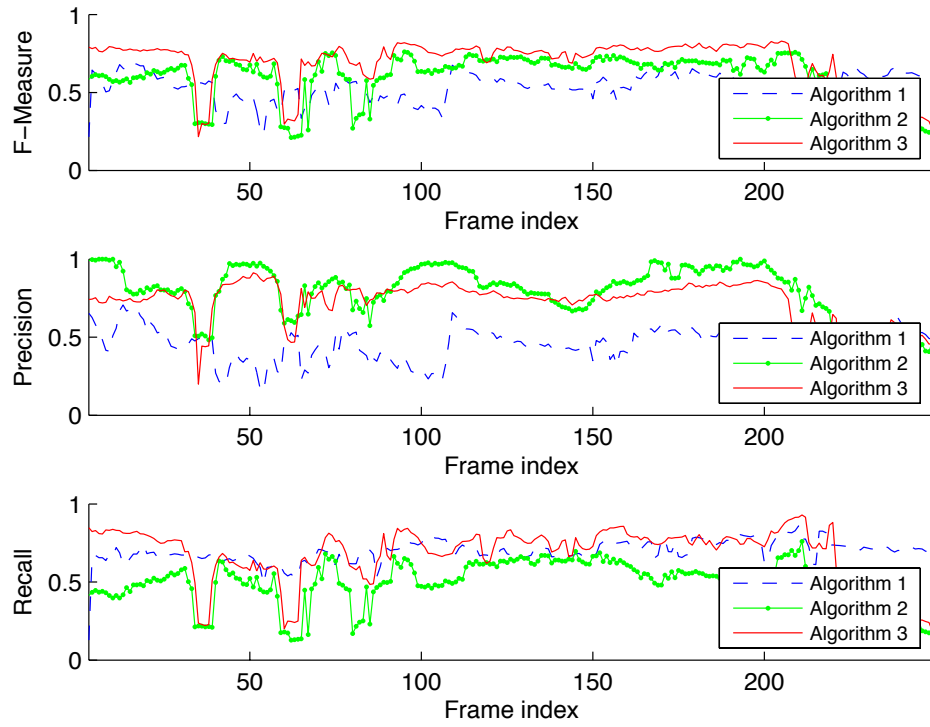
Figure 3.14: Number of foreground objects detected with the proposed algorithm and reference algorithms in sequences with multiple objects.

Table 3.5: Average F-measure for Otsu, weighted mean (wm) and hysteresis weighted mean (hwm) thresholding algorithms.

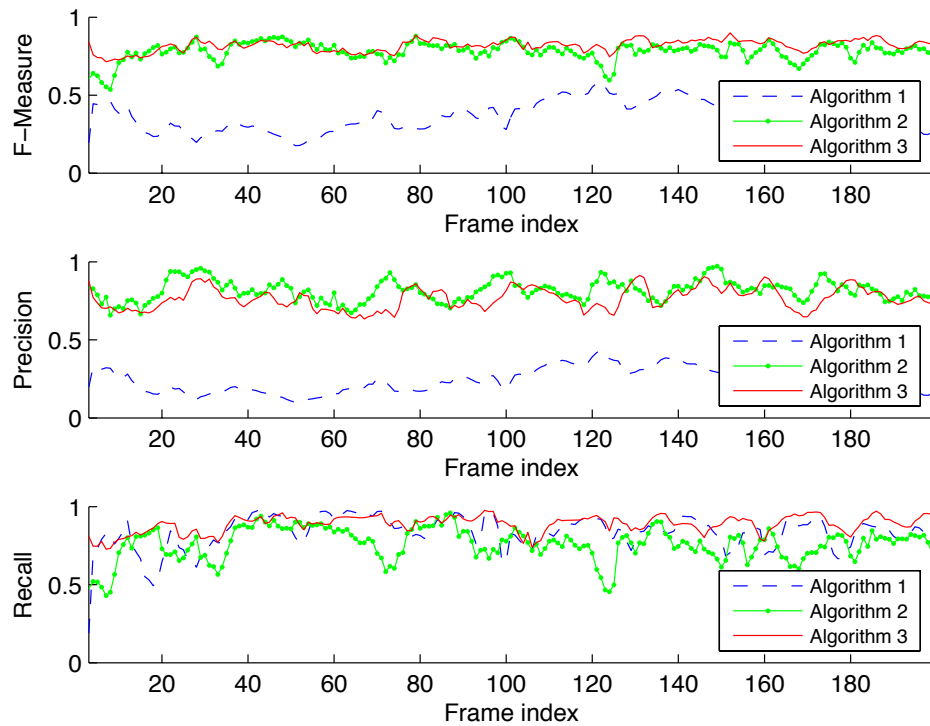
Sequence	OTSU [94]	WM [45]	HWM
<i>Allstars</i>	0.58	0.61	0.65
<i>Biathlon</i>	0.79	0.81	0.82
<i>Mountain</i>	0.82	0.82	0.84
<i>Race</i>	0.75	0.75	0.78
<i>Stefan</i>	0.69	0.71	0.73
<i>BBC fish</i>	0.75	0.76	0.81
<i>Horse</i>	0.67	0.68	0.77

for a maximum of  $\frac{m}{4} \times \frac{n}{4}$  blocks of size  $4 \times 4$  (pixels). After this calculation, the set of errors has to be sorted in order to calculate the percentiles, and this sorting results in this term. Thus, in the worst case scenario, the complexity of the proposed algorithm is  $O(n \cdot m \cdot \log(n \cdot m))$ . Algorithm 2 involves the convergence rate,  $\kappa$ , of the gradient descent [26], which determines its complexity. Assuming that  $\kappa$  is not fixed, the computational complexity of Algorithm 2 is  $O(n \cdot m \cdot \kappa)$ , whereas in case Algorithm 3 the complexity is the same as the proposed one, since the fact that most of the included steps have to be performed twice, does not change  $O$ .

Regarding runtime, the proposed algorithm needs 1.6 sec on average for a frame of a CIF sequence (*Biathlon*) under a 2.2 GHz AMD opteron 8354 with 48 GB RAM. From this time, 0.97 sec are used for GME, 0.11 sec for GMC, 0.38 sec for filtering, 0.05 sec for binarization, 0.01 sec for morphological processing and 0.12 sec for BCC. More concisely, 1.1 sec is needed for GME/C and 0.5 sec for segmentation which is implemented in MatLab. The algorithm of Kameda *et al.* [80] is not faster than the proposed algorithm, since all the steps have to be performed twice, for each direction, before combining the segmentation masks using the intersect operation. The algorithm [26] can save 75% of time in the global motion estimation step, based on the code provided as an executable by the authors. Comparing to the segmentation performance, the proposed algorithm outperforms [80] for 20% and [26] 12% in terms of F-measure. The global motion estimation algorithm that we employ here is based on the Helmholtz Tradeoff Estimator which ensures robustness against noise. This is reflected to the fact that the proposed algorithm outperforms Algorithm 1 in terms of segmentation efficiency and the possible employment of a faster global motion estimation approach would enable real-time application scenarios.



(a) Allstars



(b) Biathlon

Figure 3.15: Precision, recall and F-Measure curves per frame using proposed and reference algorithms for the *Allstars* and *Biathlon* sequences.



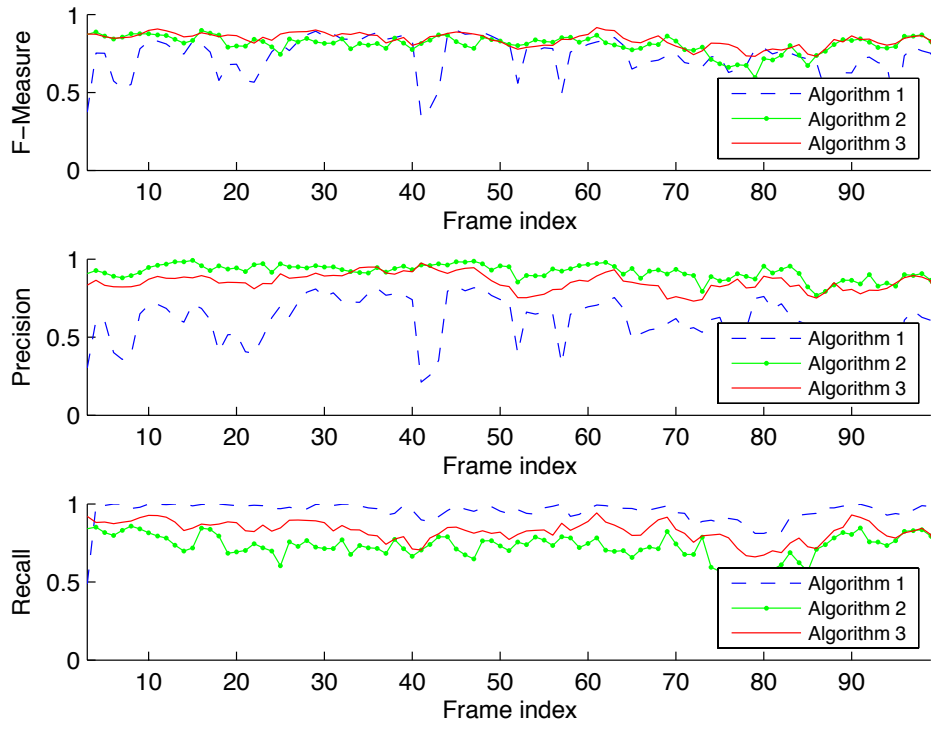
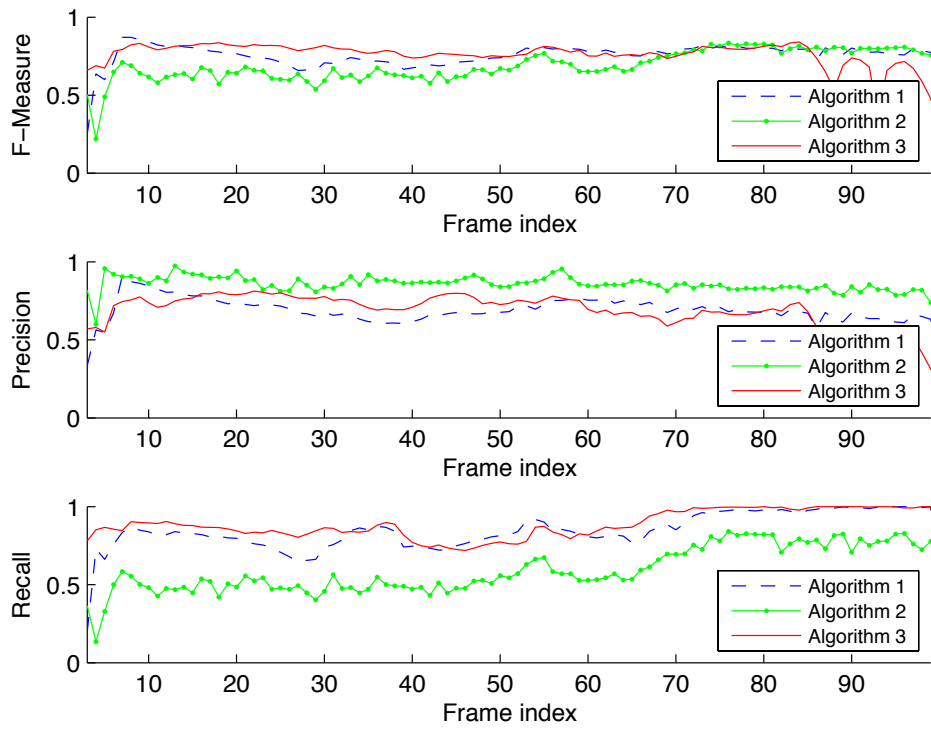
(a) *Mountain*(b) *Race*

Figure 3.16: Precision, recall and F-Measure per frame using proposed and reference algorithms for the *Mountain* and *Race* sequences.

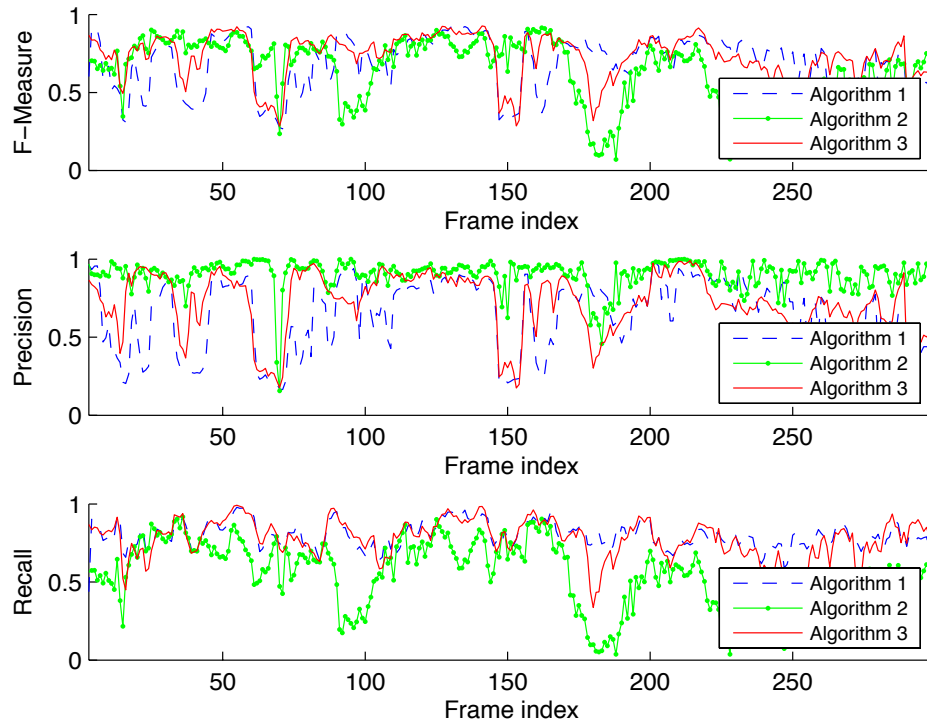
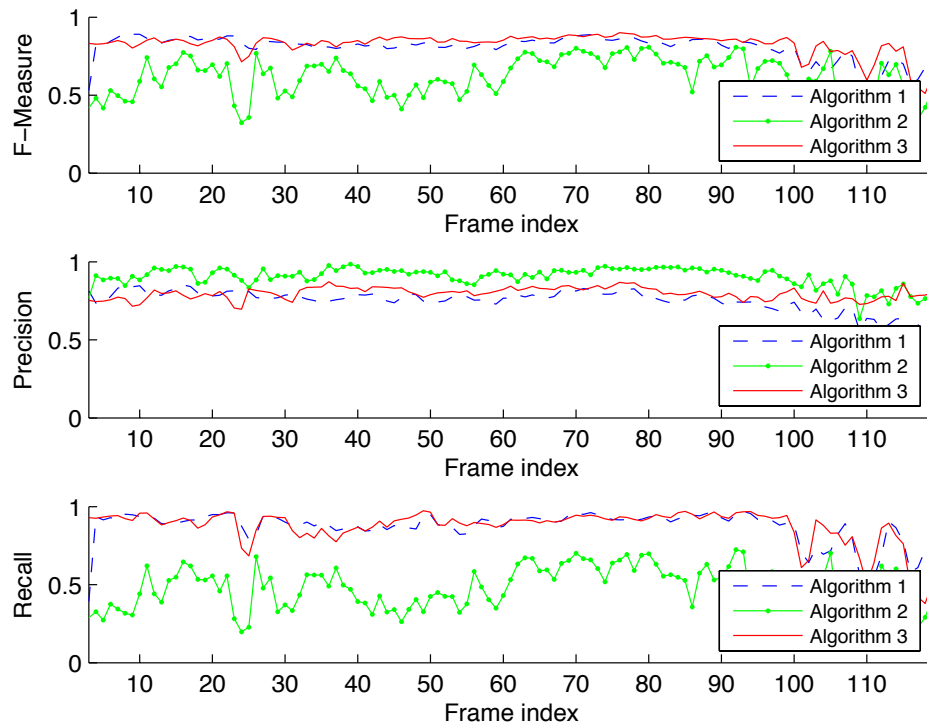
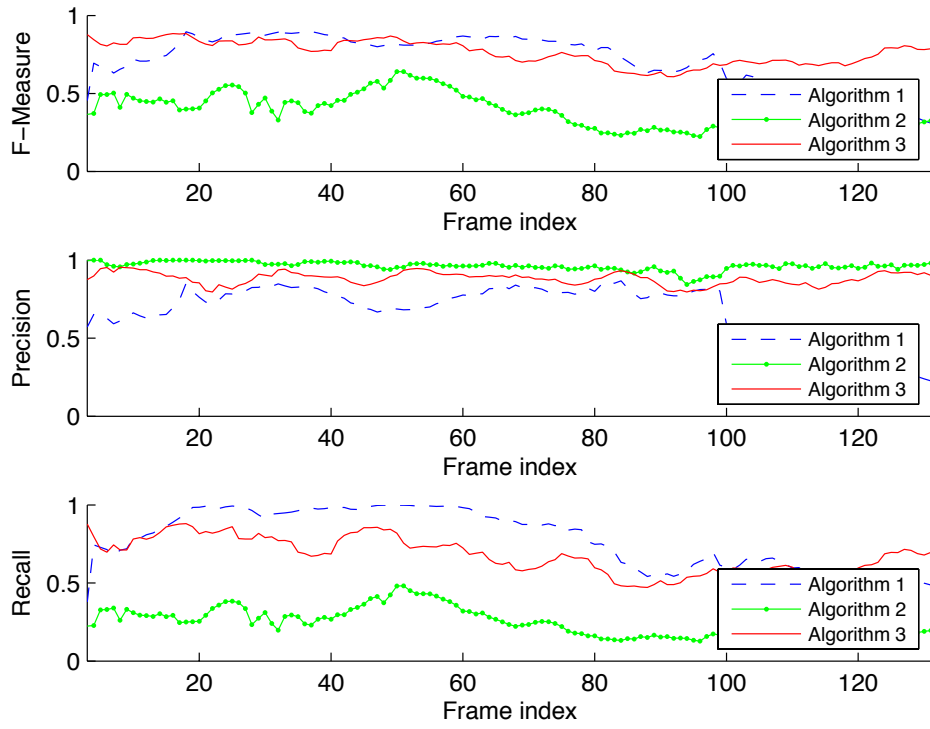
(a) *Stefan*(b) *BBC fish*

Figure 3.17: Precision, recall and F-Measure per frame using proposed and reference algorithms for the *Stefan* and *BBC fish* sequences.



(a) Horse

Figure 3.18: Precision, recall and F-Measure curves per frame using proposed and reference algorithms for the *Horse* sequence.

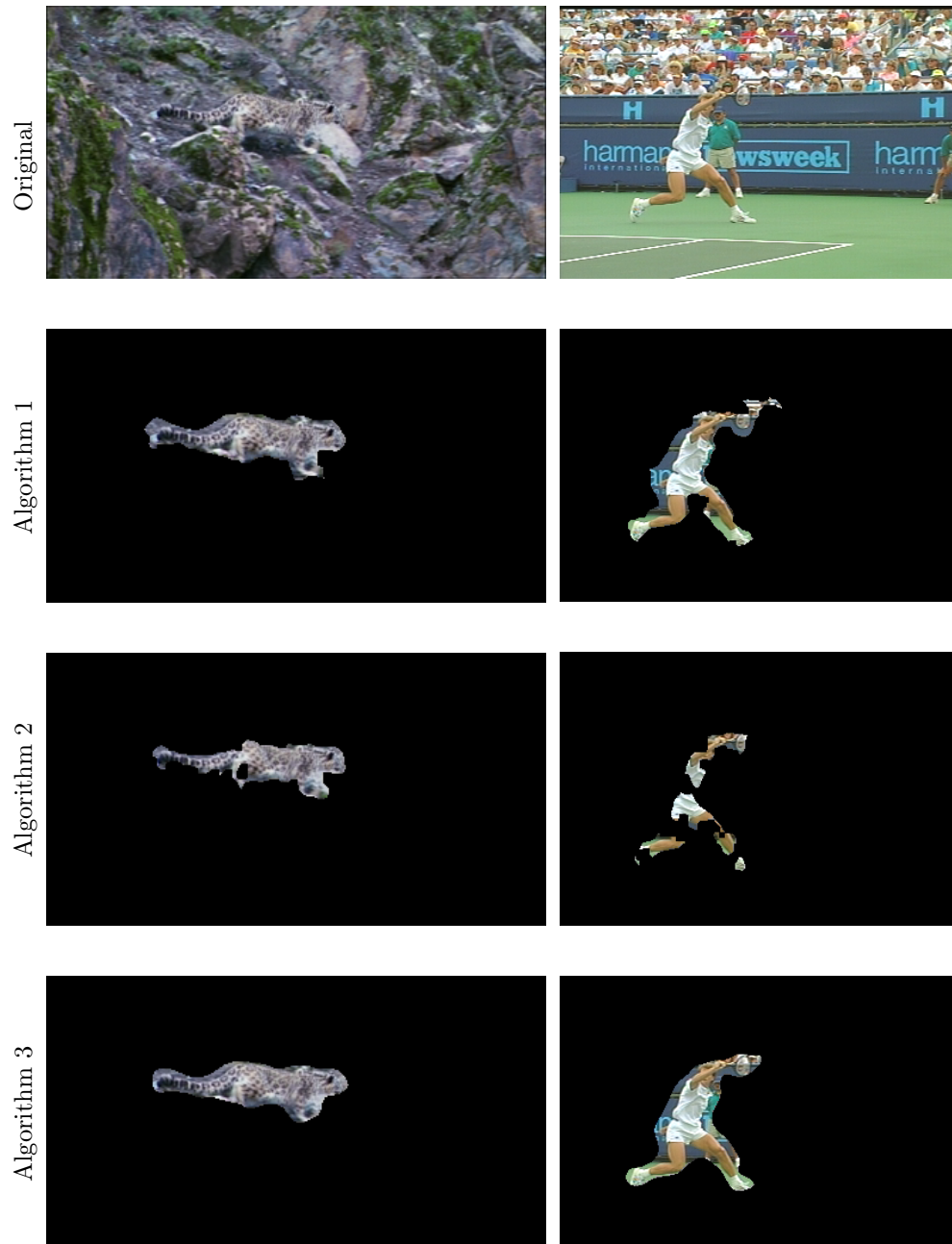


Figure 3.19: The first row shows example frames of *Mountain* (frame 95) and *Stefan* (frame 196). The second, third and fourth rows show segmentation examples using Algorithms 1, 2 and 3 respectively.

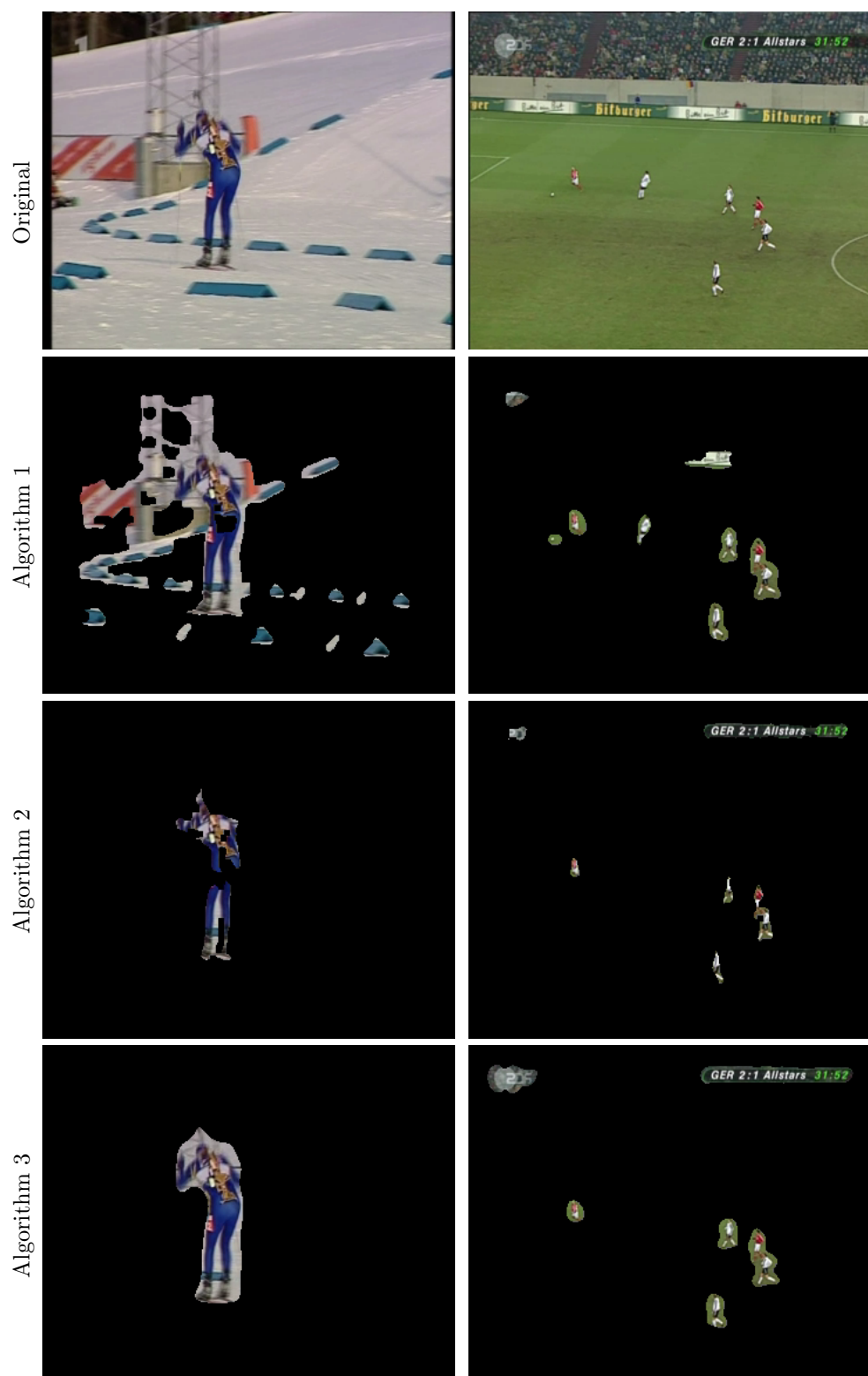


Figure 3.20: The first row shows example frames of *Biathlon* (frame 173) and *Allstars* (frame 162). The second, third and fourth rows show segmentation examples using Algorithms 1, 2 and 3 respectively.

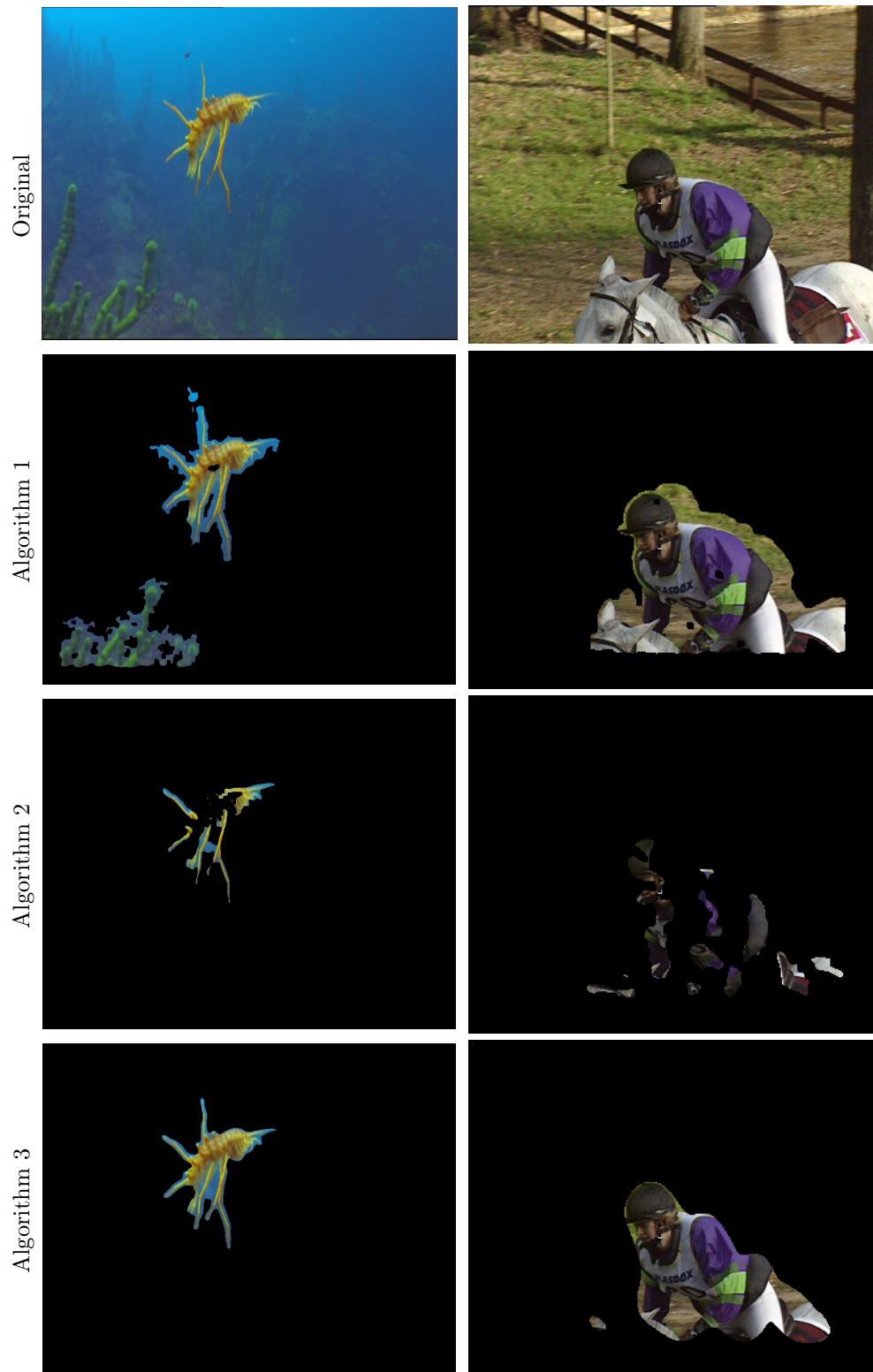


Figure 3.21: The first row shows example frames of *BBC fish* (frame 103) and *Horse* (frame 41). The second, third and fourth rows show segmentation examples using Algorithms 1, 2 and 3 respectively.

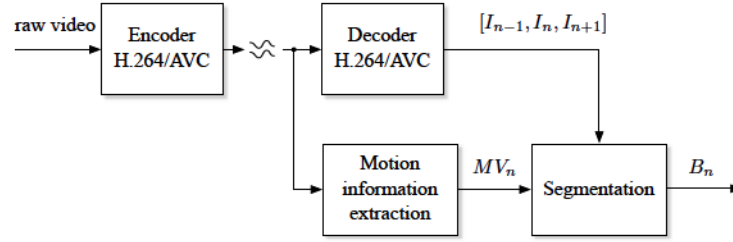


Figure 3.22: Segmentation system input when implemented at the decoder side.

### 3.5.4 Application on H.264/AVC compressed video data

In many application scenarios cameras are equipped with encoding capabilities and the reference video is not available at the decoder side for processing and extraction of content information. We test our approach with video streams from the state-of-the-art video coding standard H.264/AVC as depicted in Figure 3.22, where the input is the decoded video sequence and the motion vectors are extracted from the coded stream. The reference software KTA [53] has been used. We perform evaluation using motion vectors derived from H.264/AVC motion estimation (IPPP... GOP structure, EPZS motion estimation with  $32 \times 32$  search range,  $4 \times 4$  smallest block size, quarter-pel precision). A uniformly sampled  $4 \times 4$  MV field is obtained by macroblock splitting (e.g. when there is only one motion vector per  $16 \times 16$  macroblock, its value is assigned in every  $4 \times 4$  sub-block of it). In case of *intra* macroblocks, there is no motion information and the macroblock is omitted from global motion estimation and Gaussian filter calculation.

Table 3.6 reports the results, in the case that motion vector fields are obtained from encoding the test sequences with varying quantization parameters:  $QP \in \{4, 16, 28, 38\}$  and Figure 3.23 provides an overview. The results show that our approach is quite robust against bit rate changes, where motion information is not always representing real motion due to rate distortion optimization. By increasing  $QP$ , the number of *skip* macroblocks is also increased resulting in motion vectors with unreliable motion. Nevertheless, the results appear to be quite stable; up to  $QP = 28$  the maximum loss, in terms of F-measure compared with the  $QP = 4$  case, is 1% and for  $QP = 38$  the corresponding maximum loss is 13% for the *Horse* sequence.

In the cases of *Allstars*, *Stefan* and *BBC fish*, a slight increase (up to 2%) in terms of F-measure is observed by increasing  $QP$ . This can be explained, considering the fact that these sequences contain homogenous areas (soccer field, tennis field, blue sea) which, by increasing  $QP$ , are increasingly blurred as a consequence of the H.264/AVC deblocking filtering. This results in stronger blurring of minor details (spots in the sports field, spots in the sea etc.) and also increases the number of large macroblocks that potentially follow global motion, thus benefiting global motion estimation and eventually segmentation.

This approach can also be employed in cases of B-Frames presence. The advan-



Table 3.6: Average precision, recall and F-measure for varying quantization parameters. The PSNR column indicates average PSNR values [dB] between raw video sequences and ones coded with QP.  $\Delta_F = F_4 - F_{QP}$  and  $\Delta_{PSNR} = PSNR_4 - PSNR_{QP}$ , where  $F_4$  and  $PSNR_4$  stand for  $F$  and  $PSNR$  for  $QP = 4$  respectively.

Sequence	QP	PSNR	P	R	F	$\Delta_{PSNR}$	$\Delta_F$
<i>Allstars</i>	4	59.03	0.76	0.60	0.66	-	-
	16	47.18	0.76	0.59	0.65	-11.84	-0.01
	28	37.71	0.76	0.61	0.66	-21.31	0.01
	38	30.89	0.77	0.63	0.68	-28.14	0.02
<i>Biathlon</i>	4	59.89	0.77	0.88	0.82	-	-
	16	47.12	0.77	0.88	0.82	-12.77	0.00
	28	38.01	0.77	0.88	0.81	-21.87	-0.01
	38	31.69	0.74	0.88	0.80	-28.20	-0.02
<i>Mountain</i>	4	59.11	0.83	0.86	0.84	-	-
	16	46.11	0.83	0.86	0.84	-13.00	0.00
	28	34.53	0.82	0.87	0.84	-24.58	0.00
	38	27.01	0.81	0.87	0.83	-32.10	-0.01
<i>Race</i>	4	59.76	0.74	0.84	0.78	-	-
	16	46.57	0.74	0.84	0.78	-13.18	0.00
	28	37.43	0.74	0.84	0.78	-22.33	0.00
	38	30.89	0.73	0.79	0.74	-28.87	-0.04
<i>Stefan</i>	4	59.87	0.74	0.80	0.75	-	-
	16	46.43	0.74	0.80	0.75	-13.44	0.00
	28	35.93	0.72	0.80	0.74	-23.94	-0.01
	38	26.75	0.76	0.78	0.76	-33.13	0.01
<i>BBC fish</i>	4	59.46	0.81	0.84	0.82	-	-
	16	49.14	0.82	0.87	0.84	-10.32	0.02
	28	42.98	0.82	0.82	0.81	-16.48	-0.01
	38	36.67	0.78	0.72	0.74	-22.78	-0.08
<i>Horse</i>	4	59.99	0.90	0.66	0.76	-	-
	16	46.55	0.90	0.66	0.76	-13.45	0.00
	28	34.84	0.90	0.66	0.76	-25.16	0.00
	38	27.92	0.76	0.55	0.63	-32.07	-0.13



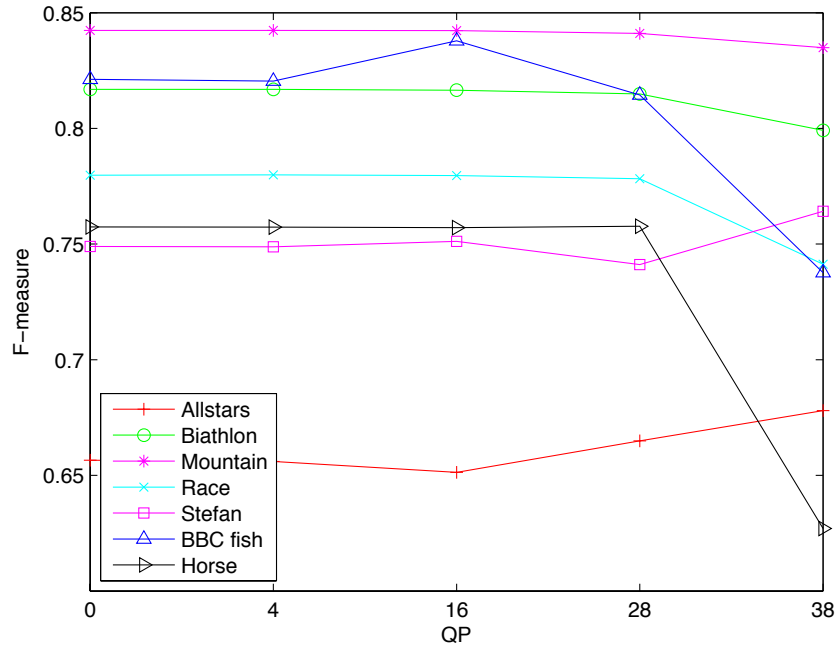


Figure 3.23: Average F-measure of segmentation at the decoder side with varying quantization parameters.

tage in this case would be the availability of motion vector fields from two directions in the encoder, and the disadvantage that the motion vector information may be prone to errors due to the larger distance between reference frames and subsequently larger displacements. Regarding I-frames, that contain no inter-frame motion displacement information, the adjacent P-frames' segmentation masks could be temporally interpolated in order to assign segmentation masks to them. Regarding the MPEG-2 compression technology, when applying our segmentation approach on MPEG-2 streams, a slight quality decrease in terms of F-measure should be expected as MPEG-2 only uses half-pixel motion compensation, instead of quarter-pixel that H.264/AVC uses, and does not use deblocking filters.

### 3.6 Chapter Summary

A motion-based object segmentation algorithm for video sequences with moving camera has been presented. The proposed algorithm is based on bidirectional inter-frame change detection. The proposed motion compensated error fusion scheme outperforms existing related ones. In addition to that, spatial error localization has been considered in the thresholding step for improving the segmentation efficiency. The issue of appropriate weight selection for weighted mean hysteresis thresholding has been addressed by employing a statistical approach. This enabled robust segmentation performance that avoids heuristics and training algorithms for parameter selection that are commonly used. Furthermore, a final post processing step has

been incorporated to enable temporal consistency of the segmentation masks using filtering of the preliminary binary masks, which is adapted according to the motion of the foreground.

The experimental evaluation has demonstrated the validity of the proposed method. It has been shown that proposed algorithm outperforms the reference algorithms and clearly improves the results in terms of F-Measure. The efficiency of the hysteresis weighted mean thresholding has shown superior performance compared to the weighted mean and the well known Otsu approach, since it accounts for spatial connectivity. By incorporating the background classification consistency step, the segmentation masks are temporally more consistent, false positives are eliminated, while foreground object detection has been more complete. In the case of multiple objects presence, the reference algorithms presented higher detection rates compared to the proposed algorithm. However, these high detection rates were followed by high false foreground detection rates. This resulted in the proposed algorithm to outperform the reference ones in terms of overall accuracy. It has also been shown that the proposed approach is quite robust under varying quantisation parameters that influence motion estimation quality.

# Content-aware Video Quality Assessment

---

## Contents

<b>4.1 Introduction</b>	<b>90</b>
4.1.1 Subjective video quality assessment	91
4.1.2 Objective video quality assessment	93
4.1.3 Advances in content-aware quality assessment	96
4.1.4 Motivation of the proposed work.	101
<b>4.2 Method M1: moving object-aware VQA Improvement</b>	<b>103</b>
4.2.1 Moving object segmentation	104
4.2.2 Foreground and background pooling	104
<b>4.3 Method M2: motion saliency for VQA Improvement</b>	<b>106</b>
4.3.1 Motion saliency model	106
4.3.2 Spatial pooling	109
4.3.3 Temporal pooling	112
<b>4.4 Experimental evaluation</b>	<b>115</b>
4.4.1 Distortion indicators through image quality models	116
4.4.2 Prediction performance indicators	118
4.4.3 Quality prediction performance of M1	119
4.4.4 Quality prediction performance of spatial pooling in M2	121
4.4.5 Quality prediction performance of temporal pooling in M2	128
4.4.6 Comparison of method M1 and method M2	130
<b>4.5 Chapter summary</b>	<b>132</b>

---

This chapter focuses on the field of video quality assessment, and specifically the improvement of video quality assessment algorithms by considering motion. The improvement refers to improving computational video quality assessment algorithms in order to be in closer agreement with the subjective evaluation of video quality. Taking into account methodologies from previous chapters, we study the effect of camera as well as object motion on the perception of distortion in video sequences. The contributions on objective video quality assessment are threefold. Firstly, we incorporate the moving-object segmentation scheme into a moving object-aware video quality assessment approach and examine related aspects. Secondly, we propose a motion

saliency model that exploits motion features on spatial level. Lastly, we propose an approach for consideration of global motion in the temporal dimension, leading to further improvements in the accuracy of video quality assessment. We discuss the benefits and drawbacks of each method and finally we perform evaluation by integrating them in existing objective quality models and also by comparing them to existing related state-of-the-art methods<sup>1</sup>.

The chapter is structured as follows. Section 4.1 begins with an introduction on video quality assessment and provides an up-to-date literature review in the topic. Section 4.2 describes the proposed moving object aware (content-aware) methodologies, whereas section 4.3 provides a description of the proposed motion saliency model for video quality prediction and the proposed approach for temporal consideration of global motion. Section 4.4 shows the validity of the proposed methods by presenting the experimental evaluation of the proposed approaches and finally section 4.5 summarises and concludes this chapter.

## 4.1 Introduction

The broad use of video in digital imaging and communication technologies has created a growing range of applications, such as video conferencing and internet protocol television (IPTV), where video content is delivered to end users. The lack of perfect communication channels, compression, and the presence of transmission errors are among the factors that have as consequence the end-user to receive impaired video content. This has triggered the increasing interest in research of *video quality assessment* (VQA) methodologies. As humans are the final judges of service quality, a key issue is the development of algorithms that efficiently assess the quality experienced by users which is widely referred to as *quality of experience* (QoE) [5]. QoE encompasses many different aspects, such as:

- quality of the video
- viewing setup and conditions, display type and properties
- quality and synchronisation of the accompanying audio
- interaction of the users with the service or display device
- aspects related to the viewers' individual interests, quality expectations and video experience.

Video quality is just one of these aspects, but it is arguably one of the most important. Methods for evaluating video quality play a critical role in quality monitoring to maintain the *quality of service* (QoS) requirements. The commonly acceptable way for assessing video quality is to conduct a large scale subjective study where a group of observers are asked to provide their personal opinions on the video, under laboratory conditions. This subjective evaluation can then be regarded as the

---

<sup>1</sup>The motion saliency model described in this chapter has been presented in [96].

ground-truth quality evaluation of the image sequence, and is usually expressed through the *mean opinion score* (MOS).

In practice, subjective experiments cannot be used in real-life scenarios, as they are time-, effort- and resource- consuming. Therefore, algorithms that evaluate the quality of visual content in an automated fashion are highly appreciated. Appropriate objective models are thus crucial to monitor the quality as experienced by the end user. Such computational metrics, are often based on some sort of comparison between the reference and the distorted signal and determine the quality degradation by accounting for signal or signal difference features. The ideal objective video quality metric should mimic the *human visual system* (HVS)<sup>2</sup>. Despite the fact that the understanding and modelling of HVS has begun quite early and the fact that there is considerable research effort on the field, the HVS itself has not been completely understood nor modelled. Existing image and video quality assessment approaches usually adopt specific assumptions related to observed properties of the HVS before proceeding to their design. In the following, we briefly present subjective video quality assessment methodologies and proceed to objective video quality approaches, before explaining the motivation of this work.

#### 4.1.1 Subjective video quality assessment

The quality of visual content as well as the performance of objective metrics is evaluated by means of specially designed subjective experiments. In subjective video quality experiments, a number of *subjects* are asked to rate the quality of the visual content presented to them. The resulting mean of all the processed individual scores, for a given sequence under specific conditions, is called mean opinion score. The mean opinion score, which has been initially introduced for quality assessment of telephony networks, is the most widely used subjective measure of video quality.

A subjective quality assessment experiment may be designed in many possible ways and there is a large variety of parameters that influence its outcome. The international telecommunication union (ITU) is constantly making efforts towards the creation of reliable frameworks that serve as a common basis for evaluation, by conducting standardisation activities. Two of the most widely used standards are the recommendations ITU-R BT.500-13 [97] and ITU-T P.910 [95] which provide specifications for the assessment of picture quality including general methods of test, the grading scales and the viewing conditions.

Depending on the availability of distorted content, ITU-R that is specialised for procedures for television pictures, has specified in ITU-R BT.500-13 [97] *single stimulus* and *double stimulus* methods. In the single stimulus continuous quality evaluation method (SSCQE) viewers rate the quality of a video sequence having watched only the impaired video stream. On the contrary, in double stimulus continuous quality scale (DSCQS) both the impaired and the reference video are presented to viewers. The DSCQS method, is claimed to be less sensitive to context, i.e. subjective ratings are less influenced by the severity and ordering of the impairments

<sup>2</sup>The HVS includes the eyes, parts of the brain and the nerve fibres connecting them.

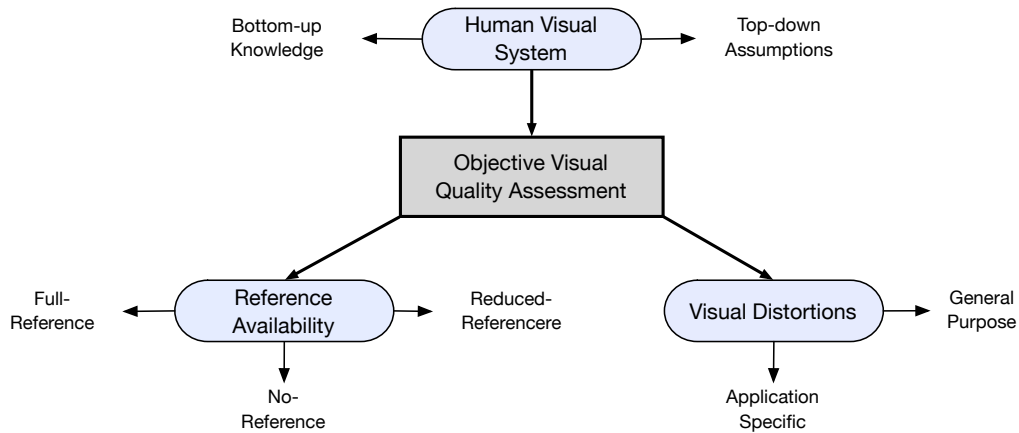


Figure 4.1: Classification of objective quality assessment approaches [101].

within the test session [98]. By contrast, the SSCQE enables viewers to dynamically rate the quality of an arbitrarily long video sequence (for instance using a slider mechanism with an associated quality scale) and is claimed to yield more representative quality estimates for quality monitoring applications. Thus, the choice of a single stimulus procedure is well suited to a large number of emerging multimedia applications, such as quality monitoring for video on demand, IPTV and internet media streaming. Additionally, it reduces the amount of time needed to conduct the study as compared to a double stimulus study, given a fixed number of human subjects.

In a similar manner, the recommendation ITU-T P.910 [95] which is indented for multimedia applications, defines the absolute category (ACR) procedure for single stimulus assessment and the degradation category rating (DCR) for double stimulus assessment. In contrast to the continuous quality scale used in SSCQE and DSCQS, in ACR and DCR the rating is provided using a discrete five-level scale. The video quality experts group (VQEG) is also conducting studies and making efforts on advanced specification of subjective evaluation related to technologies such as high definition and IPTV [99]. Regarding the parameters that influence experiments' results, a non-exhaustive list includes [100, 4]: the order of stimuli presentation, the range of video quality, test conditions such as room illumination, viewing distance, display properties (e.g. type, brightness, contrast, resolution) as well as personal aspects of the viewers, for instance their age and educational level.

In order to produce accurate and reliable results, subjective experiments require the appropriate design and precise implementation. This makes them time-, effort- and resource- consuming and therefore computational algorithms that assess video quality are highly appreciated.

### 4.1.2 Objective video quality assessment

According to the availability of the reference (undistorted signal), objective video quality assessment algorithms can be classified into *full-reference* (FR), *reduced-reference* (RR) and *no-reference* (NR) ones. In FR objective quality assessment algorithms the reference image sequence is available together with the impaired one and is used to assess its quality. As a consequence, FR approaches promise closer agreement with subjective assessment and are thus widely used. In this work we focus on FR objective quality assessment algorithms. In contrast to FR metrics, in the design of NR approaches, the reference information is omitted entirely, and are thus also referred to as "blind" metrics. Good performance of NR metrics is extremely valuable in applications such as video streaming, where (distorted) content is delivered to the users and the original signal is not available at the terminals. Nevertheless, and unlike the HVS that can judge visual quality without having knowledge of the reference, the task of designing metrics in this category is quite complex. Therefore, universal NR metrics are rare and efforts in this direction focus usually on application-specific metrics, such as ones focusing on particular sources of distortions such as blocking or blurring [102, 103, 104]. RR quality metrics [105, 106] take into account only a subset of the reference signal. In this case, only a set of extracted features of the reference image sequence is used for quality assessment, instead of the image sequence itself. In this way RR approaches, may be seen as a compromise between FR and NR approaches, as they combine advantages of each of these categories, by considering only part of the reference signal and avoiding the necessity for the whole original one.

Taking into account further aspects of visual quality metrics, the classification of approaches for objective visual quality assessment can be extended as depicted in Figure 4.1, following the classification suggested by Wang *et. al* [101]. Depending on the way HVS properties are incorporated in the design of the metrics, approaches are distinguished in bottom-up and top-down ones. Approaches in the first category emulate HVS properties by means of computational algorithms and incorporate them into the metric design. In contrast, approaches following the top-down approach, are based on high-level assumptions regarding the HVS and avoid emulating specific properties independently. An example in this case is the *structural similarity index* [107], which is later on discussed and, as other approaches in this category, treats the HVS as a black box and focuses on the input-output relation instead of dealing with individual HVS functionalities. Another aspect taken into account for classification is the considered visual distortions, based on which, general purpose and application-specific metrics are distinguished. Metrics in the first category, do not make specific assumptions related with distortions and aim to be universally applicable. On the other hand, application specific metrics, make assumptions about specific distortions met in the visual content and are thus simplified resulting in better performance for the specific application, with the cost of a non-uniform universal performance. In the following a brief overview of representative existing objective quality metrics is given.

Among the most widely used metrics in this category are the *mean square error* (MSE) and the *peak signal to noise ratio* (PSNR). Let  $R(i, n)$  and  $D(i, n)$  be the  $i$ -th pixel of the  $n$ -th frame in the reference and the distorted images respectively. The MSE between these two  $L$ -bit images is given by:

$$MSE(n) = \frac{1}{N} \sum_{i=1}^N [R(i, n) - D(i, n)]^2 \quad (4.1)$$

and the PSNR for the whole sequence of  $M$  images is given by:

$$PSNR = 10 \cdot \log_{10} \frac{(2^L - 1)^2}{\sum_{n=1}^M MSE(n)} \quad (4.2)$$

where  $N$  is the total number of pixels in the image. For 8-bit images the nominator of the logarithm becomes 255.

The wide use of PSNR is mainly attributed to its simplicity while managing to reflect satisfactorily the intended deviation between distorted and reference images. Nevertheless, PSNR has been arguably claimed [108] to be a poor predictor of visual quality perceived by humans. It is regarded as *fidelity* rather than *quality* metric, since it reflects pixel-to-pixel differentiation and does not take into account human perception, often resulting in poor consistency with subjective evaluations. Several extensions of PSNR have been proposed in the past, with *visual signal to noise ratio* (VSNR) [109] being such an example. VSNR considers both low- and mid- level properties of the human visual system, by applying the results of psychophysical experiments towards quantifying the perception of distortions in natural images. More specifically, it operates in two stages; in the first stage contrast thresholds for detection of distortions in images are computed using wavelet-based methods in order to determine whether the existing distortions are perceivable. If the distortions are not perceivable according to the calculated threshold, no further analysis is required and the image is assumed to have perfect quality. If the distortions are perceivable (above the calculated threshold) then a second stage follows that operates based on a non-sophisticated modeling of visual properties of the human visual system.

The *structural similarity index* (SSIM), proposed in 2004 [107], is based on a different philosophy from the error visibility (MSE based) approaches. According to Wang *et al.*, pixels exhibit strong dependencies, especially when they are spatially proximate, and these dependencies carry important information about the structure of the objects in the visual scene. Based on the assumption that these dependencies are indeed correlated with functionalities of the human visual system, it captures the structural similarity between the reference and distorted image to estimate the perceptual quality. Given a pair of reference and distorted images, the luminance  $lum(r, d)$ , contrast  $con(r, d)$  and structural similarity  $str(r, d)$  components are computed based on the mean, variance and covariance of small image patches. Subsequently, they are combined using a combination function  $f(\cdot)$ :

$$SSIM(r, d) = f[lum(r, d), con(r, d), str(r, d)] \quad (4.3)$$



where  $r, d$  denote the image patches from original and distorted respectively, extracted from the same image location. After the above functions have been defined (4.3) yields:

$$SSIM(r, d) = \frac{(2\mu_r\mu_d + c_1)(2\sigma_{rd} + c_2)}{(\mu_r^2\mu_d^2 + c_1)(\sigma_r^2 + \sigma_d^2 + c_2)} \quad (4.4)$$

where  $\mu_r, \mu_d$  and  $\sigma_r, \sigma_d$  denote the mean intensity and standard deviation of image signals  $r$  and  $d$  respectively,  $\sigma_{rd}$  denotes the covariance between  $r$  and  $d$ , and  $c_1, c_2$  are small constants to avoid instability.

SSIM was extended to video-SSIM [110] where it was proposed to assign higher weighting in darker areas and fast moving ones based on empirical parameter settings. Initially, the idea of SSIM was conceived in a single scale, and multiscale versions of it followed. The most widely used is the MS-SSIM proposed in [111]. It incorporates SSIM computation in  $M$  scales in order to provide more flexibility in incorporating the variations of viewing conditions, in comparison to the single-scale case. It is formulated as:

$$MS-SSIM(r, d) = lum_M(r, d)^{a_M} \cdot \prod_{j=1}^{M-1} con_j(r, d)^{\beta_j} str_j(r, d)^{\gamma_j} \quad (4.5)$$

where  $a, \beta$  and  $\gamma$  are parameters that adjust the relative importance of the three components  $lum, con$  and  $str$ .

Sheikh *et al.* introduced the *visual information fidelity criterion* (VIF)[112] for quality assessment through an information-theoretic framework. Specifically, the degradation of visual quality, seen as loss of information due to a distortion, is quantified using the information extracted from the reference image and the amount of this reference information that can be still extracted from the distorted image. Towards this direction, the images are modelled using *Gaussian scale mixtures* (GSM). Subsequently, VIF is given by:

$$VIF = \frac{\sum_{j \in \text{subbands}} I(\vec{C}^{N,j}; \vec{F}^{N,j} | s^{N,j})}{\sum_{j \in \text{subbands}} I(\vec{C}^{N,j}; \vec{E}^{N,j} | s^{N,j})} \quad (4.6)$$

where  $I(\cdot)$  denotes mutual information,  $\vec{C}$  the GSM,  $N$  the number of GSM used,  $s$  is a random field of positive scalars, and finally  $\vec{E}$  and  $\vec{F}$  denote the HVS model output for the reference and the distorted image respectively.

Another approach, the *video quality model* (VQM) [113], which is adopted by the american national standards institute (ANSI), analyses 3D spatio-temporal blocks to extract features for estimating the video quality map whereas the *motion-based video integrity evaluation* (MOVIE) metric [114] utilises properties of the visual cortex neurones to track perceptually relevant distortions both spatially and temporally and evaluates motion quality along computed motion trajectories. Relying on 3D optical flow estimation, the latter is a rather computationally complex metric.

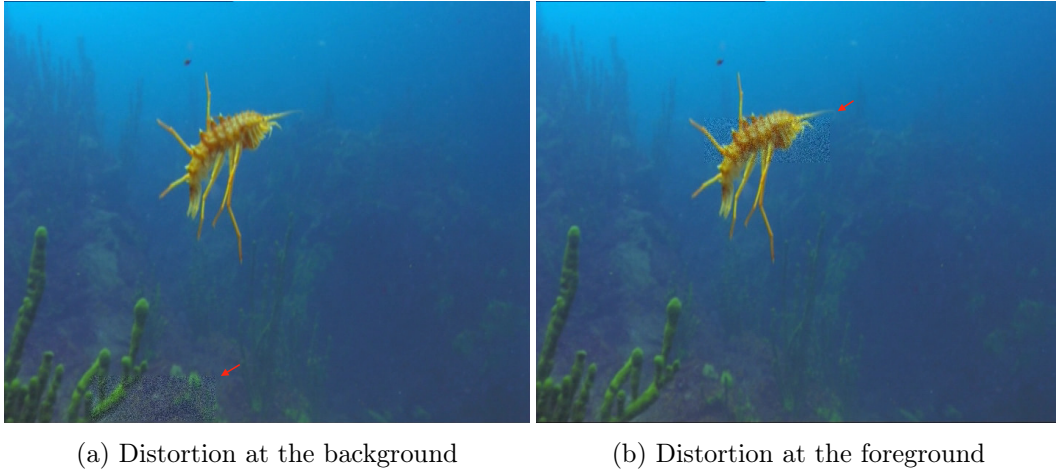


Figure 4.2: Deviation of objective and subjective quality assessment on the *BBC fish* sequence. The distorted areas are indicated with red arrows.

#### 4.1.3 Advances in content-aware quality assessment

Objective quality assessment models for image and video quality assessment compute the quality scores based on the assumption that content over space and time is of equal interest to the observer. It is assumed thus that distortions in different regions in space and time contribute equally to the overall quality perception of the video. Nevertheless, humans do not see in a way that resembles linear scanning. Rather, it is claimed to sample and process the physical world in a way that is space and temporally variant, which has led to considerable interest in visual quality assessment approaches [115] [6] [116] in recent years. However, the mechanism and functionalities of the HVS have not been completely modelled, despite the attempts of existing objective quality assessment approaches to incorporate individually particular HVS functionalities in visual quality assessment methodologies. The reasoning behind is that the HVS and the higher cognitive visual information processing is not fully understood and thus very difficult to incorporate it in an objective quality assessment metric. Thus, conventional metrics such as PSNR are still widely used for evaluating visual quality. Nevertheless, with the advances in objective quality assessment approaches that incorporate perceptual knowledge, the broad use of alternative metrics to PSNR is coming closer to realisation.

Towards understanding how traditional metrics can benefit from perceptual knowledge, we illustrate some cases indicating the shortcomings of the traditionally used PSNR with respect to the way visual content is in general assessed by humans. Let us examine, the example in Figure 4.2 which depicts a fish swimming in the seabed. The viewer will typically focus his attention mainly on the fish and secondly on the seabed. Consequently, the blurring blemish on the sea region (bottom left corner) in Figure 4.2(a) will be probably perceived only under thorough examination. On the contrary, the blurring which takes place on the region depicting



(a) PSNR = 32.533, distortion type: compression  
 (b) PSNR = 33.488, distortion type: packet loss

Figure 4.3: Deviation of objective and subjective quality assessment under transient and uniform distortion types on the *pedestrian area* sequence, LIVE video database.

the fish, in Figure 4.2(b), will be more pronouncedly perceived compared to the former case. Thus the location of the second blurring seems to play an important role on the perceived quality, resulting in the impression that Figure 4.2(b) has worse quality than Figure 4.2(a). Evaluation of the quality using PSNR is however not that revealing; both images have the same PSNR (41.5 dB).

Another example is illustrated in Figure 4.3, where an image is distorted in two different ways: Figure 4.3(a) is distorted with compression artifacts (no packet loss) introduced by H.264/AVC coding, while Figure 4.3(b) contains distortions coming from H.264/AVC coding along with packet loss (followed by error concealment). Therefore the distortions are considered spatially dispersed and localized respectively. A subjective examination of the two images would probably give the impression to the observer that the second image suffers from more severe quality degradation, compared to the first one. However, according to the PSNR the quality is worse in the first case compared to the quality of the second image. As it can be observed, the transient distortion, i.e. the packet loss, in Figure 4.3(b) takes place on a region that corresponds to a moving person, and thus becomes more noticeable than a possible packet loss in the background region. It would be useful thus to be able to distinguish perceptually "important" image regions and predict the expected impact of possible distortions on quality degradation.

Current research efforts focus on perception-aware quality metrics that are increasingly adopted in video processing systems. Such metrics try to incorporate properties of the human visual system into their design in order to achieve higher correlation with the visual perception of quality. In the following, we discuss the major approaches in this direction which is the field that the proposed work belongs.

The human visual system has attracted the interest of researchers from a physiological and psychological point of view in the last decades. Aspects of vision science [117] have been exploited in computer vision and particularly in the field of quality assessment. We mention here some of the basic characteristics that have found useful application in the field of visual quality assessment [109], namely *visual masking*,

*contrast sensitivity*, *global precedence* and *visual attention*. Visual masking is related to enhancing the influence of specific image regions on perception based on the luminance distribution and spatial localization of the visual target. This characteristic is exploited by the widely used SSIM [107], where image distortions are assumed to be the visual target. Another property is that in order a human to visually detect a target (e.g. a distortion), its contrast must be higher than a certain contrast detection threshold. Based on this, contrast sensitivity is considered to be the inverse of this threshold. Global precedence is the aspect related with the visual system's preference for integrating edges in a coarse-to-fine-scale fashion.

Another significant property of the HVS is visual attention [6]. It refers to the shift of visual focus across the visual scene to the most relevant regions. Humans can perform this procedure instantaneously and unconsciously. The mechanism behind visual attention involves complex higher cognitive processing and visual attention is regarded to be subjective and vaguely defined by the HVS. Therefore approaches that attempt to model it from an algorithmic point of view have been created. These models aim principally to predict human gaze when observing visual content. One of the most known models is the saliency model proposed by Itti *et al.* [118, 119]. It is expressed using maps that represent saliency at every spatial location in the visual field by a scalar quantity. It is based on colour, intensity and orientation features, which are calculated and combined based on the behaviour and the neuronal architecture of the early primate visual system. Specifically, the saliency maps calculation proceeds as follows: the colour image is low pass-filtered and subsampled so that nine spatial scales are created using pairs of Gaussian pyramids. The feature maps are generated by computing the difference between a centre fine scale sub-band and a surround coarser scale. The different modalities that the different features represent (colour, intensity, orientation) are addressed by using a normalisation operation (center-surround operation), which promotes feature maps with a small number of strong peaks and suppresses ones with numerous comparable peak responses. The normalised feature maps at different scales are combined into the following conspicuity maps:

$$\bar{\mathbf{I}} = \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^4 \mathcal{N}(I(c, s)) \quad (4.7)$$

$$\bar{\mathbf{C}} = \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^4 [\mathcal{N}(RG(c, s)) + \mathcal{N}(BY(c, s))] \quad (4.8)$$

$$\bar{\mathbf{O}} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \mathcal{N} \left( \bigoplus_{c=2}^4 \bigoplus_{s=c+3}^4 \mathcal{N}(O(c, s, \theta)) \right) \quad (4.9)$$

where  $\bar{\mathbf{I}}, \bar{\mathbf{C}}, \bar{\mathbf{O}}$  are the conspicuity maps for intensity, color and orientation respectively,  $I(c, s), C(c, s), O(c, s, \theta)$  are the intensity, color and orientation maps obtained at the center scale  $c$ , surround scale  $s$ , and at angle of orientation  $\theta$ .  $\mathcal{N}(\cdot)$  represents the center-surround normalization operator mentioned above, while  $\bigoplus$  denotes

the operation of point-by-point addition after interpolation to the finer scale. The saliency maps are then obtained as:

$$LS = \frac{1}{3} [\mathcal{N}(\bar{\mathbf{I}}) + \mathcal{N} [\bar{\mathbf{C}} + \mathcal{N}(\bar{\mathbf{O}}) ]]. \quad (4.10)$$

Osberger *et al.* [120] studied the influence of visual attention on image quality assessment, by weighting spatially the visible errors, depending on an importance map of the region in which they occur. Experimental evaluation, reported improvement over the conventional approach of PSNR. Following a similar perspective, in the work of [121] the authors assume "perfect" region of interest (ROI) knowledge by using a subjectively created ROI database. They validate the proposed theoretical model by performing training and validation on given image datasets that are, for this purpose, distorted with JPEG coding artifacts. The experimental evaluation confirmed the initial observation that structural degradations in the ROI have more severe impact on perceptual quality than degradations outside of it. They used four image quality models: PSNR, SSIM, the reduced-reference image quality assessment (RRIQA) [105] and the normalized hybrid image quality metric ( $\Delta$ NHIQM) [122] and it is interesting that the specific model design concluded in different parameter setting for each of the quality metrics. In [123] it is suggested that full-reference and no-reference image quality metrics could benefit from content aware information, specifically face detection. This is a very promising statement, but an elaborate description and experimental evaluation of it has been absent. In [6] the authors summarise existing methodologies on perceptual quality assessment through visual attention and point out the involved challenges. They observe that visual attention is strongly influenced by three cues that are mutually reliant and correlative; spatial location, low-level features (e.g. colour, motion, orientation, size) and higher-level feature (e.g. objects, faces). From another perspective, knowledge of the important (according to content) regions is though not enough to assess the quality of an image. Therefore, it is pointed out that the spatial distribution of the artifacts affects viewers attention as well. For instance, if a distortion is distributed densely in a small spatial region, while the rest of the image is lightly corrupted, then it is possible that this small distorted region will attract the viewers attention. Indeed, in the case of full-reference quality metrics, it is possible through comparison of the reference and the distorted images to decide where errors take place.

In [124] the authors used eye tracking data from an experimental setup as ground truth saliency to study the impact of the (perceived annoyance of) the distortion duration compared to distortion taking place on a salient region. They used PSNR and an own proposed metric, namely the temporal trajectory aware video quality measure, (TETRAVQM) [125] as video quality metrics to validate experimentally that the observers tend to distinguish annoyance levels more pronouncedly with respect to the saliency of the distorted region compared to the duration of the distortion. In [126] the authors used an image segmentation method, namely a ramp discontinuity model for multi scale segmentation, to quantify the effect of intra- and inter-region image distortions on the quality reduction of an image. They used PSNR, MS-SSIM

and VIF and they named the proposed metric segmentation-based perceptual image quality assessment (SPIQA). It is interesting to note that the authors' assumption that the size of the "important" region playing a role in the overall quality metric, proved invalid in the experimental evaluation. In similar fashion, in [127] the authors adopt an approach based on information theory to model visual saliency, and based on that they determine the weighted pooling strategy for the overall video quality metric.

Ma *et al.* [128] proposed a visual saliency estimation approach that incorporates spatial and temporal information derived from motion trajectories. Specifically a frame level Quaternion representation was proposed, that is based on the quaternion Fourier transform. This representation, which considers spatial content and temporal motion trajectories, was used to construct the visual saliency map, whose incorporation in several video quality metrics reported significant improvement in performance. The authors in [129] proposed a video quality metric that employs the structural information contained in two descriptors extracted from the 3D structure tensors, and its corresponding eigenvector, whereas in [130] a model of human visual speed perception was incorporated to model visual perception in an information communication framework.

### Temporal dimension

Approaches for video quality assessment typically deliver quality scores per frame, and this procedure is often referred to as *spatial pooling*. For a video sequence containing  $T_L$  frames, a *temporal pooling* approach is necessary to obtain the overall quality score for the video sequence. The most straightforward approach and maybe the most widely used is direct average pooling over  $T_L$  frames. Which also facilitates the wide use of image quality assessment metrics (such as PSNR) for video quality assessment. However, more sophisticated temporal pooling approaches, that exploit signal properties on the temporal dimension, have been studied.

An early comparison of temporal pooling methods is found in [131]. It was found that approaches that incorporate the *recency effect* and the *worst quality section influence* yield better results. The recency effect reflects the phenomenon that the viewers's judgement in relation to quality is strongly influenced by what they see in the last moments of the video sequence. The worst quality section influence reflects the hypothesis that the most degraded part of a sequence influences more strongly the viewers' judgment on quality. A widely employed temporal pooling strategy that enables the assignment of emphasis on highly distorted regions is the Minkowski summation [132]. This strategy was used in the perceptual distortion model (PDM) [4], where after incorporating spatial and temporal modelling aspects of the HVS towards calculating distortion information in different channels, the Minkowski summation was used in pooling spatial as well as temporal errors between the reference and distorted video sequences.

In [110] it was suggested that the larger the existing motion is, the smaller the assigned temporal weight should be. In [133] it was implied that the important value



for overall subjective quality judgement of a video sequence is not the duration of a dip in quality, rather the depth. The same position is also adopted in [134]. Temporal variations of distortions are also accounted for in the work by Ninassi *et al.* [135] by combining short-term and long-term temporal pooling techniques. Specifically, short-term pooling was identified to be particularly beneficial for improving the quality prediction performance of the quality metrics. Moorthy *et al.* [136] proposed the motion-compensated structural similarity index (MC-SSIM) that combines block-based motion estimation with SSIM [107]. Each  $8 \times 8$  block of the reference and the distorted frames is motion compensated using the corresponding preceding frame and the results are used to compute the temporal quality.

In [137] the authors defined indicators for global and local quality and, in order to obtain a metric over all frames, they proposed a temporal pooling approach that adopts the following finding, based on previous studies, which is related to the recency effect. Specifically, they take into account the phenomenon that frames in the beginning and in the end of the video sequence influence more significantly the overall quality. In [138] the authors proposed a temporal pooling approach, the *IQpooling*, where based on computed ego motion, they classify the best and worst values using k-means clustering and perform weighting based on the cardinality of each cluster. Nevertheless, it is interesting that in a recent study [139] the efficiency of the simple direct average was emphasized. Specifically, it was found that in the context of HTTP adaptive streaming, where viewing sessions last long, the direct average temporal pooling approach performs on par with sophisticated temporal pooling algorithms.

#### 4.1.4 Motivation of the proposed work.

Several approaches exist [121, 126, 127] that investigate the effect of employing *visual attention*, on image quality metrics. Additionally to visual attention, which is strongly connected to HVS properties, and low-level features such as colour and edges, motion is a feature that is often neglected in the design of visual quality metrics. The very important role of motion and human perception dates back at the beginning of last century, when Koffka and Wertheimer formulated the Gestalt principle and specifically the *grouping law of common fate* [54], according to which

*"Humans tend to perceive elements moving in the same direction as being more related than elements that are stationary or that move in different directions."*

Based on the established connection between motion and perception and considering that moving regions will likely attract the viewer's attention, the main idea of this chapter is to exploit motion for video quality assessment, in spatial and temporal level. We exploit motion between successive video frames by considering motion features along temporal trajectories. The basic motivation is the observation that the way humans perceive and evaluate distortions is not independent of the semantic

information contained in a video frame. The examples in Figures 4.2 and 4.3 highlighted this observation. If the distortion occurs in a non-salient region, the human observer may hardly notice the quality deterioration. However, in the case that the distortion affects a salient region, the human observer is more likely to notice the distortion and have the impression of a more "bad quality" video scene. These examples indicate the deficiency of considering spatially uniform approaches for quality assessment.

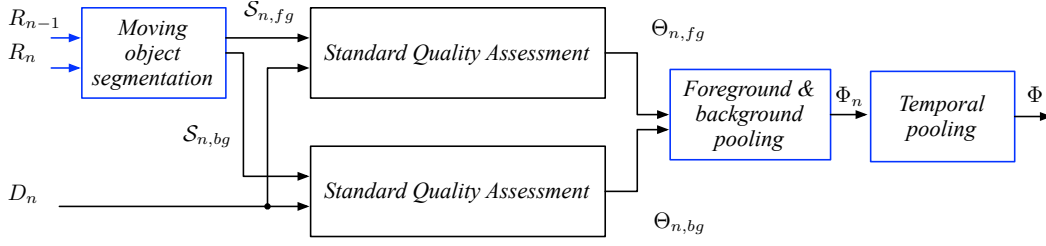
Two questions have arisen at this point: a) what is a reliable way to identify important content information based on motion in a sequence of image frames? and b) how can this information be exploited towards visual quality assessment? In the following, based on the assumption that object motion in a video sequence attracts visual attention, we deal with the first question by exploiting the foreground object motion as well as global motion information. Towards reaching an answer to the second question, we study several approaches that are detailed described in the next sections.

Another important aspect regarding the assessment of video quality is temporal correlation. The consideration of only spatial correlations is satisfactory for image quality assessment, more sophisticated considerations are required though in the case of video quality assessment, where temporal correspondences constitute a determining factor. Often in the latter case, correspondences between frames are ignored. Thus, suitable approaches specially designed for video quality assessment that take into consideration motion and especially global motion need to be developed.

### Proposed approaches

In *Method M1* the distortion maps are independently computed in the foreground and the background regions to obtain quality measures for each region. The distortions in the foreground and background regions are then participating in the determination of the frame quality level by taking into account each region's properties, resulting in a moving object-aware quality metric. In the second approach, *Method M2*, the foreground and the background regions are not known, and the distortion map undergoes spatial weighting according to an importance map. We propose a novel motion saliency model (*motion saliency*) that is derived from motion information included in successive frames and is inline with properties of human visual properties. The motion saliency model is subsequently incorporated in the VQA framework at spatial level as the importance map. By assigning higher weighting to regions with salient motion, we expect these regions to attract humans' attention, and thus make distortions on these areas more perceivable. Further, in the framework of *Method M2*, we propose a third approach for modelling of the global motion in the temporal dimension and study its impact on perceived quality. Finally we evaluate the performance of our proposed methods by comparing them with related state-of-the-art approaches.



Figure 4.4: *Method 1* framework overview.

## 4.2 Method M1: moving object-aware VQA Improvement

This section introduces a framework to incorporate the moving object segmentation approach, proposed in chapter 3, into a moving object-aware quality model, in order to improve the performance of existing quality metrics.

Figure 4.4 illustrates an overview of the visual quality assessment framework of *Method M1*. The first step is to segment the reference (undistorted) image  $R_n$  and extract the foreground and background regions, denoted as  $\mathcal{S}_{fg}$  and  $\mathcal{S}_{bg}$  respectively, of the current frame  $n$ , using the preceding frame  $n - 1$ . In the next step, based on the derived segmentation mask, the corresponding segments in the undistorted and distorted frames are compared using standard quality assessment approaches, in order to obtain an indication of the distortions' distribution in the frame. Subsequently the segments' quality values,  $\Theta_{n,fg}$  and  $\Theta_{n,bg}$ , are incorporated in the foreground/background pooling stage to obtain the quality indication at frame level (local quality). Finally, at the temporal pooling stage, the local quality values are pooled using direct average to obtain the overall quality indication  $\Phi$ . Frame indices are following omitted for brevity.

As already pointed out, our main motivation is to avoid the assumption that content attracts equally the viewer's attention over the frame. The proposed weighted model incorporates moving object segmentation in its formulation, so that it depends on the moving content. Specifically, the content is characterised by temporal features (i.e. motion) that determine indirectly the spatial location boundaries. Thus, the issue of content-aware quality assessment is addressed by assigning (motion-driven) spatially varying weights on the impact of conventional quality metrics.

The overall quality is computed as a weighted sum of the segment's quality metrics (Figure 4.4). Each segment's quality indicator is independently computed, using a standard quality assessment approach, and contributes to the *moving object-aware quality model* (MOAQM) by taking into consideration region's features. Let  $\Theta$  be the conventional image quality metric as a general definition and  $\Phi$  the content-aware quality metric. In the general case the segmentation aware quality assessment

metric  $\Phi$  can be expressed according to [127] as:

$$\Phi = \frac{\sum_{i=1}^N w_i \cdot \Theta_i}{\sum_{i=1}^N w_i} \quad (4.11)$$

where for each region  $i$ ,  $\Theta_i$  is the corresponding local quality measure and  $w_i = f(s_i)$  is the weight, which depends on the region's saliency  $s_i$ . The challenges are thus to express  $s_i$  and to find the appropriate function  $f$  that yields a content-aware quality metric  $\Phi$  that correlates well with subjective evaluation.

This formulation may involve several segments of an image and assignment of various weights. Here we incorporate the moving object segmentation scheme described in the previous chapter and assign different weighting to each segment  $\mathcal{S}_{fg}$  and  $\mathcal{S}_{bg}$ . Thus, the (local) moving object-aware quality model  $\Phi_n$  of the  $n$ -th frame is given as:

$$\Phi_n = \frac{\omega_{fg} \cdot \Theta_{fg} + \omega_{bg} \cdot \Theta_{bg}}{\omega_{fg} + \omega_{bg}} \quad (4.12)$$

where  $\Theta_{fg}$  is the local quality measure in the foreground region,  $\Theta_{bg}$  is the local quality measure in the background region. Of course, the case where  $\omega_{fg} = \omega_{bg} = 1$  leads to the conventional metric.

#### 4.2.1 Moving object segmentation

The algorithm for bidirectional motion-based object segmentation using hysteresis, overviewed in Figure 3.1, is used to segment each reference frame of the image sequence into foreground  $\mathcal{S}_{fg}$  and background  $\mathcal{S}_{bg}$  segments. According to this segmentation mask the corresponding regions of the reference and the distorted frames are compared (to measure their quality level). Each segment is then assigned a weight  $\omega_{fg}$  and  $\omega_{bg}$  respectively according to the corresponding properties. An example of segmentation mask is shown in Figures 4.5 (a) - 4.5 (b).

#### 4.2.2 Foreground and background pooling

At the frame level, in order to determine each segment's contribution to the quality measure, we examine the following approaches. Each segment's contribution is quantified using an extracted feature, namely: motion, motion combined with the segment's size, saliency, and saliency combined with the segment's size. Each approach is described in the following.

**Motion** As the goal is to incorporate an aspect of the relative motion in the scene (between objects and background), we employ the global motion compensated error. Specifically, equation (3.7) is used for the quantification of the moving object's motion in relation to the camera motion. Hence, the average global motion compensated error of each segment, which is normalized in equation (4.12) is taken into

account. In this way, the weight  $\omega_i$ ,  $i \in \{bg, fg\}$  is set equal to  $mot_i$ :

$$mot_i = \frac{1}{|\mathcal{S}_i|} \sum_j^{|\mathcal{S}_i|} E_j \quad (4.13)$$

where  $|\mathcal{S}_i|$  is the number of pixels in segment  $i$  and  $E_j$  (see equation (3.7)) denotes the global motion compensated error in pixel position  $j$  based on the previous frame.

**Motion and segment size** In this case the size of each segment is taken into account together with the calculated motion. We consider that the smaller segment plays a more important role in the quality assessment, thus the segment's weight is given as:

$$\omega_i = mot_i \cdot \frac{1}{siz_i} \quad (4.14)$$

where  $i \in \{bg, fg\}$  and

$$siz_i = \frac{|\mathcal{S}_i|}{|R_n|} \quad (4.15)$$

is the relative size of the background or foreground segments ( $\mathcal{S}_{fg}$  or  $\mathcal{S}_{bg}$  respectively) of reference frame  $R_n$  and  $|R_n|$  is the number of pixels in  $R_n$ .

**Local saliency Itti-Koch-Niebur** The LS-IKN model [118], presented in subsection 4.1.3, accounts for the behaviour and the neuronal architecture of the early primate visual system, and is based on colour, intensity and orientation features. It is expressed using maps that indicate saliency at each spatial location in the visual field by a scalar quantity.

By the incorporation of the local saliency model, equation (4.10), we aim at weighing each segment's impact on the quality model design, exploiting the different modalities (colour, intensity, orientation of edges) that the LS-IKN model involves. Specifically, we employ the average local saliency into each segment as the weight  $\omega_i = lsal_i$ , where  $i \in \{bg, fg\}$  and :

$$lsal_i = \frac{1}{|\mathcal{R}_i|} \sum_j^{|\mathcal{R}_i|} LS_j. \quad (4.16)$$

**Local saliency and segment size** Here, the segment's size is taken into account together with the segment's saliency in the following way, similarly to (4.14).

$$\omega_i = lsal_i \cdot \frac{1}{siz_i} \quad (4.17)$$

where  $siz_i$  is the relative size of the background or foreground segments defined in (4.15). Figure 4.5 shows an example of a segmented frame and the weights assigned to the corresponding segments.

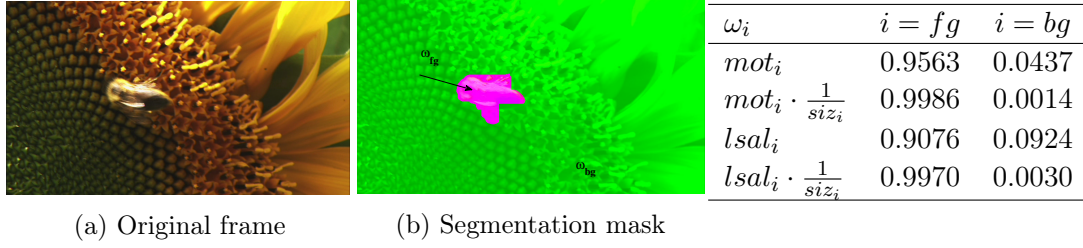


Figure 4.5: Example using method M1. The segmentation algorithm is applied on the original frame (a), and the foreground and background segments (b),  $\mathcal{R}_{fg}$  and  $\mathcal{R}_{bg}$  respectively, are weighted towards a segmentation-aware quality model design. *Sun flower* sequence, frame 60, LIVE video database.

### 4.3 Method M2: motion saliency for VQA Improvement

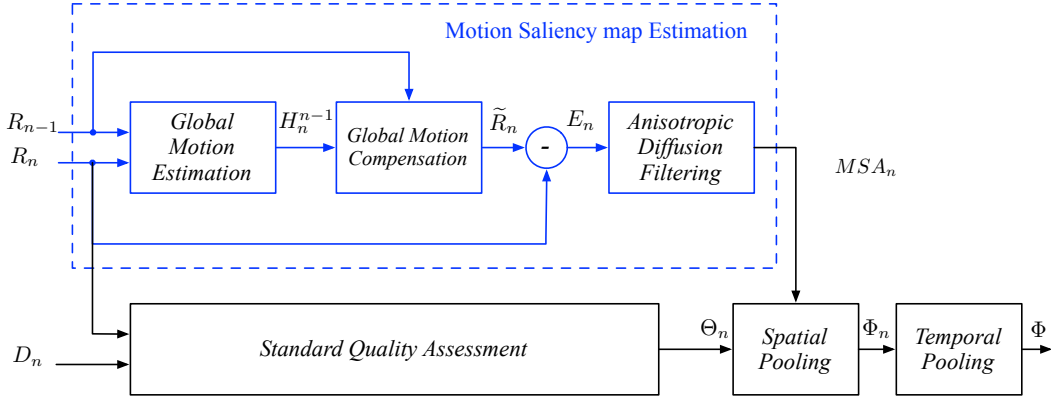
The main idea of the proposed method M2 is to detect regions that contain significant relative motion between frames, and emphasize their effect to the image quality index in the spatial pooling stage that will follow. This means that if a distortion occurs in a region that contains motion, it is expected to attract the attention of the viewer and to have thus negative impact on the quality assessment in comparison to a distortion that occurs in a region not containing motion. In contrast to M1, here the foreground and background segment regions assumed not to be known.

Under the assumption that the background (i.e. camera) motion is the dominant motion between two frames of a video sequence, the foreground motion is likely to attract visual attention, according to the properties of the HVS that are explored with respect to this point of view in [130]. Based on this observation we propose the following strategy, which is illustrated in Figure 4.6 and is subsequently described.

At first stage, the motion model  $H_n^{n-1}$  between two successive frames of the reference sequence  $R_{n-1}$  and  $R_n$  is computed. Based on  $H_n^{n-1}$  and  $R_{n-1}$ , the estimated frame  $\tilde{R}_n$  is computed and subsequently subtracted from  $R_n$ . This results in the global motion compensated absolute error frame  $E_n$  where high error energy indicates motion of the foreground area.  $E_n$  is subsequently filtered using anisotropic diffusion resulting in the  $MSA_n$  map that assigns a weight to each pixel location. In the spatial pooling step the standard quality assessment measure,  $MSA_n$  is used as a significance map and is combined with  $\Theta_n$  yielding the local motion saliency-aware model  $\Phi_n$ . Finally, the local quality metrics are combined in the temporal pooling stage to result in the overall quality measure  $\Phi$ .

#### 4.3.1 Motion saliency model

The eight-parameter perspective motion model is used at the first stage to describe the background motion between two successive frames of the reference sequence  $R_{n-1}$  and  $R_n$ . This is realised using a well-known feature-based global motion estimation approach [33] which includes detection of feature points, the computation of correspondences between two sets of features for successive frames, and finally

Figure 4.6: Method  $M2$  framework overview.

the estimation of motion model ( $H_n^{n-1}$ ) parameters. Considering that these feature correspondences represent motion between this pair of images, the global motion is estimated. Following the algorithms for feature detection and feature-based global motion estimation are described, before proceeding to the presentation of the proposed motion saliency estimation.

**Feature correspondences** The features correspondences are established between feature points that are detected in a pair of images, a procedure referred to as feature tracking. Here, we employ the well known *Kanade - Lucas - Tomasi* (KLT) tracking algorithm [140] which is overviewed in the following. It is a window-based approach that minimizes the squared error differences between the current and a reference window. The displacement vector  $\mathbf{d}$  of a feature point  $\mathbf{x}$ , between frames  $I$  and  $J$ , is assumed to be small and it can therefore be approximated by a translational motion model. Assuming image noise  $n(\mathbf{x})$ , the relationship of two corresponding features is defined as:

$$J(x) = I(\mathbf{x} - \mathbf{d}) + n(\mathbf{x}). \quad (4.18)$$

For small motions,  $I(\mathbf{x} - \mathbf{d})$  can be approximated by a first degree Taylor series with two dimensional gradient vector  $\mathbf{g}$ :

$$I(\mathbf{x} - \mathbf{d}) = I(\mathbf{x}) - \mathbf{g} \cdot \mathbf{d}. \quad (4.19)$$

The optimal displacement vector  $\mathbf{d}$  is found by minimizing the mean squared error  $\varepsilon$  between a search window  $\mathcal{W}$  around the feature in the two frames

$$\varepsilon = \int_{\mathcal{W}} [I(\mathbf{x} - \mathbf{d}) - J(\mathbf{x})]^2 d\mathbf{x} = \int_{\mathcal{W}} (h - \mathbf{g} \cdot \mathbf{d})^2 d\mathbf{x} \quad (4.20)$$

where  $h = I(\mathbf{x}) - J(\mathbf{x})$ . By interpreting the error function as  $\varepsilon(\mathbf{d})$ , its minimum is given by differentiation with respect to  $\mathbf{d}$  and setting the result equal to zero.

$$\int_{\mathcal{W}} (h - \mathbf{g} \cdot \mathbf{d}) \mathbf{g} dA = 0 \quad (4.21)$$

<u>Objective</u>	Robust estimation of the optimal model $Model_{opt}$ from $N$ feature pairs.
<u>Algorithm</u>	<ul style="list-style-type: none"> <li>(i) Choose randomly a minimal number of feature pairs and fit the model to this sample.</li> <li>(ii) Use threshold <math>\tau</math> to determine the number of inliers (consensus set) <math>S_i</math> amongst all feature pairs that have a distance greater than threshold <math>\tau</math>.</li> <li>(iii) If <math>S_i &gt; S_{max}</math>, then <math>Model_{opt} = Model_i</math> and <math>S_{max} = S_i</math>.</li> <li>(iv) Repeat (ii), (iii) until <math>M</math> subsample-based estimations have been performed.</li> <li>(v) Reestimate <math>Model_{opt}</math> using all inliers of the consensus set, in order to improve the estimated model.</li> </ul>

Table 4.1: RANSAC approach description. The minimal amount of samples needed  $M$ , i.e. the number of iterations to be performed, is given by (3.6).

since  $\mathbf{d}$  is assumed constant within  $\mathcal{W}$  and  $(\mathbf{g} \cdot \mathbf{d})\mathbf{g} = (\mathbf{g}\mathbf{g}^T)\mathbf{d}$ , it yields

$$\left( \int_{\mathcal{W}} \mathbf{g}\mathbf{g}^T dA \right) \mathbf{d} = \int_{\mathcal{W}} h\mathbf{g} dA \quad (4.22)$$

which is a system of two unknowns in two scalar equations.

**Global motion model estimation** Based on the detected features, we use the *random sample consensus* (RANSAC) [141] approach for fast and accurate model estimation. RANSAC is a non-deterministic approach that estimates model parameters from a set of points containing inliers and outliers. An overview of the RANSAC algorithm is provided in Table 4.1. Given a pair of images, the employment of features to estimate global motion has considerable advantages over the use of motion vectors for this scope. Mainly due to the fact that the resolution of motion vectors is usually limited to half or quarter pixel, whereas feature based global motion estimation has the advantage that features can be tracked at much higher resolution than quarter pixel, thus enabling better performance. Moreover if features are selected properly, the estimation is based only on these specific features and avoids using motion vectors that correspond to the entire images.

**Motion saliency model** Based on the connection between motion and perception and considering that moving areas will likely attract the viewer's attention the main idea is to exploit them for video quality estimation. Furthermore, studies on the human visual system have shown that the human retina is highly space variant in processing and sampling of visual information [6]. The accuracy is highest in the central point of focus, the fovea, and the peripheral visual field is perceived

with lower accuracy. Therefore, we consider the locations of the highest motion compensated error energy as the central points of focus, and to address the gradually decreasing focus, the error maps are low-pass filtered resulting in the motion saliency map:

$$MSA(x, y, n) = \alpha * |\hat{R}(x, y, n) - R(x, y, n)| \quad (4.23)$$

where  $x, y$  are the pixel coordinates in the horizontal and vertical direction, and  $n$  is the frame number. Anisotropic diffusion filtering ( $\alpha$ ), which is already described in section 3.3.1 is interestingly related to the neural dynamics of brightness perception [88, 142] that the anisotropic diffusion equation in equation (3.10) presents. Anisotropic diffusion filtering [87] offers a non-linear and space-variant filtering of the error frame, that while having a low pass character preserves the edges of the image. In this way higher weighting can be assigned to regions that have moved between two successive frames and we expect that they are more likely to attract visual attention in comparison to other areas that have not moved (or have moved with the background). As shown in Figure 4.7 the proposed motion saliency estimation approach can detect the motion of the foreground as depicted in the MSA map as brighter areas. Of course motion is not the only feature that attracts visual attention. Other features such as contrast, colour and structural information will be considered implicitly through the incorporation in standard objective metrics that will be described in the following.

### 4.3.2 Spatial pooling

Conventional image quality metrics generate a quality index  $\Theta$  between a reference and a distorted image ( $R$  and  $D$  respectively) and then consider that every pixel contributes equally to the overall image metric by averaging over all pixel locations. For reasons that have been already discussed, towards avoiding uniform spatial pooling, we employ a weighted pooling strategy where the estimated motion saliency maps are incorporated in conventional image quality metrics in frame level. The weighted mean for single scale metrics in  $(x, y)$  location of the  $n$ -th frame [130] is formulated as:

$$\Phi(x, y, n) = \frac{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} w(x, y, n) \cdot \Theta(x, y, n)}{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} w(x, y, n)} \quad (4.24)$$

where  $\Phi$  is the weighted metric and  $N_x, N_y$  are the frame dimensions. The proposed motion saliency maps  $MSA(x, y, n)$  are used as weighting maps  $w(x, y, n)$ . For multiscale models, that use  $M$  scales, the weighting map is scaled correspondingly and



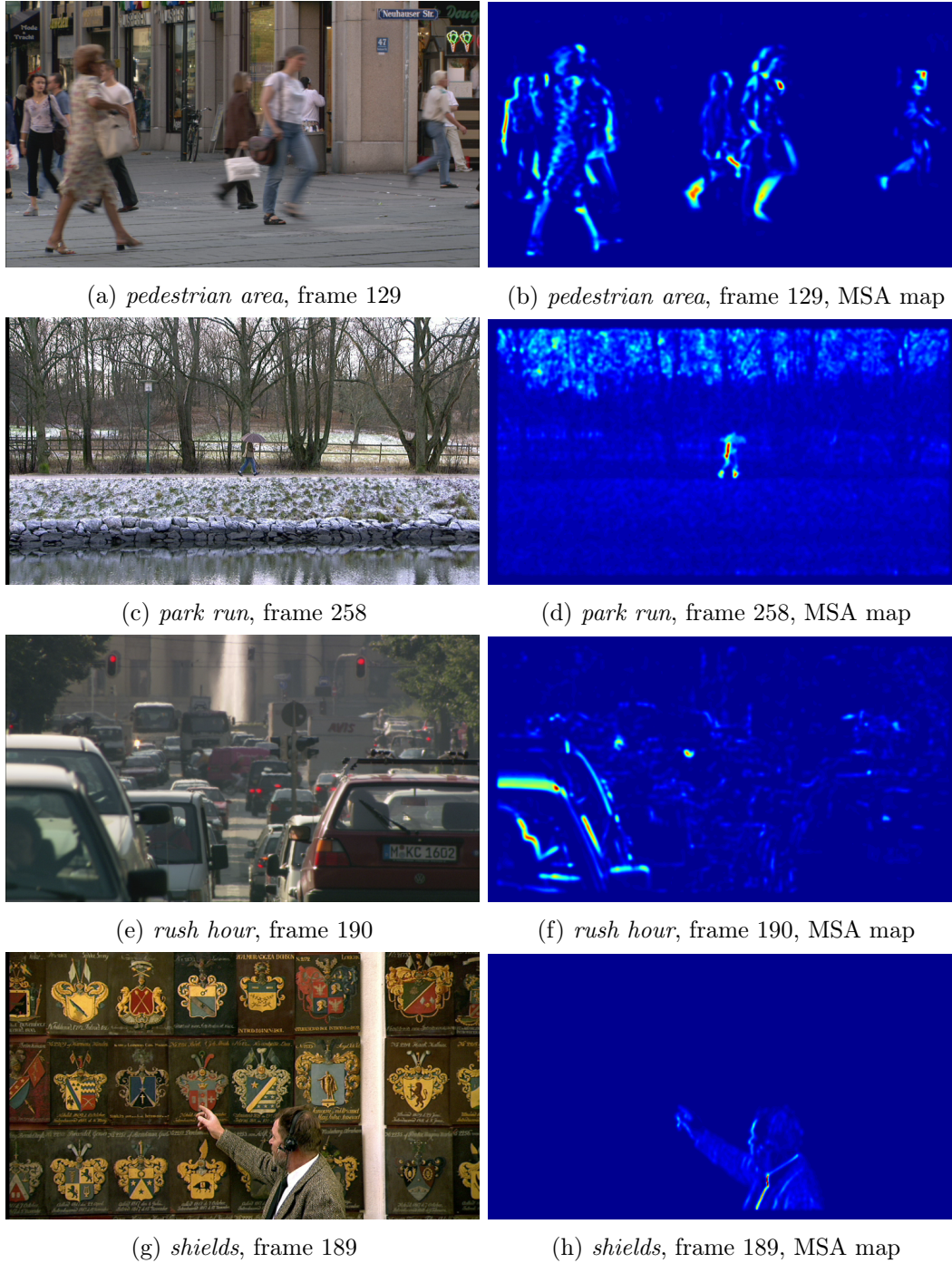


Figure 4.7: The first column depicts example reference frames  $R_n$  of the LIVE video database. The second column depicts the corresponding motion saliency maps  $MSA_n$  as heat maps, where warmer regions indicate higher motion saliency.



the overall metric is calculated as follows:

$$\Phi(x, y, n) = \prod_{j=1}^M \frac{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} w(x, y, n) \Theta(x, y, n)}{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} w(x, y, n)} \quad (4.25)$$

It is noted that the weighting map in method M2 assigns a weight to each pixel and not to regions of pixels.

We use several motion saliency models in order to compare the performance of the proposed motion saliency model in the experimental evaluation of method M2. These models are listed below.

**Motion saliency model (MSA)** The proposed saliency algorithm as presented in section 4.3.1.

**Local saliency Itti-Koch-Niebur (LS)** The LS-IKN model [118] as already discussed in section 4.2 is an indicator of local saliency, incorporating modalities such as colour, intensity and orientation of edges.

**Visual saliency model (VS)** The visual saliency estimation algorithm proposed by Ma *et al.* [128] is based on a Quaternion representation for each frame which incorporates spatial and temporal information derived from motion vectors. The Quaternion Fourier transform (QFT) uses hypercomplex numbers, namely Quaternions, to represent colour images. Moreover the VS model based on the QFT incorporates the three colour channels in a holistic manner, and not for each colour channel separately. In [128] the authors employ four image features: the luminance, the vertical motion vector component, the horizontal motion vector component and the corresponding motion prediction error to construct the Quaternion representation. This representation is subsequently filtered using Gaussian filtering to construct the proposed visual saliency maps which are incorporated in several video quality metrics for improved quality assessment.

**Short term moving object segmentation masks (STMOS)** In this case we employ the segmentation maps produced using the bidirectional motion-based object segmentation algorithm using hysteresis proposed in chapter 3 that is overviewed in Figure 3.1. In this way pixels in the background are assigned zero weight, whereas pixels belonging in the foreground are assigned  $w = 1$ . Thus, only foreground pixels participate in the spatial pooling stage.

**Filtered short term moving object segmentation masks (fSTMOS)** To avoid radically discarding pixels belonging to the background, as described in the above mentioned case, we create a map that lies in the range of  $[0, 1]$  and avoids

using "hard" decision weights  $\{0,1\}$ . The centre of the foreground regions is assigned with the highest scale, and the values continuously decrease while moving towards the foreground boundaries. In the background region, moving away from the foreground boundaries results in continuous decrease of the weights. To achieve this, the segmentation mask is Gaussian filtered with  $\sigma = 20$ .

**LS - STMOS** In this case short term moving object segmentation masks (STMOS) are used in combination with the local saliency (LS) Itti-Koch-Niebur model. In this case the segmentation masks are combined in frame level with the local saliency maps using per element multiplication. This is also driven by the motivation to avoid the "hard" decision resulting from the binary segmentation maps and enhance it using the well known local saliency model proposed in [118]. Pixels in the background regions here are zero weighted, in contrast to the fSTMOS case.

### 4.3.3 Temporal pooling

After the local weighted quality scores of every frame  $\Phi_n$  are computed, temporal pooling follows where the local scores are considered over the  $T_L$  frames of the sequence to yield the overall quality score  $\Phi$ .

The proposed temporal pooling approach is based on the variation of motion characteristics over time. Assuming that the perception of distortions in the temporal dimension is affected by several factors, such as the velocity of the camera, the velocity of the moving objects and the object size, the temporal pooling approach should take into account such factors. This can be done by weighting the frame-level (local) quality scores across time. In the general case the weighted average temporal pooling is given by:

$$\Phi_{wa} = \frac{\sum_{n=1}^{T_L} \omega_n \cdot \Phi_n}{\sum_{n=1}^{T_L} \omega_n} \quad (4.26)$$

where  $\omega_n$  is the weight assigned in the quality score of the  $n$ -th frame  $\Phi_n$ . In the special case of  $\omega_n = 1$ , for  $n = 1, 2, \dots, T_L$ ,  $\sum_{n=1}^{T_L} \omega_i = T_L$  and (4.26) gives the direct average which is the typical method to perform temporal pooling.

**Proposed approach** In order to account for the perceived quality degradation due to global motion, we propose the following approach. As observed in the description of the basic transformations involved in the perspective model (described in section 2.1.1) certain parameters are closely related with specific transformations exclusively. The parameter  $h_1$  reveals rotation and/or scaling,  $h_2$  and  $h_5$  indicate translation in the horizontal and vertical direction respectively,  $h_3$  corresponds to rotation, while the rest of the parameters ( $h_0, h_4, h_6, h_7$ ) are related with more than one basic transformations. Figure 4.8(a) shows the estimated global motion parameters for a test sequence over the entire length.

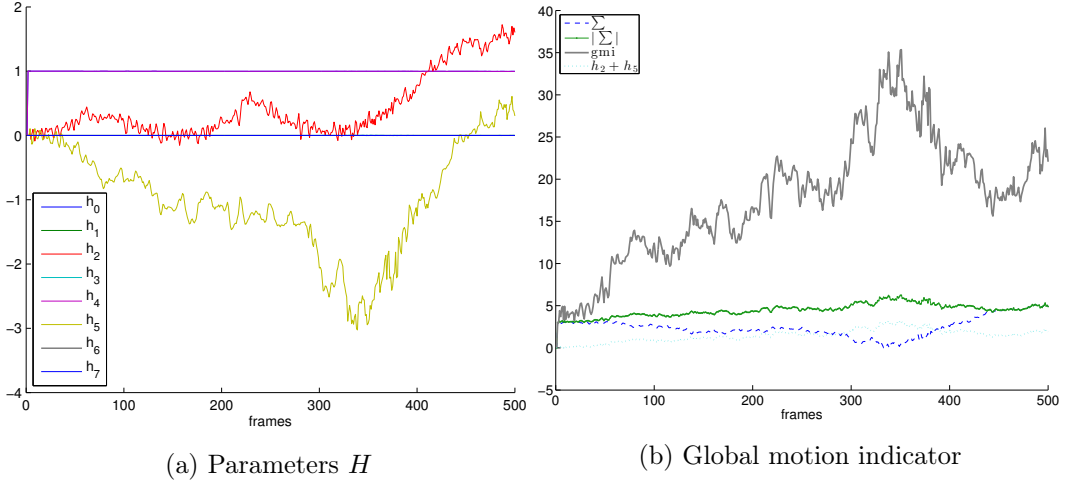


Figure 4.8: Global motion parameters and the proposed global motion indicator over frames for the *mobile calendar* sequence, LIVE database.

We make the hypotheses that a) the perception of distortions is affected more by translational motion and b) the faster the camera moves the greater the impact of distortions on perceived quality. Thus, we define the *global motion indicator* (gmi) as:

$$gmi(n) = \mathbf{F} \cdot (h_0 \ h_1 \ h_2 \ h_3 \ h_4 \ h_5 \ h_6 \ h_7)^T \quad (4.27)$$

where  $h_k$ ,  $k = 0, \dots, 7$  denote the elements of the eight-parameter homography of the  $n$ -th frame derived from global motion estimation using RANSAC as described in section 4.3.1 and  $\mathbf{F}$  is the enhancement matrix defined as:

$$\mathbf{F} = (1 \ 1 \ f \ 1 \ 1 \ f \ 1 \ 1). \quad (4.28)$$

Considering the behaviour of global motion coordinates in Figure 4.8(a), we set  $f = 10$ . The *gmi* represents thus the attributes of global motion for the scope of weighted temporal pooling here. Figure 4.8(b) depicts the *gmi* over all frames along with the case where  $f = 1$  (denoted as  $\sum$ ), the absolute value of it (denoted as  $|\sum|$ ), as well as the case of summation of only the translational parameters in the horizontal and vertical direction ( $h_2 + h_5$ ) for comparison.

Further, we describe the Minkowski summation pooling and a temporal pooling approach based on the temporal pooling function that are commonly adopted in video quality assessment approaches and will be used in the experimental evaluation section for comparison with the proposed approach.

**Minkowski temporal pooling** The temporal pooling strategy using the Minkowski summation [132] has been widely used in quality metrics design. It is described as follows:

$$\Phi_{mink} = \left[ \frac{1}{T_L} \sum_{n=1}^{T_L} [\Phi(n)]^\beta \right]^{\frac{1}{\beta}}. \quad (4.29)$$

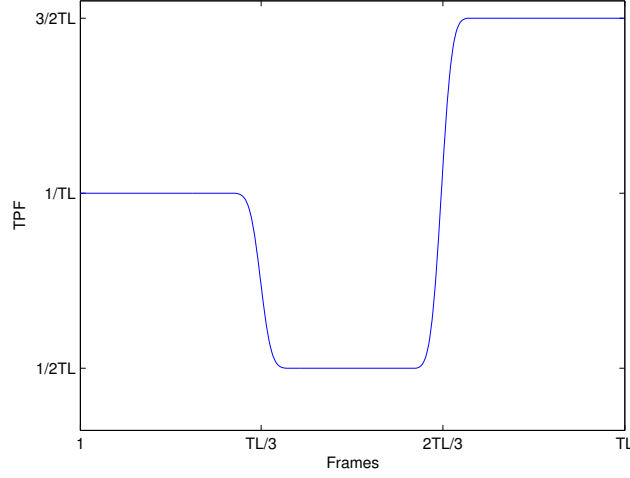


Figure 4.9: The temporal pooling function proposed in [137].  $TL$  denotes the sequence length (in frames).

For  $\beta = 1$ , it equals direct average. The exponent  $\beta$  is often suggested to yield good results when set to 2. Often, a few high distortions may draw the viewer's attention more than many lower ones [4]. This behavior can be emphasized well with Minkowski temporal pooling and especially as the exponent  $\beta$  increases.

**Temporal pooling function** Existing studies have shown that frames at the beginning and at the end of a video sequence have greater impact on the overall perceived quality. This is also reflected to the tendency to perceive better quality when increasing quality (over time) is observed. The temporal pooling scheme proposed by You *et al.* [137] is using the *temporal pooling function* (TPF) to estimate the overall quality score as follows:

$$\Phi_{tpf} = \left(1 + \frac{1}{TV}\right) \cdot \sum_{n=1}^{T_L} [\Phi(n) \cdot \text{TPF}(n)] \quad (4.30)$$

where  $TV$  is the total variation of the spatial quality metric over all frames. The  $TPF$ , illustrated in Figure 4.9, denotes the filtered version of the  $P(n)$  function defined in (4.31), after Gaussian filtering.

$$P(n) = \begin{cases} \frac{1}{T_L} & \text{for } n \leq \frac{T_L}{3} \\ \frac{1}{2T_L} & \text{for } \frac{T_L}{3} < n \leq \frac{2T_L}{3} \\ \frac{3}{2T_L} & \text{for } n \geq \frac{2T_L}{3} \end{cases} \quad (4.31)$$

where  $n$  is the frame index and  $T_L$  the length of the video sequence.

In the following we compare the proposed temporal pooling approach based on global motion indicator to the standard direct average temporal pooling approach,

the Minkowski summation pooling and a temporal pooling approach based on the temporal pooling function.

## 4.4 Experimental evaluation

**Test dataset** For the performance evaluation of the proposed approaches and towards reproducible research, we employ the *LIVE video quality database* [143] which is publicly available. This database is provided by the laboratory for image and video engineering (LIVE) of the University of Texas at Austin, USA. It contains 150 distorted video sequences obtained from 10 uncompressed reference videos ( $768 \times 432$  pixels, 3206 frames totally) of natural scenes. The distorted videos have been created using four commonly encountered distortion types: MPEG-2 compression, H.264/AVC compression, simulated transmission of H.264/AVC compressed bitstreams through error-prone IP networks, and through error-prone wireless networks. Each video has been assessed by 38 human subjects in a single stimulus study with hidden reference removal, where the subjects scored the video quality on a continuous quality scale (single stimulus continuous procedure). The difference scores of a given subject are computed by subtracting the score assigned by the subject to the distorted video sequence from the score assigned by the same subject to the corresponding reference video sequence.

Briefly, the following post-processing of the subjective scores takes place. Let  $s_{ijk}$  and  $s_{ij_{ref}k}$  denote the scores assigned by subject  $i$  to distorted video  $j$  and the reference one  $j_{ref}$  respectively in session  $k = \{1, 2\}$ . The difference scores are computed per session as:

$$d_{ijk} = s_{ij_{ref}k} - s_{ijk} \quad (4.32)$$

and they are converted to z-scores per session [143] as:

$$z_{ijk} = \frac{d_{ijk} - \mu_{ik}}{\sigma_{ik}} \quad (4.33)$$

where:

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} d_{ijk} \quad (4.34)$$

$$\sigma_{ik} = \sqrt{\frac{1}{N_{ik} - 1} \sum_{j=1}^{N_{ik}} (d_{ijk} - \mu_{ik})^2} \quad (4.35)$$

and  $N_{ik}$  denotes the number of videos watched by subject  $i$  in session  $k$ . Statistically unreliable subjects, according to ITU-R [97] are excluded from the procedure, z-scores are rescaled to lie in  $[0, 100]$  and finally the *difference mean opinion score* (DMOS) of each video is computed as the mean of the rescaled standardised difference scores (z-scores) of the statistically reliable subjects.

$$DMOS_j = \frac{1}{M} \sum_{i=1}^M z'_{ij} \quad (4.36)$$

where the finally reliable participating subjects are  $M = 29$ .

**Mapping to predicted DMOS** Each objective quality metric may produce values in a different range compared to the subjective scores. To facilitate comparison of the various models in the final stage, non linear regression analysis is performed in order to map each objective video quality metric output ( $\Phi$ ) to the subjective rating (DMOS) scale. This is performed using a four-parameter, monotonic logistic function as suggested in [99]:

$$\Phi' = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-\left(\frac{\Phi - \beta_3}{|\beta_4|}\right)}} \quad (4.37)$$

where  $\Phi'$  and  $\Phi$  are the predicted (mapped) and initial metrics respectively. The optimal parameter vector  $B = \{\beta_1, \beta_2, \beta_3, \beta_4\}$  is found using nonlinear least square optimization. Specifically, minimizing the least square error between the vector **DMOS** that contains the subjective scores ( $DMOS_j, j = 1, 2, \dots, DB_L$ ) and the vector  $\Phi'$  that contains the fitted objective scores ( $\Phi'_j, j = 1, 2, \dots, DB_L$ ) for the whole test database with  $DB_L$  videos. For the initialisation of the parameter vector  $B$  we use:

$$B_0 = \left[ \frac{\max(\mathbf{DMOS})}{\sigma(\mathbf{DMOS})}, \frac{\min(\mathbf{DMOS})}{\sigma(\mathbf{DMOS})}, \mu(\Phi), 1 \right] \quad (4.38)$$

where  $\sigma$  stands for standard deviation and  $\mu(\Phi) = \frac{1}{DB_L} \sum_j^{DB_L} \Phi_j$ .

The logistic function in equation (4.37) aims at mapping the range of a quality metric onto the range of the subjective scores obtained from the subjective quality assessment experiment. This is useful because each metric may produce predictions in a different range compared to the subjective score, which may result in non meaningful metric representation. Furthermore, non linear regression serves to remove nonlinearities due to the subjective rating process. It is found that human observers tend to make more pronounced distinctions between two quality levels of lightly distorted images than between two highly distorted ones [7]. Thus, non linear regression using an appropriate logistic function takes account for this phenomenon.

Figure 4.10 illustrates examples of fitted objective scores  $\Phi'$  versus subjective  $DMOS$  along with the best fitting logistic function. Subsequently, for evaluation the fitted objective scores ( $\Phi'$ ) and the subjective ones ( $DMOS$ ) are used for the calculation of the prediction performance indicators (described in section 4.4.2).

#### 4.4.1 Distortion indicators through image quality models

The proposed methodologies  $M1$ ,  $M2$  as well as the proposed temporal pooling approach are incorporated in several quality assessment models, namely MSE, SSIM [107], MS-SSIM [111] and VIF [112]. These models were described in section 4.1.2 and are used in the experimental evaluation as following described.

**Mean square error** The squared error map provided by equation (4.1) serves as distortion indication map (quality index) in the case of MSE.

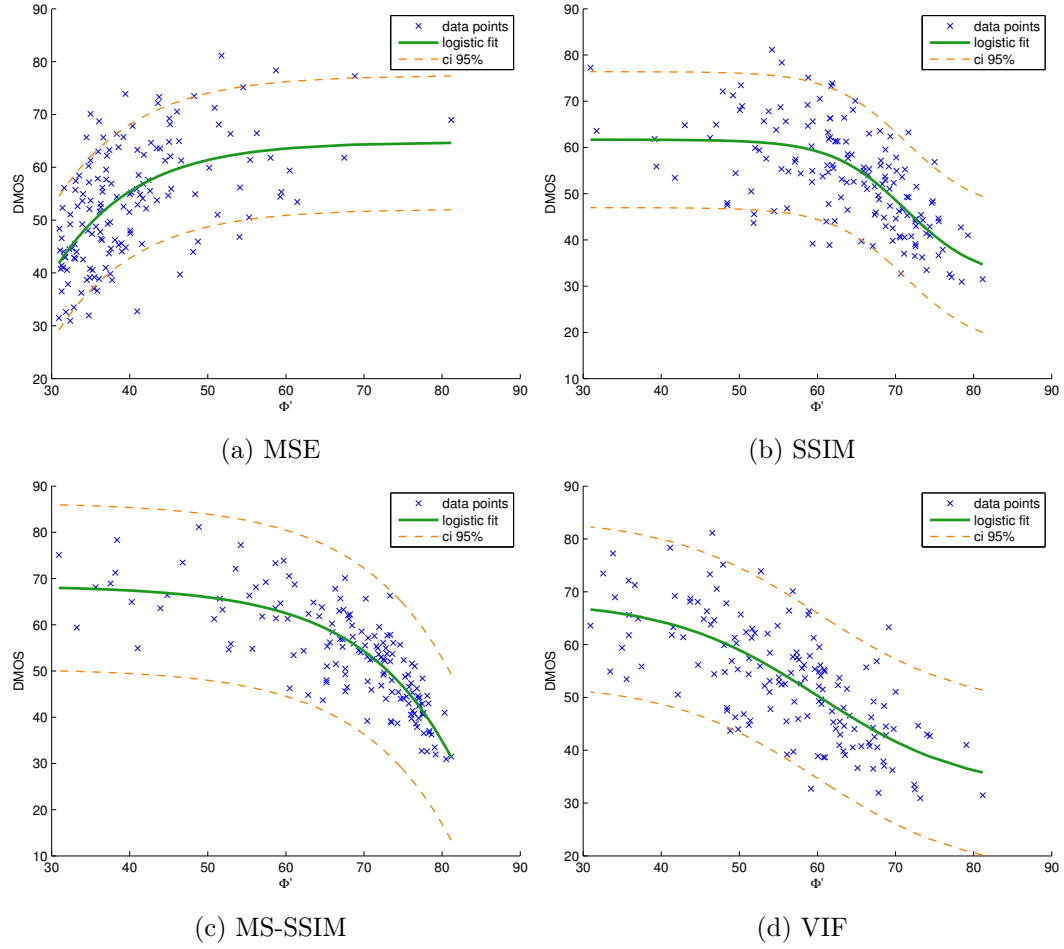


Figure 4.10: Scatter plots of subjective DMOS versus predicted objective quality scores  $\Phi'$  (MSE, SSIM, MS-SSIM and VIF) for the LIVE video quality database (blue cross marks) along with the best fitting logistic function (green line) and the corresponding 95% confidence interval (*ci*, orange dotted lines).

**Structural similarity index** The 2D SSIM index map is employed as distortion indication map (quality index) for SSIM.

**Multiscale SSIM** MS-SSIM incorporates SSIM evaluations in different scales. To this end, the single-scale SSIM index maps are weighted in each scale, and the weighted scaled indexes are subsequently combined as described in [111] using equation (4.25) to obtain the distortion indication map.

**Visual Information Fidelity criterion** In the case of VIF, the mutual information (between the input and the output of the HVS channel) for the reference image and the mutual information (between the input and the output of the HVS channel) for the distorted image are separately weighted using equation (4.24). They are subsequently scaled and finally combined over multiple scales, to finally output the distortion indication map.

#### 4.4.2 Prediction performance indicators

The quality prediction performance of the metrics is evaluated following the recommendation of the video quality experts group (VQEG) [99], which define the following prediction performance indicators:

- Prediction accuracy: the ability of an objective quality model to predict the subjective quality rating with low error.
- Prediction monotonicity: the degree to which the objective quality model maintains prediction accuracy over the range of video test sequences.

According to the VQEG recommendation [99], for  $K$  sequences, the prediction accuracy is determined using the Pearson linear correlation coefficient:

$$\rho_p = \frac{\sum_{k=1}^K (\phi_k - \bar{\phi})(s_k - \bar{s})}{\sqrt{\sum_{k=1}^K (\phi_k - \bar{\phi})^2} \sqrt{\sum_{k=1}^K (s_k - \bar{s})^2}} \quad (4.39)$$

where  $\phi_k, s_k$  are the predicted score and the subjective rating corresponding to the  $k$ -th sequence respectively, and  $\bar{\phi}, \bar{s}$  are the corresponding averages of each set.  $\rho_p$  is an indicator of the strength and the direction of the linear relationship between the pairs of predicted scores and subjective evaluations.

The prediction monotonicity is expressed using the Spearman rank order correlation coefficient:

$$\rho_s = \frac{\sum_{k=1}^K (\chi_k - \bar{\chi})(\gamma_k - \bar{\gamma})}{\sqrt{\sum_{k=1}^K (\chi_k - \bar{\chi})^2} \sqrt{\sum_{k=1}^K (\gamma_k - \bar{\gamma})^2}} \quad (4.40)$$



where  $\chi_k, \gamma_k$  denote the ranks of the predicted scores and the subjective scores respectively and  $\bar{\chi}, \bar{\gamma}$  correspond to the midranks of each of these sets. The Spearman correlation coefficient quantifies whether changes in one variable are followed by changes (increase or decrease) in the other variable, irrespective of the magnitude of the changes. In this way it is used as a prediction's monotonicity reflector.

Moreover, we employ the root mean square error:

$$RMSE = \sqrt{\frac{1}{K} \sum_{k=1}^K [DMOS_k - DMOS_{pk}]^2} \quad (4.41)$$

between  $DMOS$  and predicted  $DMOS_p$ .

Larger  $\rho_p$  and  $\rho_s$  indicate better correlation between objective and subjective scores, while smaller RMSE is indicator of better performance.

#### 4.4.3 Quality prediction performance of M1

In the  $M1$  scheme (presented in section 4.2), the first stage includes segmenting the frames of the video sequence, using our proposed motion segmentation approach. Figure 4.5 illustrated such an example. In the next stage, in contrast to the conventional objective quality assessment algorithms, where each image region participates equally in the determination of quality level, in  $M1$  each segment may have a different impact in the quality assessment procedure. This impact is defined in this work by the inter-segment interaction parameter, which is expressed as the weight  $\omega_i$ , where  $i = \{fg, bg\}$  used in equation (4.12) for the following cases:

- $\omega_i = lsal_i$  uses the local saliency model LS-IKN model [118] that incorporates colour, intensity and orientation features
- $\omega_i = mot_i$  uses the motion of each segment to determine inter-segment interactions
- $\omega_i = lsal_i \cdot \frac{1}{siz_i}$  accounts for the size of each segment together with its local saliency
- $\omega_i = mot_i \cdot \frac{1}{siz_i}$  accounts for the size of each segment together with its estimated motion.

Table 4.2 shows the results of method M1 using the above weights  $\omega_i$ . For each evaluation model we highlight the two best results with boldface. From the evaluation we observe that  $\omega_i = mot_i$  performs well in most cases. The incorporation of the segment's size does not improve the performance compared to  $\omega_i = mot_i$  or  $\omega_i = lsal_i$  respectively, except for the case of MS-SSIM where there is a relatively large improvement. We also observe that the motion feature as indicator of the inter-segment interactions, appears to be more effective in most cases compared to local saliency. Finally, the approaches show similar performance, with a small tendency of  $\omega_i = mot_i$  to perform better in this scheme.

Table 4.2: VQA metrics performance comparison of each case of foreground/background pooling used in method  $M1$ , described in section 4.2.2, on LIVE video database. The direct average approach has been used in this case for temporal pooling. The two best results are highlighted with boldface.

Algorithm	$\rho_p$	$\rho_s$	RMSE
MSE	<b>0.5614</b>	<b>0.5391</b>	<b>9.0839</b>
MSE - M1, $\omega_i = lsal_i$	0.5387	0.5281	9.2485
MSE - M1, $\omega_i = mot_i$	<b>0.5440</b>	<b>0.5371</b>	<b>9.2111</b>
MSE - M1, $\omega_i = lsal_i \cdot \frac{1}{siz_i}$	0.5334	0.5242	9.2856
MSE - M1, $\omega_i = mot_i \cdot \frac{1}{siz_i}$	0.5366	0.5298	9.2630
SSIM	0.5411	0.5231	9.2315
SSIM - M1, $\omega_i = lsal_i$	0.5887	0.5686	8.8733
SSIM - M1, $\omega_i = mot_i$	<b>0.5958</b>	<b>0.5744</b>	<b>8.8159</b>
SSIM - M1, $\omega_i = lsal_i \cdot \frac{1}{siz_i}$	0.5925	0.5674	8.8429
SSIM - M1, $\omega_i = mot_i \cdot \frac{1}{siz_i}$	<b>0.5942</b>	<b>0.5698</b>	<b>8.8295</b>
MS-SSIM	0.7556	0.7474	7.1911
MS-SSIM - M1, $\omega_i = lsal_i$	0.7029	0.6911	7.8077
MS-SSIM - M1, $\omega_i = mot_i$	0.7047	0.6944	7.7882
MS-SSIM - M1, $\omega_i = lsal_i \cdot \frac{1}{siz_i}$	<b>0.7800</b>	<b>0.7733</b>	<b>6.8687</b>
MS-SSIM - M1, $\omega_i = mot_i \cdot \frac{1}{siz_i}$	<b>0.7760</b>	<b>0.7690</b>	<b>6.9236</b>
VIF	0.5322	<b>0.5297</b>	9.2936
VIF - M1, $\omega_i = lsal_i$	<b>0.5463</b>	<b>0.5440</b>	<b>9.1943</b>
VIF - M1, $\omega_i = mot_i$	0.5245	0.5177	9.3458
VIF - M1, $\omega_i = lsal_i \cdot \frac{1}{siz_i}$	<b>0.5336</b>	0.5202	<b>9.2838</b>
VIF - M1, $\omega_i = mot_i \cdot \frac{1}{siz_i}$	0.5109	0.5007	9.4362

#### 4.4.4 Quality prediction performance of spatial pooling in M2

Method M2, specifically spatial pooling, has been described in section 4.3.2. In contrast to *M1* no segmentation map is available and weighted spatial pooling is performed in each pixel location. The proposed approach for determination of weight is motion saliency (MSA). Pixel positions that present significant relative motion between frames are assigned heavier weighing, since we expect that if a distortion occurs in a region that contains obvious motion, it is expected to attract the attention of the viewer and thus to have more significant impact on the quality assessment. In this way higher weighting can be assigned to regions that have moved between two successive frames and we expect that they are more likely to attract visual attention in comparison to other areas that have not moved (or have not moved in relation to the background). We compare the following cases where:

- **LS-IKN** denotes the local saliency Itti-Koch-Niebur model [118]
- **VS** denotes the visual saliency model proposed by Ma *et al.* [128]
- **STMOS** denotes the short term moving object segmentation masks
- **fSTMOS** denotes the filtered short term moving object segmentation masks
- **MSA** is the proposed motion saliency model described in section 4.3.1
- **LS-STMOS** denotes the short term moving object segmentation masks (STMOS) incorporated with the local saliency Itti-Koch-Niebur model.

Figures 4.11 - 4.12 present the significance maps of the above mentioned cases. The original frame is shown in 4.11 (a) and the overlaid heat maps of the various weighting maps used for Method M2 are also illustrated. Figure 4.11 (b) illustrates the local saliency LS-IKN model map. Figure 4.11 (c) shows the STMOS mask where a binary decision is used to account for the foreground and background regions. The use of the filtered STMOS mask in Figure 4.11(d) avoids the "hard" boundaries and in this case the mask is smoothed along the foreground borders both in the direction towards the centre of the foreground object and towards outside it. In similar fashion, in the case of Figure 4.11(f) the mask is smoothed in the direction towards the centre of the foreground object, but in this case the background remains unaltered having zero impact as the segmentation mask defines. Figure 4.11(e) depicts the proposed motion saliency map that is not as solid in comparison to the LS-IKN model and also accounts for moving edges that are discarded in the case of STMOS masks.

Table 4.3 reports the performance evaluation of the examined cases of spatial pooling employed in method M2, using various objective VQA algorithms. MSA is the proposed method. VS denotes the visual saliency model proposed in [128], whereas local saliency denotes the employment of the IKN local saliency maps proposed in [118] for weighting in the same manner as described in the previous section. For each evaluation model we highlight the best results with boldface.

Table 4.3: VQA metrics performance comparison of each case of spatial pooling method  $M2$  on LIVE video database. The direct average approach is used in this case for temporal pooling. Data for VS is taken from [128].

Algorithm	$\rho_p$	$\rho_s$	RMSE
MSE	0.5614	0.5391	9.0839
MSE - M2 - LS-IKN [118]	0.5429	0.5262	9.2184
MSE - M2 - VS [128]	<b>0.6295</b>	<b>0.6268</b>	<b>8.5310</b>
MSE - M2 - STMOS	0.5175	0.5055	9.3931
MSE - M2 - fSTMOS	0.5128	0.5049	9.4238
MSE - M2 - MSA	0.5669	0.5593	9.0427
MSE - M2 - LS-STMOS	0.5222	0.5087	9.3619
SSIM	0.5411	0.5231	9.2315
SSIM - M2 - LS-IKN [118]	0.5995	0.5764	8.7855
SSIM - M2 - VS [128]	0.6308	0.6187	8.5310
SSIM - M2 - STMOS	0.5876	0.5602	8.8819
SSIM - M2 - fSTMOS	0.5802	0.5559	8.9406
SSIM - M2 - MSA	<b>0.6470</b>	<b>0.6334</b>	<b>8.3698</b>
SSIM - M2 - LS-STMOS	0.6142	0.5873	8.6630
MS-SSIM	0.7556	0.7474	7.1911
MS-SSIM - M2 - LS-IKN [118]	0.7597	0.7483	7.1382
MS-SSIM - M2 - VS [128]	0.7583	0.7468	7.1570
MS-SSIM - M2 - STMOS	0.7887	0.7819	6.7487
MS-SSIM - M2 - fSTMOS	0.7895	0.7813	6.7372
MS-SSIM - M2 - MSA	<b>0.8009</b>	<b>0.7964</b>	<b>6.5726</b>
MS-SSIM - M2 - LS-STMOS	0.7700	0.7620	7.0037
VIF	0.5322	0.5297	9.2936
VIF - M2 - LS-IKN [118]	0.6790	0.6687	8.0587
VIF - M2 - STMOS	0.6915	0.6863	7.9294
VIF - M2 - fSTMOS	0.6925	0.6885	7.9195
VIF - M2 - MSA	<b>0.6946</b>	<b>0.6959</b>	<b>7.8968</b>
VIF - M2 - LS-STMOS	0.6756	0.6637	8.0933

It is observed that (motion and local) saliency weighted models perform better compared to non-weighted models. Motion saliency spatial pooling proves to be more effective compared to local saliency pooling. The proposed motion saliency model in method M2 outperforms the local saliency model and also the recently proposed visual saliency approach. The VS approach, is based mainly on motion modelling whereas the LS-IKN model accounts mainly for colour and contrast variations. In the experimental evaluation it is suggested that motion is more beneficial compared to colour or contrast features for enhancing objective quality assessment algorithm, since the VS as well as the MSA model outperform the LS-IKN model. Furthermore, we studied the incorporation of the significance map as hard decision (*STMOS*, where the significance map is based on binary decision) as well as two cases where the weights are attenuating towards the inner object the boundaries. In the latter cases, *fSTMOS* and *LS – STMOS* it was assumed that the centres of the object regions have the strongest impact on the quality assessment procedure. Even though it was suspected that a hard decision (*STMOS*) wouldn't be more efficient than the soft decision cases (*fSTMOS* and *LS – STMOS*) it is observed that the performance of *fSTMOS* is comparable with the one of *STMOS*, which is also true for the case of *LS – STMOS*.

Figure 4.13 illustrates example scatter plots of predicted DMOS using standard and weighted objective metrics using the proposed method MSA for spatial pooling M2 (green and blue marks respectively) versus DMOS. It can be observed, especially for the case of MS-SSIM 4.13(c), that the proposed weighting method yields scores that are closer to the angle bisector of  $45^\circ$  compared to the standard metric. Confirming that the proposed approach is effective for video quality assessment.

### Study on each distortion class

To examine the effect of the proposed weighting on different distortion types, we present in Table 4.4 the performance improvement, in terms of Spearman rank order correlation coefficient, introduced by the proposed method for each distortion class separately. As expected, our proposed approach contributes on average more in cases of transient distortions (in the presence of packet losses, classes #1 and #2) compared to cases with uniformly distributed distortions (no packet losses, classes #3 and #4). The average improvement in terms of  $\rho_s$  for distortion classes #1 and #2 is 0.0881, whereas for classes #3 and #4 is 0.0784, whereas the overall trend of outperformance of motion saliency spatial pooling remains unchanged across the various distortion types.

Table 4.4: Performance improvement in terms of  $\rho_s$  of our proposed method using motion saliency (MSA) over standard metrics using the spatial pooling method M2 on LIVE database for each distortion class.

#	Distortion class	MSE	SSIM	MS-SSIM	VIF
1	H264 + wireless	-0.0291	0.1328	0.0638	0.1538
2	H264 + IP	0.1139	0.1166	0.0206	0.1326
	average (#1,#2)	0.0424	0.1247	0.0422	0.1432
3	H264	0.0251	0.1099	0.0901	0.1546
4	MPEG2	0.0238	0.1110	0.0662	0.0463
	average (#3,#4)	0.0245	0.1105	0.0782	0.1005
	All data	0.0202	0.1103	0.0490	0.1662

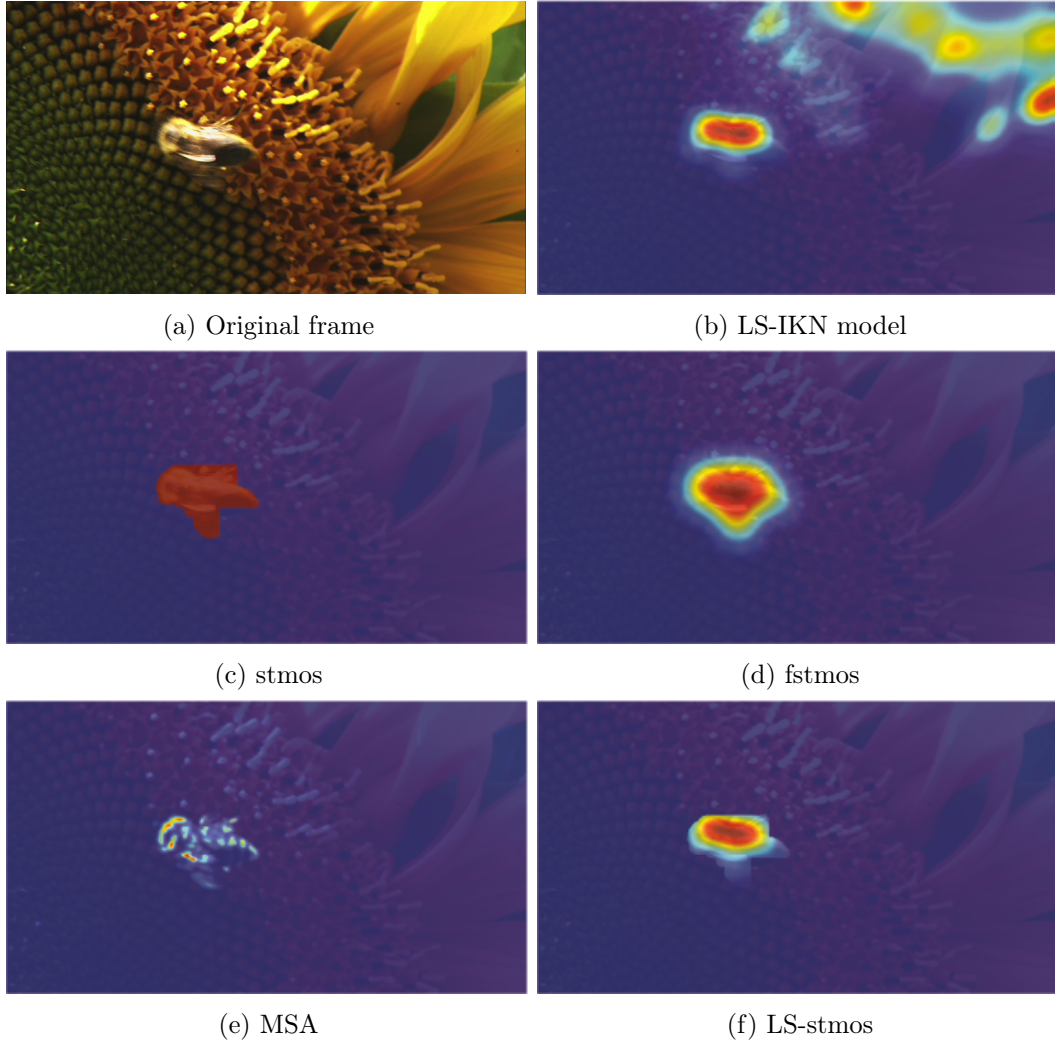


Figure 4.11: Original frame 60 of *sun flower* of the LIVE database and weighting maps, depicted as over imposed heat maps, used in Method M2. (a) Original frame, (b) local saliency IKN model map (LS-IKN), (c) short term moving object segmentation mask (STMOS), (d) filtered STMOS mask (fSTMOS), (e) proposed motion saliency map (MSA) and finally (f) combined local saliency IKN and STMOS (LS-STMOS) map.

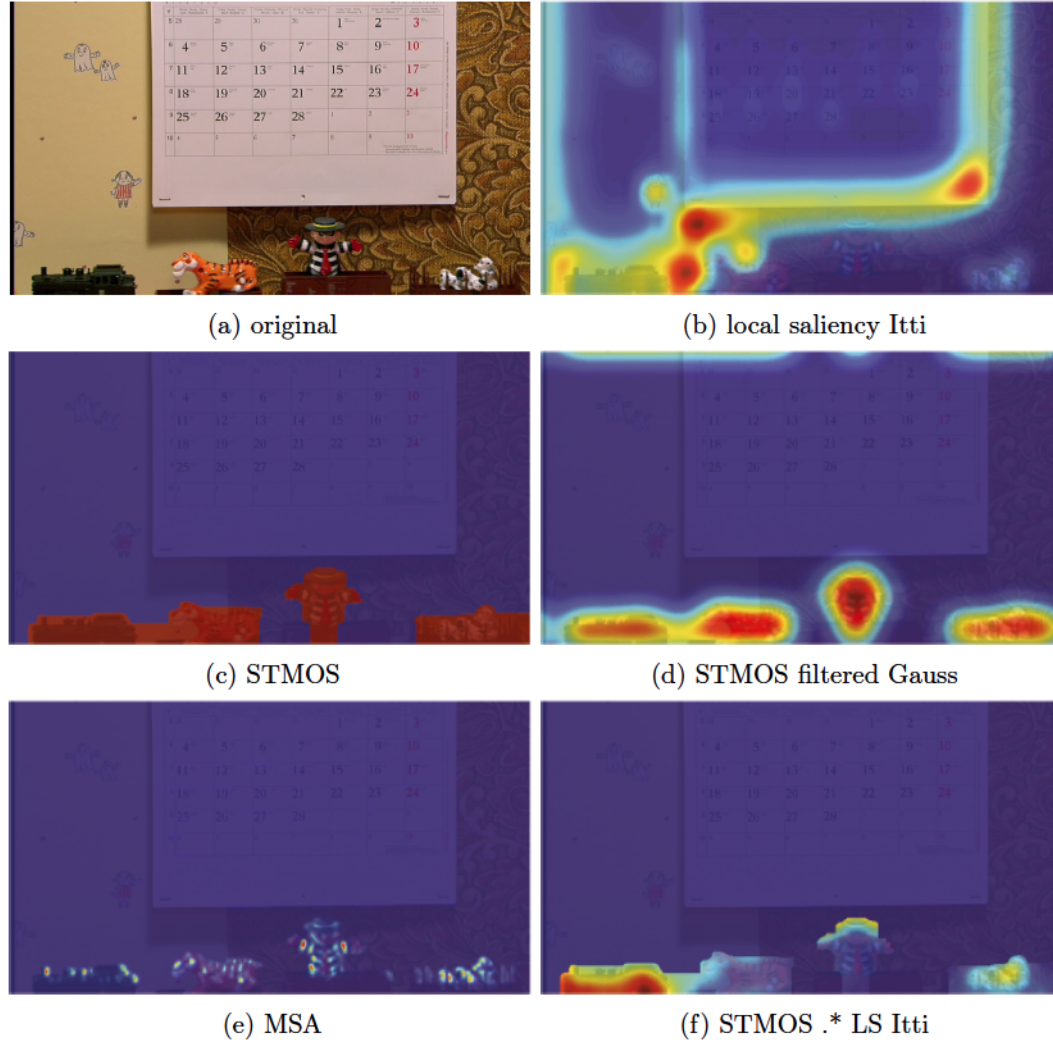


Figure 4.12: Original frame 433 of *mobile calendar* of the LIVE database and weighting maps, depicted as over imposed heat maps, used in Method M2. (a) Original frame, (b) local saliency IKN model map (LS-IKN), (c) short term moving object segmentation mask (STMOS), (d) filtered STMOS mask (fSTMOS), (e) proposed motion saliency map (MSA) and finally (f) combined local saliency IKN and STMOS (LS-STMOS) map.



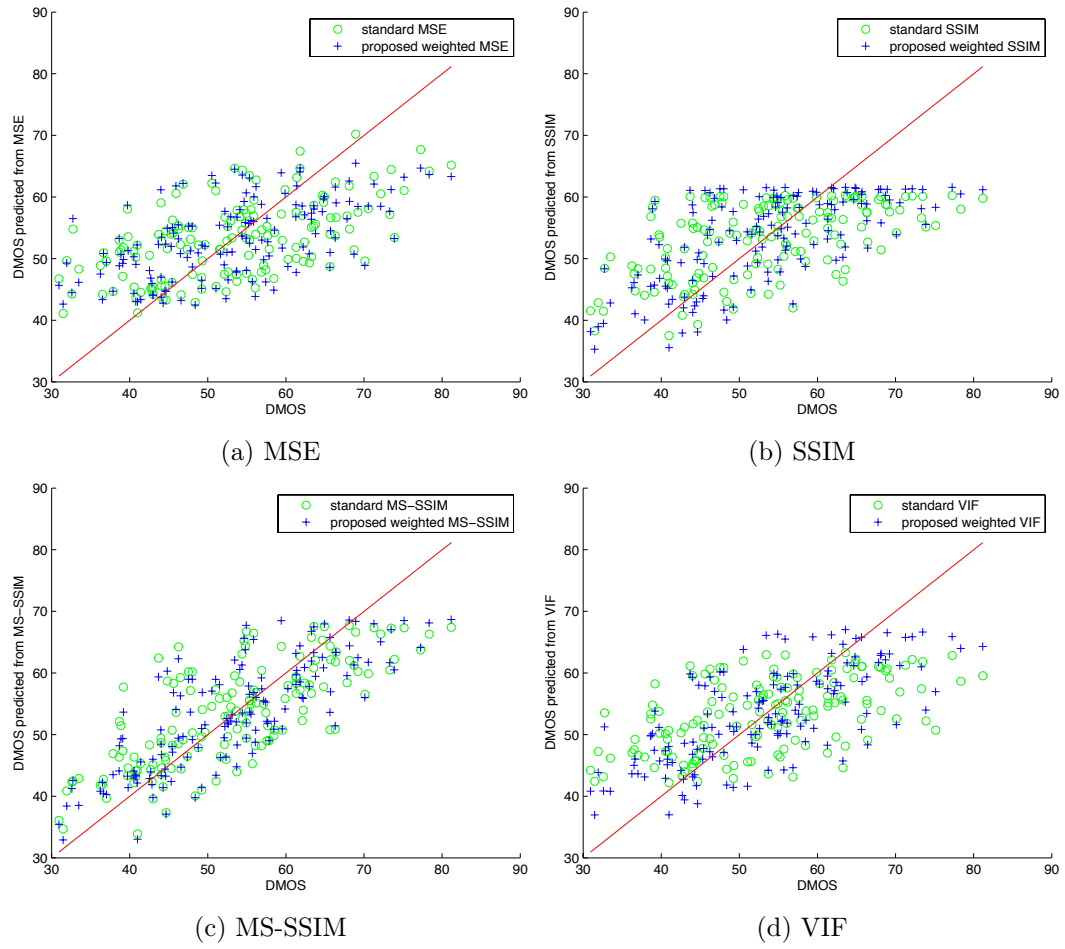


Figure 4.13: Scatter plots of DMOS versus predicted DMOS using standard and proposed method (blue and green marks respectively) on LIVE video database.

#### 4.4.5 Quality prediction performance of temporal pooling in M2

Temporal pooling follows the spatial pooling stage, as the local weighted quality scores  $T_L$  frames have to be taken into account to output the overall quality score. The proposed temporal pooling approach, weights the frame-level quality scores across time based on the variation of global motion on the temporal dimension, assuming that large camera motion causes distortions to have a greater impact on perceived video quality and that the perception of distortions is affected mostly by translational motion. The proposed temporal pooling approach as well the related approaches used for comparison have been described in section 4.3.3. Considering MSA as the spatial pooling strategy, and towards evaluating the proposed temporal pooling scheme, four temporal pooling approaches are examined:

- $\mathbf{T}_{da}$  is the widely used direct average approach.
- $\mathbf{T}_{gms}$  is the approach using the global motion indicator where  $f = 1$  in equation (4.28) to allow for uniform participation of all global motion parameters temporal pooling.
- $\mathbf{T}_{gmi}$  denotes the proposed method using the proposed global motion indicator with  $f = 10$  forcing the translational components of global motion to have a stronger influence on temporal pooling.
- $\mathbf{T}_{mink}$  denotes the widely used Minkowski summation, where we obtained the best results with the exponent  $\beta = 2$  and
- $\mathbf{T}_{tpf}$  the temporal pooling function is finally reported as proposed in [137].

Table 4.5 presents the results. The proposed temporal pooling approach  $T_{gmi}$  employing the global motion indicator performs better compared to the other approaches. It brings up to 0.0207 improvement in the case of SSIM in terms of Pearson correlation coefficient, compared to the direct average. The direct average approach  $T_{da}$  is in each case outperformed, whereas the performance of  $T_{mink}$ ,  $T_{tpf}$  and  $T_{pa}$  depends on the objective metric.

$T_{gmi}$  and  $T_{gms}$  are the two best performing approaches, which suggests that taking into consideration the estimated global motion brings improvement in the performance of the quality metrics. The global motion indicator expresses the proportional relation between the existing global motion and the temporal weight. This finding is in contrast to [110] where it was suggested that the larger the existing global motion is, the smaller the assigned temporal weight should be. In our point of view, it seems to be more reasonable distortions to be more profoundly perceived in cases of large global motion, but this may be also be dependent on the kind of distortions. For instance blurring may be more noticeable in static or slowly moving scenes compared to fast moving scenes, however, blockiness may be more noticeable in cases of large camera movement compared to cases with slow camera movement.

Table 4.5: Temporal pooling performance comparison. The spatial pooling method *M2* has been used. Results on the LIVE video database.

Algorithm	$\rho_p$	$\rho_s$	RMSE
MSE - M2 - MSA - $T_{da}$	0.5669	0.5593	9.0427
MSE - M2 - MSA - $T_{gms}$	0.5700	0.5643	9.0193
MSE - M2 - MSA - $T_{gmi}$	<b>0.5748</b>	<b>0.5676</b>	<b>8.9825</b>
MSE - M2 - MSA - $T_{mink}$	0.5488	0.5400	9.1766
MSE - M2 - MSA - $T_{tpf}$ [137]	0.5685	0.5609	9.0304
SSIM - M2 - MSA- $T_{da}$	0.6470	0.6333	8.3697
SSIM - M2 - MSA - $T_{gms}$	0.6558	0.6387	8.2876
SSIM - M2 - MSA - $T_{gmi}$	<b>0.6678</b>	<b>0.6420</b>	<b>8.1710</b>
SSIM - M2 - MSA - $T_{mink}$	0.6455	0.6302	8.3843
SSIM - M2 - MSA - $T_{tpf}$ [137]	0.6386	0.6217	8.4477
MS-SSIM - M2 - MSA- $T_{da}$	0.8010	0.7964	6.5724
MS-SSIM - M2 - MSA - $T_{gms}$	0.8087	0.8007	6.4568
MS-SSIM - M2 - MSA - $T_{gmi}$	<b>0.8155</b>	<b>0.8096</b>	<b>6.3527</b>
MS-SSIM - M2 - MSA - $T_{mink}$	0.8037	0.7977	6.5314
MS-SSIM - M2 - MSA - $T_{tpf}$ [137]	0.7892	0.7834	6.7410
VIF - M2 - MSA- $T_{da}$	0.6946	0.6958	7.8968
VIF - M2 - MSA - $T_{gms}$	0.6983	0.6972	7.8577
VIF - M2 - MSA - $T_{gmi}$	<b>0.7092</b>	<b>0.7121</b>	<b>7.7391</b>
VIF - M2 - MSA - $T_{mink}$	0.6808	0.6798	8.0408
VIF - M2 - MSA - $T_{tpf}$ [137]	0.6846	0.6801	8.0020

To conclude, our proposed approach based on the global motion indicator outperforms the direct average and commonly used related approaches that are commonly used, namely the Minkowski summation, the temporal pooling function. This indicates that more sophisticated approaches that account for the motion features over the temporal dimension, can be very beneficial for video quality assessment. It justifies thus our initial motivation that considering temporal dependencies between frames is a more suitable approach for assessing the quality of video sequences compared to the case of using image quality metrics. In the video quality assessment procedure, motion plays a critical role and by applying image quality metrics on frame level and subsequently fusing the local measures using direct average this important aspect is ignored.

#### 4.4.6 Comparison of method M1 and method M2

Comparing the performance of the two spatial pooling strategies Method M1 and Method M2, we observe the following. The incorporation of the moving object segmentation approach into the spatial pooling stage of VQA metrics, by means of foreground/background pooling in method M1 has shown that the benefits are not as consistent as expected. The improvement using a significance map in the spatial pooling stage, as in method M2, proved to be higher and more consistent. Several cases have been studied suggesting that accounting for motion saliency based on the proposed modelling is improving the correlation with subjective ratings. This relies on the fact that in method M2 the weights are assigned to pixel positions instead of regions (scheme adopted in method M1) which enables more detailed and accurate determination of salient areas. Moreover, this is also attributed to inaccuracies occurred in the moving object segmentation step, where filtering and thresholding may cause moving object regions to be misclassified, for instance mistakenly labeled as background or falsely either dilated or eroded.

Table 4.6 provides a comparative overview of the performance of Method 1, Method 2 and the conventional (content-unaware) metrics. M2 outperforms in each case method M1 as well as the corresponding conventional metrics. Introducing the  $T_{gmi}$  temporal pooling to method M2 improves the performance even more, and MS-SSIM-M2-MSA- $T_{gmi}$  presents the highest performance among the metrics.

Moreover the performance of the state-of-the-art video quality assessment models: MC-SSIM [136], VQM [113] and MOVIE index [114] are juxtaposed in Table 4.6 for comparison. The proposed method for the case of MSA-weighted MS-SSIM using the gmi (referred as MS-SSIM - M2 - MSA -  $T_{gmi}$ ) outperforms the state-of-the-art motion models, which confirms the validity and the encourages further perspectives of the proposed approach.

Table 4.6: Comparison of methods  $M1$  and  $M2$  on LIVE video quality database. MC-SSIM, VQM and MOVIE performance as reported in [136].

Algorithm	$\rho_p$	$\rho_s$
MSE	0.5614	0.5391
MSE - M1, $\omega_i = mot_i$	0.5440	0.5371
MSE - M2 - MSA	<b>0.5669</b>	<b>0.5593</b>
MSE - M2 - MSA - $T_{gmi}$	<b>0.5748</b>	<b>0.5676</b>
SSIM	0.5411	0.5231
SSIM - M1, $\omega_i = mot_i$	0.5958	0.5744
SSIM - M2 - MSA	<b>0.6470</b>	<b>0.6334</b>
SSIM - M2 - MSA - $T_{gmi}$	<b>0.6678</b>	<b>0.6420</b>
MS-SSIM	0.7556	0.7474
MS-SSIM - M1, $\omega_i = lsal_i \cdot \frac{1}{siz_i}$	0.7800	0.7733
MS-SSIM - M2 - MSA	<b>0.8009</b>	<b>0.7964</b>
MS-SSIM - M2 - MSA - $T_{gmi}$	<b>0.8155</b>	<b>0.8096</b>
VIF	0.5322	0.5297
VIF - M1, $\omega_i = lsal_i$	0.5463	0.5440
VIF - M2 - MSA	<b>0.6946</b>	<b>0.6959</b>
VIF - M2 - MSA - $T_{gmi}$	<b>0.7092</b>	<b>0.7121</b>
MC-SSIM [136]	0.6976	0.6791
VQM [113]	0.7236	0.7026
MOVIE [114]	0.8102	0.7861

## 4.5 Chapter summary

In this chapter we studied the problem of enhancing objective quality metrics' performance by improving their correlation with subjective quality scores. Particularly, in the first part of the chapter we employed our motion segmentation algorithm using bidirectional change detection and hysteresis towards enhancing objective video quality assessment metrics and also studied several possibilities regarding the incorporation segmentation features into a moving object-aware VQA scheme. Further, we proposed a novel motion saliency estimation method for video sequences considering motion between successive frames, and their corresponding parametric camera motion representation. This motion saliency model was incorporated in the spatial pooling stage of several objective video quality metrics and it has been shown that it outperforms existing state-of-the-art approaches. Finally, we have proposed a temporal pooling approach that enables further improvement of objective metrics by exploring global motion in the temporal dimension.

Experimental evaluation has shown that in the case of spatial pooling, (motion and local) saliency improves objective quality assessment models. Specifically, motion saliency has proved to be more effective compared to local saliency and outperforms existing motion saliency approaches. Therefore, it can be concluded that motion saliency is a powerful approach and motion is more powerful compared to colour or contrast features for improving objective quality assessment algorithms.

In video quality assessment, motion plays a critical role and by applying image quality assessment metrics on frame level and subsequently fusing the local measures using average this important aspect is ignored. With respect to temporal pooling, the proposed global motion indicator, which reflects a proportional relation between global motion and temporal weighting, outperformed existing related approaches. This indicates that sophisticated approaches that account for motion features over the temporal dimension, are very beneficial for video quality assessment. It justifies also our initial motivation that temporal dependencies between frames should be taken into consideration for assessing the quality of video sequences.

To conclude, it has been shown that the discrepancy between objective metrics and subjective evaluation is reduced, which is an indicator that objective quality models benefit from the proposed approaches. Having explored several aspects of incorporating motion and especially global motion into VQA, motion seems to be an important aspect that affects the perception of visual quality assessed by humans.

# Conclusion

---

## Contents

---

<b>5.1 Summary of the thesis</b>	<b>133</b>
<b>5.2 Discussion and outlook</b>	<b>135</b>

---

In the previous chapters novel work on analysis and processing of global motion in video sequences captured by a moving camera has been presented. The basic objectives have been to propose an improved global motion estimation algorithm, to extract moving object segmentation regions and finally to exploit this knowledge towards improving objective video quality assessment approaches. This chapter summarises the thesis, discusses approaches implemented within it and considers the general framework in which this thesis falls. Further, improvements and extensions that can be subject for future research are discussed and conclude the thesis.

## 5.1 Summary of the thesis

In the first part of this work a new approach for improving the parametric global motion estimation based on motion vectors and exploiting the variable sizes of the corresponding blocks has been proposed. In a contemporary video codec environment, typically a region which can be described by homogenous motion is assigned one motion vector in contrast to a heterogeneous region, described by multiple motions, that is assigned multiple motion vectors. The latter is translated in multiple smaller blocks that each one corresponds to a motion vector. In this way, motion diversities in an image determine the size and shape - which is often predefined - of the assigned blocks. In the case of block-based parametric global motion estimation approaches it is critical to distinguish and discard blocks belonging to foreground. Existing approaches do not consider block partitioning characteristics and thus valuable information that can be exploited for outlier rejection is often neglected. Towards addressing this, the block size variability was taken into account to select appropriate blocks for global motion estimation.

In the case of the binary partition tree, improvements in the performance of global motion estimation in terms of accurate background reconstruction were achieved by making an appropriate selection and controlling the influence of participating blocks in global motion estimation. Improvement is also reported in comparison to the case of fixed-size blocks. Subjective evaluation based on segmentation performance also reflected the benefits of the proposed approach. Furthermore, it has

been shown that using a parametric global motion model, improvements in conventional motion prediction are achieved. This is especially beneficial in the background region and in cases of no moving object's presence. A possible exploitation of this can be in compression schemes, since it may enable compression of the background region with less necessary information compared to conventional motion prediction using motion vectors.

Following the first part of the thesis, a motion-based object segmentation algorithm for video sequences with moving camera has been presented. The proposed algorithm exploits short-term motion information between frames towards change detection. It is based on bidirectional inter-frame change detection using a motion compensated error fusion scheme that outperforms previously proposed fusion schemes. In addition to that, spatial error localization is considered in the thresholding step for improving the segmentation efficiency. As the hysteresis thresholding introduces the requirement of two thresholds instead of one, a chi-squared test on results produced by different thresholding parameters has been used to select the appropriate weights, out of a given set of candidates. This enables robust segmentation performance that avoids the requirement of empirically defining the thresholding parameter and training algorithms that are commonly adopted for parameter selection. Furthermore, a final post-processing step has been incorporated to enable temporal consistency of the segmentation masks using filtering of the preliminary outcome, which is adapted according to the motion of the foreground. The experimental evaluation demonstrated the validity of the proposed approach in comparison with existing related approaches and it was additionally shown that its performance is quite stable in a codec framework, under varying quantisation parameters that influence motion estimation quality.

In the last part of this thesis, approaches for enhancing state-of-the-art objective video quality metrics' performance in terms of improving their consistency with subjective evaluations have been proposed. Towards this direction, the incorporation of the derived motion-related information into content-aware video quality assessment schemes was proposed, by exploring two different frameworks. Specifically, in method M1 the proposed motion segmentation algorithm using bidirectional change detection and hysteresis has been incorporated into a moving object-aware video quality assessment scheme. Several possibilities regarding the consideration of segmentation features into the moving object-aware video quality assessment scheme have been compared. The experimental evaluation showed that motion serves as a powerful indicator of the inter-segment interactions and is a valuable feature towards enhancing objective video quality assessment metrics.

Further, a novel motion saliency estimation method for video sequences considering motion between successive frames and their corresponding parametric camera motion representation has been proposed. The motion saliency model was incorporated in the spatial pooling stage of several objective video quality metrics and it was shown that it overperformed existing state-of-the-art approaches. Finally, a temporal pooling approach that enables further improvement of objective video quality assessment metrics has been proposed. This approach accounts for global motion



in the temporal dimension that is a neglected feature in common approaches on this topic, and it shown to outperform them. Experimental evaluation has shown that in this way objective metrics are more consistent with subjective evaluation scores, which confirms that our proposed approaches are beneficial for video quality assessment.

## 5.2 Discussion and outlook

It has been shown that exploiting characteristics of the block assignment is beneficial for parametric global motion estimation. For this, the size of the blocks has been mainly used in this work. An interesting direction for future study is to incorporate edge and texture information in the weighting of participating blocks in global motion estimation. Given that large homogenous blocks may contain misleading motion vector information (motion vectors that do not correspond to real motion) their influence should be reduced, and instead, increase the impact of more reliable large blocks that contain edges and higher texture.

Under circumstances, motion prediction benefits from parametric global motion model representations, especially in the background region of a scene that can be better predicted compared to the conventional motion vector prediction. It is noted here that the global parametric model is derived from the actual motion vector field. So a question under study at this point has been whether and how can a region be better predicted using a product of post-processing than using the initial information itself directly. It has been shown that indeed such an improvement is achievable. The justification of this effect, relies on outlier detection and robust regression. It often occurs that motion vectors are erroneously depicting real motion, which can be caused (i) due to miscalculation, e.g. in block matching or (ii) by rate distortion requirements in the case they are calculated within a codec framework. By incorporating robust regression, the goal is to exclude these errors from the model estimation and thus eliminate their influence. Estimators can deal with data containing up to a certain percentage of outliers, which is also known as *breakdown point*. Least squares has a 50% breakdown point, M-estimator close to 50%, Helmholtz tradeoff estimator 80% and RANSAC greater than 50%.

The motion-based object segmentation algorithm exploits bidirectional information between successive frames. The exploitation of two frames instead of one results in more stable segmentation results. A thresholding approach that incorporates multiple improvements on existing approaches is proposed. Along with spatial connectivity, the issue of optimal weight selection for weighted mean hysteresis thresholding is addressed towards avoiding heuristics and data training. It is to be noted that the proposed algorithm requires the tuning by the user of a small number of parameters. Particularly, as discussed in chapter 3 we circumvent the requirement of a thresholding parameter and instead require a broader set of parameters, which enables a more wide application. In the author's point of view, a fully automatic generic algorithm that can deal well with all kinds of video sequences is utopian. Depending

on the application, the desired segmentation results may also vary for a given video sequence. Of course fully automatic approaches have been proposed to deal within specific application frameworks that perform very effectively. For instance in cases where specific patterns are requested (e.g. faces) or in cases where strict assumptions are made (e.g. static cameras in surveillance scenarios) existing approaches are able to address impressively the segmentation task. However, the good performance is narrowed within the specific requirements.

Furthermore, a final post-processing step was incorporated to enable temporal consistency of the segmentation masks. This was achieved using filtering of the preliminary binary masks that is adapted according to the motion of the foreground. It has been shown that the proposed motion compensated error fusion scheme outperforms previously proposed ones. The experimental evaluation demonstrated the validity of the proposed method.

The influence of video compression on the quality of the segmentation results has been also studied. It has been shown that by increasing the compression rate, particularly the quantisation parameter, segmentation accuracy presents only a slight decrease in segmentation accuracy. This shows that segmentation performance is quite robust under the variation of quantisation parameters, which influences the motion estimation quality. This relies on the parametric global motion estimation approach that, by increasing the quantisation parameter up to a certain point, does not decrease the performing accuracy mainly due to increasingly blurring of homogenous regions (due to deblocking filtering).

An interesting direction for future research on this topic is the combination of the proposed motion-based segmentation approach with approaches based on texture and colour features. This may offer advantages in terms of segmentation accuracy with the cost of increased computational complexity. Specific deficiencies of approaches based on solely motion, colour, or texture features can be addressed by combining them. Particularly, problematic cases that can be addressed include: (i) moving and stopping objects - often encountered in motion-based approaches, (ii) changing illumination conditions - from which colour-based approaches suffer, and (iii) vague boundary definitions - which is a major deficiency of texture-based approaches. Moreover, there is room for improvement regarding implementation issues and the computational time can be reduced by optimization of the implementation, especially in the stage of parametric global motion estimation.

The consideration of motion information, specifically the relative motion between global and local one, in a video quality assessment framework shows very promising results for further development of motion-perception aware video quality assessment metrics. Indeed the rather good performance obtained by the weighted spatial pooling using the motion saliency map suggests that the relative motion comprises a valuable clue towards enhancing video quality assessment approaches. The impact of the proposed motion saliency model has shown significant improvement in the agreement with subjective evaluations in the experimental evaluation, where several cases have been studied. It has thus been shown that accounting for motion saliency based on the proposed modelling improves the correlation with the

subjective evaluations.

Considering global motion has been beneficial for video quality assessment with its incorporation in the temporal dimension. The proposed global motion indicator presented good performance in comparison to the commonly used direct average and existing state-of-the-art techniques. This suggests that sophisticated approaches that account for motion features over the temporal dimension, can be beneficial for video quality assessment. Such a claim is reasonable since in video quality assessment, motion plays a critical role. By applying image quality metrics on frame level, and subsequently fusing the local measures using direct average, this important aspect is ignored. Considering temporal dependencies between frames is thus a more suitable approach for assessing the quality of video sequences compared to the case of using image quality metrics.

The incorporation of the moving object segmentation approach into the spatial pooling, by means of foreground and background pooling showed that this approach is not as efficient as the case where a saliency map is used as significance map for spatial pooling. This might rely on the fact that in the latter case the weights are assigned to pixel positions instead of regions which enables more detailed and accurate determination of salient areas. Moreover, this is also attributed to inaccuracies occurred in the moving object segmentation step, where filtering and thresholding may cause moving object regions to be misclassified, for instance mistakenly labeled as background or falsely either dilated or eroded. Having studied the incorporation of motion directly (by means of weighted spatial pooling) and indirectly (by using motion-based segmentation masks) it is concluded that motion shows promising results in enhancing video quality assessment approaches.

To discuss possible further perspectives in this direction, an interesting direction is to combine the impact of global motion in the spatial and temporal dimension. This can be achieved by designing a proper motion-perception based metric that will encapsulate a spatio-temporal consideration of global motion. For instance, in this thesis the global motion indicator, used for temporal pooling, is assigned on the entire frame. This can be extended by assigning a motion indicator to each segment (or an alternatively defined region) of the frame, allowing for spatio-temporal pooling where each segment's motion has an individual impact on the temporal pooling. Furthermore, with respect to temporal pooling, specific types of temporal artefacts that are typically met to compressed video sequences, such as jerkiness (i.e. the perception of still images, instead of moving ones, in a video sequence) or flickering may be studied in order to further exploit motion information.

Other aspects for further exploration include the development of a reduced-reference metric considering for motion features in the spatial and temporal dimension. Reduced-reference metrics are a good compromise between full-reference and no-reference ones. They combine the advantages of full-reference metrics: being accurate and generic, and the advantages of no-reference ones: having broad application in communications today. Thus, the author's opinion is that research on the topic of reduced-reference metrics that account for motion is very promising. Furthermore, enhancing the proposed global motion indicator for temporal pooling towards

adopting more sophisticated approaches for consideration of global motion is an interesting topic for future exploration. First, it would be interesting to study the combination of an attenuation function, which can be based on human visual perception, with the global motion indicator. Secondly, the use of a sliding window over time to account for global motion would enable the flexibility to adjust the influence of global motion alterations on the perception of temporal distortions.

# Description of datasets

---

A brief description of the evaluation datasets used in this thesis is given in this Appendix.

## A.1 Video Dataset 1

Test sequences that no segmentation ground truth is available.

<b>Birds</b>	Description:	Birds flying on the sky
	Camera motion:	Camera pan and tilt
	Number of objects:	Up to six objects
	Background:	Varying texture background
	Size:	$720 \times 576$ , 110 frames



frames 24 & 97

---

---

<b>Monaco</b>	Description:	Camera pan over the harbor of Monaco
	Camera motion:	Slow camera pan
	Number of objects:	None
	Background:	Highly textured background
	Size:	$352 \times 288$ , 150 frames



frames 18 &amp; 135

---

<b>Foreman</b>	Description:	Close up of a talking man
	Camera motion:	No camera motion (first part), camera pan (second part)
	Number of objects:	One large foreground object (first part) and none (second part)
	Background:	Lightly textured background (first part), highly textured background (second part)
	Size:	$352 \times 288$ , 300 frames

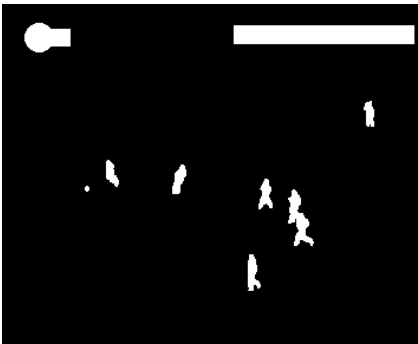
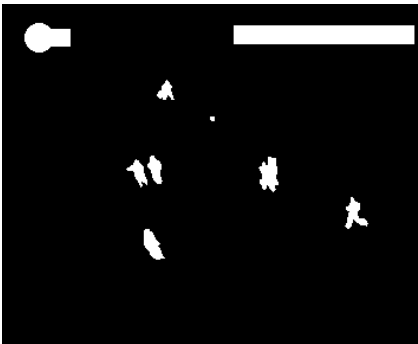


frames 34 &amp; 97

A.2 Video Dataset 2

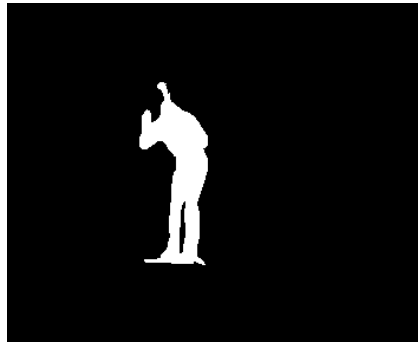
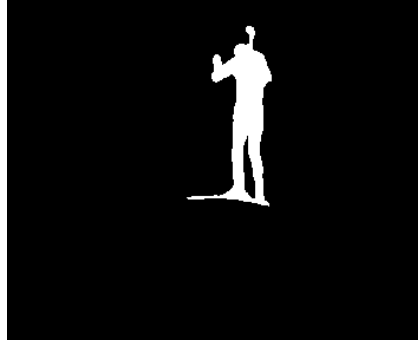
Test sequences that segmentation ground truth is available.

Allstars	Description:	Soccer players playing on the field
	Camera motion:	Slow camera pan and tilt
	Number of objects:	Up to eight small objects
	Background:	Lightly textured background
	Size:	$352 \times 288$ , 250 frames
	SI/TI index:	75/15



original frames 97 (first row) & 162 (second row)      ground truth frames 97 (first row) & 162 (second row)

Biathlon	Description:	A biathlon athlete skiing
	Camera motion:	Fast camera pan and slow zoom
	Number of objects:	One medium sized object
	Background:	Lightly textured background
	Size:	$352 \times 288$ , 200 frames
	SI/TI index:	89/21

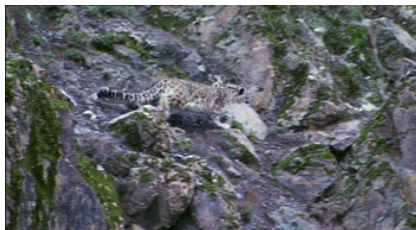


original frames 95 (first row) &  
173 (second row)

ground truth frames 95 (first  
row) & 173 (second row)

### Mountain

Description:	A leopard climbing down on rocks
Camera motion:	Camera pan, tilt and zoom
Number of objects:	One medium object size
Background:	Highly textured background
Size:	$352 \times 288$ , 100 frames
SI/TI index:	79/34



original frames 81 (first row) &  
95 (second row)

ground truth frames 81 (first  
row) & 95 (second row)



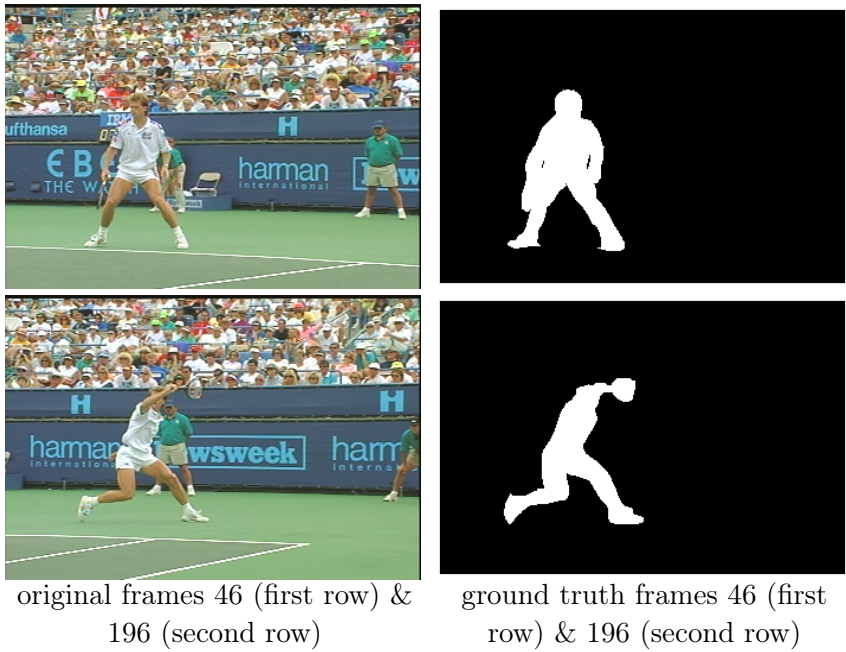
<b>Race</b>	Description:	Go-kart type cars moving across the field
	Camera motion:	Fast camera pan
	Number of objects:	Three objects that undergo size variations (due to perspective)
	Background:	Moderately textured background
	Size:	$544 \times 336$ , 100 frames
	SI/TI index:	104/52



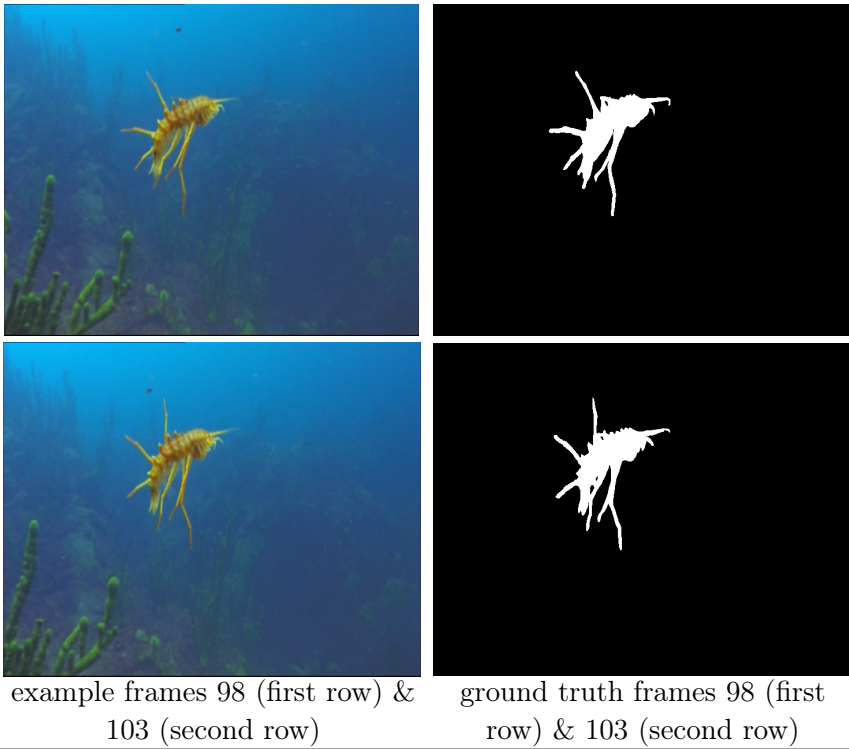
original frames 15 (first row) &  
25 (second row)

ground truth frames 15 (first  
row) & 25 (second row)

<b>Stefan</b>	Description:	A tennis player playing on a tennis field and watching crowd
	Camera motion:	Fast camera pan and zoom
	Number of objects:	Up to two objects; a large one and presence of a much smaller one in several frames (ball)
	Background:	Moderately textured background
	Size:	$352 \times 240$ , 300 frames
	SI/TI index:	153/49



BBC fish	Description:	A fish swimming in the seabed
	Camera motion:	Camera pan, tilt and zoom
	Number of objects:	One medium sized object
	Background:	Lightly textured background
	Size:	720 × 576, 120 frames
	SI/TI index:	29/14



<b>Horse</b>	Description:	A person riding a horse and jumping obstacles.
	Camera motion:	Fast camera pan, fast tilt and zoom
	Number of objects:	One large object
	Background:	Highly textured background
	Size:	352 × 288, 120 frames
	SI/TI index:	114/46



example frames 29 (first row) &  
41 (second row)

ground truth frames 29 (first  
row) & 41 (second row)

---

### A.3 LIVE video database

The LIVE video database [143] has been developed at the University of Texas at Austin and is publicly available<sup>1</sup>. It contains 150 distorted videos obtained from 10 uncompressed reference videos ( $768 \times 432$  pixels, 3206 frames totally) of natural scenes. The distorted videos are created using four commonly encountered distortion types. These include MPEG-2 compression, H.264/AVC compression, simulated transmission of H.264/AVC compressed bitstreams through error-prone IP networks, and through error-prone wireless networks.

Each video was assessed by 38 human subjects in a single stimulus study with hidden reference removal, where the subjects scored the video quality on a continuous quality scale. The difference scores of a given subject are computed by subtracting the score assigned by the subject to the distorted video sequence from the score assigned by the same subject to the corresponding reference video sequence. Following the difference mean opinion score (DMOS) of each video is computed as the mean of the rescaled standardized difference scores (Z-scores) of statistically reliable subjects.

#### A.3.1 Sequences Description

**Blue Sky** Circular camera motion showing a blue sky and some trees.  
 $768 \times 432$ , 217 frames, 25 fps



**River Bed** Still camera, showing a river bed containing some pebbles and water.  
 $768 \times 432$ , 250 frames, 25 fps



**Pedestrian area** Still camera, showing some people walking about in a street intersection.  
 $768 \times 432$ , 250 frames, 25 fps



<sup>1</sup>[http://live.ece.utexas.edu/research/quality/live\\_video.html](http://live.ece.utexas.edu/research/quality/live_video.html)

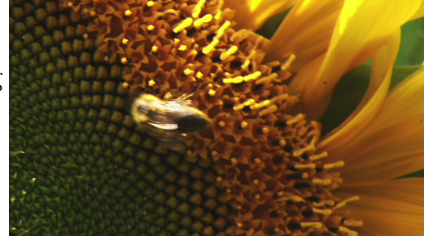


**Tractor**

Camera pan, showing a tractor moving across some fields.  
 $768 \times 432$ , 250 frames, 25 fps

**Sunflower**

Still camera, showing a bee moving over a sunflower in close-up.  
 $768 \times 432$ , 250 frames, 25 fps

**Rush hour**

Still camera, showing rush hour traffic on a street.  
 $768 \times 432$ , 250 frames, 25 fps

**Park run**

Camera pan, a person running across a park.  
 $768 \times 432$ , 500 frames, 50 fps

**Shields**

Camera pans at first, then becomes still and zooms in; shows a person walking across a display pointing at it.  
 $768 \times 432$ , 500 frames, 50 fps

**Mobile& Calendar**

Camera pan, toy train moving horizontally with a calendar moving vertically in the background.  
 $768 \times 432$ , 500 frames, 50 fps



### A.3.2 Distortion types

The distorted videos are created using four commonly encountered distortion types.

**MPEG-2 Compression** This distortion category is referred to as "MPEG-2". Four MPEG-2 compressed videos corresponding to each reference video exist in this database with compression rates that vary from 700 kbps to 4 Mbps, depending on the reference sequence. The MPEG-2 reference software available by the international organization for standardization (ISO) was used.

**H.264/AVC Compression** This distortion category is referred to as "H.264". There are four H.264/AVC compressed videos corresponding to each reference one, with compression rates that vary from 200 kbps to 5 Mbps. The JM reference software (version 12.3) by the joint video team (JVT) was used.

**Transmission Over IP Networks** This distortion category is referred to as "IP". Three IP videos corresponding to each reference one in the database exist that were created by simulating IP losses on an H.264/AVC compressed video stream. The JM reference software (version 12.3) by the JVT was used, with compression rates 0.5 – 7 Mbps. Four IP error patterns supplied by the video coding experts group (VCEG), with loss rates of 3%, 5%, 10%, and 20% were used and the error patterns were obtained from real-world experiments on congested networks and are recommended by the VCEG to simulate the Internet backbone performance for video coding experiments.

**Transmission Over Wireless Networks** This distortion category is referred to as "wireless". The JM reference software (version 12.3) by the JVT was used, with compression rates 0.5 – 7 Mbps. Four videos corresponding to each reference one exist that were created by simulating losses sustained by an H.264/AVC compressed video stream in a wireless environment.





# List of Figures

1.1	Overview of the proposed system. . . . .	2
2.1	Perspective projection example where parallel lines intersect at an ideal point. . . . .	11
2.2	Transformation matrices . . . . .	12
2.3	Feature matching based image registration. . . . .	13
2.4	Block partitioning of frame 22 of the <i>Stefan</i> sequence. . . . .	17
2.5	Blocks of frame 22 of the <i>Stefan</i> sequence. . . . .	21
2.6	Background PSNR in dB for the <i>Stefan</i> sequence. . . . .	25
2.7	Background PSNR in dB for the <i>Biathlon</i> sequence. . . . .	26
2.8	Background PSNR in dB for the <i>Mountain</i> sequence. . . . .	26
2.9	Background PSNR improvement and employed blocks per frame for the <i>Stefan</i> sequence. . . . .	27
2.10	Background PSNR improvement and employed blocks per frame for the <i>Biathlon</i> sequence. . . . .	27
2.11	Background PSNR improvement and employed blocks per frame for the <i>Mountain</i> sequence. . . . .	28
2.12	Background PSNR improvement and size of employed blocks per frame for the <i>Stefan</i> sequence. . . . .	28
2.13	Background PSNR improvement and size of employed blocks per frame for the <i>Biathlon</i> sequence. . . . .	29
2.14	Background PSNR improvement and size of employed blocks per frame for the <i>Mountain</i> sequence. . . . .	29
2.15	Segmentation results for the <i>Stefan</i> and <i>Biathlon</i> test sequences. . .	32
2.16	Segmentation results for the <i>Mountain</i> sequence. . . . .	33
2.17	Possible partition of macroblocks for motion compensation in H.264/AVC. . . . .	34
2.18	Filtering for quarter pixel accurate motion compensation. . . . .	35
2.19	Block diagram of the examined scheme using adaptive mode selection. .	37
2.20	Block partitioning of frame 23 of <i>Stefan</i> sequence using H.264/AVC, reference software KTA [53]. . . . .	41
2.21	Example frames showing block mode allocation in the case of $QP = 38$ for several test sequences. . . . .	43
2.22	PSNR using adaptive mode selection with $ds = 16$ compared to motion vector prediction for each test sequence, for varying $QP$ . . . . .	44
2.23	PSNR using adaptive mode selection with $ds = 4$ compared to motion vector prediction for each test sequence, for varying $QP$ . . . . .	45
2.24	PSNR improvement using adaptive mode selection over motion vector prediction, using $ds = 16$ for all test sequences for varying $QP$ . . . . .	46

2.25	PSNR improvement using adaptive mode selection over motion vector prediction, using $ds = 4$ for all test sequences for varying $QP$ . . . . .	46
3.1	Proposed segmentation system overview and examples. . . . .	52
3.2	Global motion estimation algorithm using the Helmholtz tradeoff estimator and two motion models. . . . .	54
3.3	Example global motion compensated error frames of the <i>Stefan</i> sequence. . . . .	57
3.4	Example of global motion compensated error frames for luminance and chrominance components as well as combination of them for the <i>Mountain</i> sequence. . . . .	58
3.5	Thresholding example for frame 20 of the <i>Stefan</i> sequence. . . . .	60
3.6	Estimated ground truth the first processed frame of the <i>Biathlon</i> sequence, $L = 28$ . . . . .	62
3.7	Chi-square test for finding the optimal weight pair for weighted mean thresholding for <i>Biathlon</i> sequence. . . . .	63
3.8	Test dataset on the spatial - temporal perceptual information plane. . . . .	68
3.9	Reference algorithm 2 system overview. . . . .	69
3.10	Percentage of frames with quality above 75% in terms of F-measure. Comparison of reference and proposed algorithms. . . . .	71
3.11	Comparison of F-measure distributions on the test dataset between the proposed and reference algorithms. . . . .	72
3.12	Original frame, segmentation example of proposed algorithm without and with BCC for <i>Race</i> sequence. . . . .	73
3.13	Dependence of performance in terms of F-measure on parameter $\kappa$ in anisotropic diffusion filtering. . . . .	75
3.14	Number of foreground objects detected with the proposed algorithm and reference algorithms in sequences with multiple objects. . . . .	76
3.15	Precision, recall and F-Measure curves per frame using proposed and reference algorithms for the <i>Allstars</i> and <i>Biathlon</i> sequences. . . . .	78
3.16	Precision, recall and F-Measure per frame using proposed and reference algorithms for the <i>Mountain</i> and <i>Race</i> sequences. . . . .	79
3.17	Precision, recall and F-Measure per frame using proposed and reference algorithms for the <i>Stefan</i> and <i>BBC fish</i> sequences. . . . .	80
3.18	Precision, recall and F-Measure curves per frame using proposed and reference algorithms for the <i>Horse</i> sequence. . . . .	81
3.19	Segmentation examples of <i>Mountain</i> and <i>Stefan</i> sequences using Algorithms 1, 2 and 3 . . . . .	82
3.20	Segmentation examples of <i>Biathlon</i> and <i>Allstars</i> sequences using Algorithms 1, 2 and 3 . . . . .	83
3.21	Segmentation examples of <i>BBC fish</i> and <i>Horse</i> sequences using Algorithms 1, 2 and 3. . . . .	84
3.22	Segmentation system input when implemented at the decoder side. . . . .	85

3.23	Average F-measure of segmentation at the decoder side with varying quantization parameters. . . . .	87
4.1	Classification of objective quality assessment approaches [101]. . . . .	92
4.2	Deviation of objective and subjective quality assessment on the <i>BBC fish</i> sequence. . . . .	96
4.3	Deviation of objective and subjective quality assessment under transient and uniform distortion types on the <i>pedestrian area</i> sequence . . . . .	97
4.4	<i>Method 1</i> framework overview. . . . .	103
4.5	Example using method M1 . . . . .	106
4.6	<i>Method M2</i> framework overview. . . . .	107
4.7	Motion saliency maps $MSA_n$ as heat maps . . . . .	110
4.8	Global motion parameters and the proposed global motion indicator over frames for the <i>mobile calendar</i> sequence . . . . .	113
4.9	The temporal pooling function proposed in [137] . . . . .	114
4.10	Scatter plots of subjective DMOS versus predicted objective quality scores . . . . .	117
4.11	Example frame of the <i>sun flower</i> sequence and weighting maps, depicted as over imposed heat maps, used in Method M2 . . . . .	125
4.12	Example frame of the <i>mobile calendar</i> sequence and weighting maps, depicted as over imposed heat maps, used in Method M2 . . . . .	126
4.13	Scatter plots of DMOS versus predicted DMOS using standard and proposed method . . . . .	127



# List of Tables

2.1	Transformations allowed in motion models. . . . .	12
2.2	Mean background PSNR (in dB) and average block size (in pixels) of the participating blocks (RRBS) comparing reference and proposed algorithms. . . . .	30
2.3	Segmentation approach [45] used for preliminary evaluation. . . . .	31
2.4	PSNR for motion vector prediction, global motion prediction, adaptive mode selection and the corresponding improvements for $bs = 16$ and $bs = 4$ . . . . .	42
3.1	Chi-square test for optimal weight selection for hysteresis thresholding. . . . .	64
3.2	Test sequences and results of experimental evaluation in terms of average precision, recall and F-measure of reference and proposed algorithms. . . . .	71
3.3	Contribution of the background classification consistency to the performance improvement in terms of average precision, recall and F-measure. . . . .	73
3.4	Computational complexity of each step of the proposed algorithm. . . . .	75
3.5	Average F-measure for Otsu, weighted mean and hysteresis weighted mean thresholding algorithms . . . . .	77
3.6	Average precision, recall and F-measure for varying quantization parameters. . . . .	86
4.1	RANSAC approach description. . . . .	108
4.2	VQA metrics performance comparison of each case of foreground/background pooling used in method $M1$ . . . . .	120
4.3	VQA metrics performance comparison of each case of spatial pooling method $M2$ on LIVE video database . . . . .	122
4.4	Performance improvement in terms of $\rho_s$ of our proposed method using motion saliency over standard metrics using the spatial pooling method $M2$ for each distortion class. . . . .	124
4.5	Temporal pooling performance comparison . . . . .	129
4.6	Comparison of methods $M1$ and $M2$ . . . . .	131



# Publications

---

- . M. G. Arvanitidou, M. Tok, A. Glantz, A. Krutz and T. Sikora, *Motion-based object segmentation using hysteresis and bidirectional inter-frame change detection in sequences with moving camera*, Elsevier Signal Processing: Image Communication Journal, 28 (10), 1420 - 1434, 2013.
- . M. G. Arvanitidou, M. Tok, A. Krutz and T. Sikora, *Short-term motion-based object segmentation*, proceedings of the IEEE International Conference on Multimedia and Expo, Barcelona Spain, July 2011
- . M. G. Arvanitidou and T. Sikora, *Motion saliency for spatial pooling of objective video quality metrics*, proceedings of the International workshop on Quality of Experience for Multimedia Content Sharing, Berlin Germany, July 2012
- . F. Kaiser, M. G. Arvanitidou and T. Sikora, *Audio Similarity Matrices Enhancement in an Image Processing Framework*, proceedings of the international workshop on Content-Based Multimedia Indexing, Madrid Spain, June 2011.
- . M. Tok, A. Glantz, M. G. Arvanitidou, A. Krutz and T. Sikora, *Compressed domain global motion estimation using the Helmholtz Tradeoff Estimator*, proceedings of the IEEE International Conference on Image Processing, Hong Kong, September 2010
- . M. G. Arvanitidou, A. Glantz, A. Krutz, T. Sikora, M. Mrak and A. Kondo, *Global motion estimation using variable block sizes and its application to object segmentation*, proceedings of the IEEE International Workshop on Image Analysis for Multimedia Interactive Services, London UK, May 2009
- . A. Astaras, M. G. Arvanitidou, I. Chouvarda, V. Kilintzis, V. Koutkias, E. MontÚn SŒnchez, G. Stalidis, A. Triantafyllidis and N. Maglaveras, *An integrated biomedical telemetry system for sleep monitoring employing a portable body area network of sensors (SENSATION)*, proceedings of the IEEE Engineering in Medicine and Biology society Conference (EMBC), Vancouver Canada, August 2008





# Bibliography

- [1] A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel, and T. Sikora, “Image sequence analysis for emerging interactive multimedia services-the european cost 211 framework,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 7, pp. 802–813, Nov. 1998. (Cited on pages 1 and 50.)
- [2] A. Smolic, “Globale bewegungsbeschreibung und video mosaiking unter verwendung parametrischer 2-d modelle, schätzverfahren und anwendungen,” Ph.D. dissertation, RWTH Aachen, 2001. (Cited on pages 1 and 2.)
- [3] A. Krutz, “From sprites to global motion temporal filtering,” Ph.D. dissertation, Technische Universität Berlin, 2010, nr. 26. (Cited on pages 1 and 4.)
- [4] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. John Wiley & Sons, 2005. (Cited on pages 1, 4, 92, 100 and 114.)
- [5] S. Winkler and P. Mohandas, “The evolution of video quality measurement: From psnr to hybrid metrics,” *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, Sept. 2008. (Cited on pages 1 and 90.)
- [6] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, “Visual attention in quality assessment,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 50–59, Nov. 2011. (Cited on pages 1, 4, 96, 98, 99 and 108.)
- [7] U. Engelke, “Modelling perceptual quality and visual saliency for image and video communications,” Ph.D. dissertation, Blekinge Institute of Technology Sweden, 2010. (Cited on pages 1, 4 and 116.)
- [8] P. Bouthemy, M. Gelgon, and F. Ganansia, “A unified approach to shot change detection and camera motion characterization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 7, pp. 1030–1044, Oct. 1999. (Cited on pages 2 and 14.)
- [9] F. Dufaux and J. Konrad, “Efficient, robust, and fast global motion estimation for video coding,” *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 497–501, Mar 2000. (Cited on pages 2 and 14.)
- [10] Y. Su, M.-T. Sun, and H. V., “Global motion estimation from coarsely sampled motion vector field and the applications,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 232–242, Feb. 2005. (Cited on pages 2 and 15.)
- [11] D. S. Farin, “Automatic video segmentation employing object/camera modeling techniques,” Ph.D. dissertation, Technische Universiteit Eindhoven, 2005. (Cited on pages 2, 4 and 15.)

- 
- [12] M. Tok, A. Glantz, A. Krutz, and T. Sikora, “Monte-carlo-based parametric motion estimation using a hybrid model approach,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 607–620, 2013. (Cited on pages 3, 14 and 15.)
  - [13] P. J. Huber, *Robust Statistics*. John Wiley & Sons, 1981. (Cited on pages 3 and 19.)
  - [14] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. John Wiley & Sons, 1987. (Cited on pages 3, 19 and 54.)
  - [15] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011, vol. 114. (Cited on page 3.)
  - [16] O. E. Parmenides, *On Nature (Περὶ Φύσεως)*, ca. B.C. 475. (Cited on page 3.)
  - [17] A. Amer, “Object and event extraction for video processing and representation in on-line video applications,” Ph.D. dissertation, Ottawa University, 2001. (Cited on page 4.)
  - [18] D. Zhang and G. Lu, “Segmentation of moving objects in image sequence: A review,” *Circuits, Systems, and Signal Processing, Springer*, vol. 20, no. 2, pp. 143–183, Mar. 2001. (Cited on pages 4, 15 and 49.)
  - [19] L. T. To, “Video object segmentation using phase-based detection of moving object boundaries,” Ph.D. dissertation, University of New South Wales, 2005. (Cited on pages 4 and 15.)
  - [20] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec 2012. (Cited on pages 4 and 15.)
  - [21] T. Wiegand, G. Sullivan, G. Bjntegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, July 2003. (Cited on pages 4, 15, 35 and 36.)
  - [22] F. Kaiser, M. G. Arvanitidou, and T. Sikora, “Audio similarity matrices enhancement in an image processing framework,” in *proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI)*, Madrid, Spain, Jun. 2011. (Cited on page 6.)
  - [23] J. Foote, “Visualizing music and audio using self-similarity,” in *proceedings of the ACM International Conference on Multimedia*. ACM, 1999, pp. 77–80. (Cited on page 6.)

- [24] A. Smolic, T. Sikora, and J.-R. Ohm, “Long-term global motion estimation and its application for sprite coding, content description, and segmentation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1227–1242, Dec. 1999. (Cited on pages 10 and 11.)
- [25] T. Sikora, “Trends and perspectives in image and video coding,” *proceedings of the IEEE*, vol. 93, no. 1, pp. 6–17, Jan. 2005. (Cited on page 10.)
- [26] B. Qi, M. Ghazal, and A. Amer, “Robust global motion estimation oriented to video object segmentation,” *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 958–967, June 2008. (Cited on pages 10, 14, 50, 69, 71 and 77.)
- [27] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2000, vol. 2. (Cited on page 11.)
- [28] M. Kunter, “Advances in sprite-based video coding towards universal usability,” Ph.D. dissertation, Technische Universität Berlin, 2008. (Cited on pages 11 and 13.)
- [29] A. Krutz, M. Frater, M. Kunter, and T. Sikora, “Windowed image registration for robust mosaicing of scenes with large background occlusions,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, Atlanta, GA, USA, Oct. 2006, m. Frater: Australian Defence Force Academy, Canberra, Australia. (Cited on page 14.)
- [30] A. M. Tourapis, “Enhanced predictive zonal search for single and multiple frame motion estimation,” in *Electronic Imaging*. International Society for Optics and Photonics, 2002, pp. 1069–1079. (Cited on page 15.)
- [31] A. Smolic, M. Hoeyneck, and J.-R. Ohm, “Low-complexity global motion estimation from p-frame motion vectors for mpeg-7 applications,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, vol. 2, 2000, pp. 271–274. (Cited on pages 15, 16, 19, 22 and 23.)
- [32] E. Saez, J. M. Palomares, J. I. Benavides, and N. Guil, “Global motion estimation algorithm for video segmentation,” in *SPIE Visual Communications and Image Processing*, vol. 5150, 2003, pp. 1540–1550. (Cited on page 15.)
- [33] D. Farin and P. H. de Witha, “Evaluation of a feature-based global-motion estimation system,” in *Visual Communications and Image Processing*, vol. 5960, 2005. (Cited on pages 15 and 106.)
- [34] Y.-M. Chen, I. V. Bajic, and P. Saeedi, “Motion segmentation in compressed video using markov random fields,” in *proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, jul. 2010, pp. 760–765. (Cited on pages 15 and 49.)

- 
- [35] A. Alatan, E. Tuncel, and L. Onural, “A rule-based method for object segmentation in video sequences,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, vol. 2, okt 1997, pp. 522–525. (Cited on page 15.)
  - [36] C.-T. Hsu, Y.-S. Tsai, M.-H. Hsieh, and Y.-N. Chien, “Video object segmentation based on global motion estimation/compensation,” in *proceedings of the International Conference on Consumer Electronics*, 2001, pp. 168–169. (Cited on page 15.)
  - [37] J. Heuer and A. Kaup, “Global motion estimation in image sequences using robust motion vector field segmentation,” in *proceedings of the ACM Multimedia*, Oct. 1999, pp. 261–264. (Cited on page 15.)
  - [38] S. Treetasanatavorn, J. Heuer, U. Rauschenbach, K. Illgner, and A. Kaup, “Temporal video segmentation using global motion estimation and discrete curve evolution,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, vol. 1, 24-27 Oct. 2004, pp. 385–388. (Cited on page 16.)
  - [39] Y. Chen and I. Bajic, “A joint approach to global motion estimation and motion segmentation from a coarsely sampled motion vector field,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, p. 1, 2011. (Cited on page 16.)
  - [40] B. Kamolrat, W. A. C. Fernando, M. Mrak, and A. Kondoz, “Flexible motion model with variable size blocks for depth frames coding in colour-depth based 3d video coding,” in *proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, June 2008. (Cited on page 16.)
  - [41] M. Servais, T. Vlachos, and T. Davies, “Motion-compensation using variable-size block-matching with binary partition trees,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, vol. 1, Sept. 2005, pp. I–157. (Cited on page 16.)
  - [42] W. Press, S. Teukolsky, W. Vetterling, and B. Flannary, *Numerical recipes in C*. Cambridge University Press, 1992. (Cited on page 19.)
  - [43] M. Haller, A. Krutz, and T. Sikora, “Robust global motion estimation using motion vectors of variable size blocks and automatic motion model selection,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, Hong Kong, Sep. 2010. (Cited on page 21.)
  - [44] M. Arvanitidou, A. Glantz, A. Krutz, T. Sikora, M. Mrak, and A. Kondoz, “Global motion estimation using variable block sizes and its application to object segmentation,” in *proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, May 2009, pp. 173–176. (Cited on page 23.)

- [45] A. Krutz, M. Kunter, M. Mandal, M. Frater, and T. Sikora, “Motion-based object segmentation using sprites and anisotropic diffusion,” in *proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Jun. 2007. (Cited on pages 31, 59, 61, 74, 77 and 155.)
- [46] G. Sullivan and T. Wiegand, “Video compression - from concepts to the H.264/AVC standard,” *proceedings of the IEEE*, vol. 93, no. 1, pp. 18–31, 2005. (Cited on page 35.)
- [47] I. E. Richardson, *H.264 and MPEG-4 Video Compression: Video Coding for Next-Generation Multimedia*. John Wiley & Sons, 2004. (Cited on page 35.)
- [48] A. Smolic, Y. Vatis, H. Schwarz, P. Kauff, U. Gölz, and T. Wiegand, “Improved video coding using long-term global motion compensation,” in *Electronic Imaging*. International Society for Optics and Photonics, 2004, pp. 343–354. (Cited on page 36.)
- [49] H. Yu, Z. Lin, and F. Teo, “An efficient coding scheme based on image alignment for H.264/AVC,” in *proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, 2009, pp. 629–632. (Cited on page 37.)
- [50] T. Wiegand, E. Steinbach, and B. Girod, “Affine multipicture motion-compensated prediction,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 197–209, 2005. (Cited on page 37.)
- [51] R. Kordasiewicz, M. Gallant, and S. Shirani, “Affine motion prediction based on translational motion vectors,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 10, pp. 1388–1394, 2007. (Cited on page 37.)
- [52] A. Glantz, M. Tok, A. Krutz, and T. Sikora, “A block-adaptive skip mode for inter prediction based on parametric motion models,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 1201–1204. (Cited on page 37.)
- [53] K. T. A. KTA, “H.264/AVC,” ITU-T VCEG, available at: <http://www.tnt.uni-hannover.de/~vatis/cta/>, Tech. Rep., Nov. 2010. (Cited on pages 38, 41, 85 and 151.)
- [54] K. Koffka, *Principles of Gestalt Psychology*. Hartcourt Brace Jovanovich, New York, 1935. (Cited on pages 48 and 101.)
- [55] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, “Image change detection algorithms: a systematic survey,” *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, Mar. 2005. (Cited on page 48.)
- [56] M. G. Arvanitidou, M. Tok, A. Krutz, and T. Sikora, “Short-term motion-based object segmentation,” in *proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, Barcelona, Spain, Jul. 2011. (Cited on pages 48, 70 and 73.)

- 
- [57] M. G. Arvanitidou, M. Tok, A. Glantz, A. Krutz, and T. Sikora, “Motion-based object segmentation using hysteresis and bidirectional inter-frame change detection in sequences with moving camera,” *Signal Processing: Image Communication*, vol. 28, no. 10, pp. 1420 – 1434, 2013. (Cited on pages 48 and 73.)
  - [58] A. M. Tekalp, *Digital video processing*. Prentice Hall, 1995, vol. 1. (Cited on page 48.)
  - [59] L. Zappella, X. Lladó, and J. Salvi, “Motion segmentation: A review,” in *proceedings of the Conference on Artificial Intelligence Research and Development*. IOS Press, 2008, pp. 398–407. (Cited on page 49.)
  - [60] T. Aach, A. Kaup, and R. Mester, “Statistical model-based change detection in moving video,” *Elsevier Signal Processing*, vol. 31, no. 2, pp. 165–180, 1993. (Cited on page 49.)
  - [61] H. Shen, L. Zhang, B. Huang, and P. Li, “A map approach for joint motion estimation, segmentation, and super resolution,” *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 479–490, 2007. (Cited on page 49.)
  - [62] Y. Tsaig and A. Averbuch, “Automatic segmentation of moving objects in video sequences: a region labeling approach,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 7, pp. 597–612, Jul. 2002. (Cited on page 49.)
  - [63] J. Weber and J. Malik, “Rigid body segmentation and shape description from dense optical flow under weak perspective,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 139–143, 1997. (Cited on page 49.)
  - [64] A. Bugeau and P. Perez, “Detection and segmentation of moving objects in complex scenes,” *Computer Vision and Image Understanding*, vol. 113, no. 4, pp. 459 – 476, 2009. (Cited on page 49.)
  - [65] T. Senst, V. Eiselein, M. Pätzold, and T. Sikora, “Efficient real-time local optical flow estimation by means of integral projections,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, Brussels, Belgium, Sep. 2011, pp. 2393–2396. (Cited on page 49.)
  - [66] M. P. Kumar, P. H. Torr, and A. Zisserman, “Learning layered motion segmentations of video,” *International Journal of Computer Vision*, vol. 76, no. 3, pp. 301–319, 2008. (Cited on page 49.)
  - [67] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, “Bilayer segmentation of live video,” in *proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 53–60. (Cited on page 49.)



- [68] J. Y. Wang and E. H. Adelson, “Representing moving images with layers,” *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 625–638, 1994. (Cited on page 50.)
- [69] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: a factorization method,” *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992. (Cited on page 50.)
- [70] T. Morita and T. Kanade, “A sequential factorization method for recovering shape and motion from image streams,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 8, pp. 858–867, 1997. (Cited on page 50.)
- [71] J. Costeira and T. Kanade, “A multi-body factorization method for motion analysis,” in *proceedings of the IEEE International Conference on Computer Vision*, 1995, pp. 1071–1076. (Cited on page 50.)
- [72] Y. Sheikh, O. Javed, and T. Kanade, “Background subtraction for freely moving cameras,” in *proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 1219–1225. (Cited on page 50.)
- [73] S.-C. S. Cheung and C. Kamath, “Robust techniques for background subtraction in urban traffic video,” in *SPIE*, vol. 5308, 2004, pp. 881–892. (Cited on page 50.)
- [74] M. Piccardi, “Background subtraction techniques: a review,” in *proceedings of the IEEE International Conference on Systems, man and cybernetics*, vol. 4, 2004, pp. 3099–3104. (Cited on page 50.)
- [75] S. Berrabah, G. De Cubber, V. Enescu, and H. Sahli, “Mrf-based foreground detection in image sequences from a moving camera,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, Oct. 2006, pp. 1125–1128. (Cited on page 50.)
- [76] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in *proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999. (Cited on page 50.)
- [77] R. H. Evangelio, M. Pätzold, I. Keller, and T. Sikora, “Adaptively splitted gmm with feedback improvement for the task of background subtraction,” *IEEE Transactions on Information Forensics & Security*, vol. 9, no. 5, pp. 863–874, May 2014. (Cited on page 50.)
- [78] N. Goyette, P. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, “Change detection.net: A new change detection benchmark dataset,” in *proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2012, pp. 1–8. (Cited on page 50.)

- [79] C. Kim and J.-N. Hwang, "Fast and automatic video object segmentation and tracking for content-based applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 122–129, Feb 2002. (Cited on pages 50, 52 and 56.)
- [80] Y. Kameda and M. Minoh, "A human motion estimation method using 3-successive video frames," *proceedings of International Conference on Virtual Systems*, Jan 1996. (Cited on pages 51, 55, 69, 71 and 77.)
- [81] I. Patras, "Object-based video segmentation with region labeling," Ph.D. dissertation, TU Delft, 2001. (Cited on page 51.)
- [82] M.-Y. Shih, Y.-J. Chang, B.-C. Fu, and C.-C. Huang, "Motion-based background modeling for moving object detection on moving platforms," in *proceedings of the International Conference on Computer Communications and Networks*, Aug. 2007, pp. 1178–1182. (Cited on pages 51 and 55.)
- [83] J.-C. Huang, T.-S. Su, L.-J. Wang, and W.-S. Hsieh, "Double-change-detection method for wavelet-based moving-object segmentation," *Electronics Letters*, vol. 40, no. 13, pp. 798–799, Jun 2004. (Cited on pages 51 and 55.)
- [84] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989. (Cited on page 51.)
- [85] H. Liu, X. Chen, Y. Chen, and C. Xie, "Double change detection method for moving-object segmentation based on clustering," in *proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2006, p. 4 pp. (Cited on page 51.)
- [86] M. Tok, A. Glantz, M. G. Arvanitidou, A. Krutz, and T. Sikora, "Compressed domain global motion estimation using the helmholtz tradeoff estimator," in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, Hong Kong, Sep. 2010. (Cited on pages 52 and 53.)
- [87] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, Jul. 1990. (Cited on pages 56, 58 and 109.)
- [88] J. Weickert, *Anisotropic diffusion in image processing*. Teubner Stuttgart, 1998, vol. 1. (Cited on pages 58 and 109.)
- [89] P. Rosin and T. Ellis, "Image difference threshold strategies and shadow detection," in *British Machine Vision Conference*. BMVA Press, Jul. 1995, pp. 347–356. (Cited on page 60.)
- [90] E. Hancock and J. Kittler, "Adaptive estimation of hysteresis thresholds," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 1991, pp. 196–201. (Cited on page 61.)



- [91] Y. Yitzhaky and E. Peli, “A method for objective edge detection evaluation and detector parameter selection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 1027 – 1033, aug. 2003. (Cited on pages 61 and 62.)
- [92] R. Medina-Carnicer, F. Madrid-Cuevas, A. Carmona-Poyato, and R. Muñoz-Salinas, “On candidates selection for hysteresis thresholds in edge detection,” *Pattern Recognition*, vol. 42, no. 7, pp. 1284–1296, 2009. (Cited on page 61.)
- [93] D. Farin, T. Haenselmann, S. Kopf, G. Kühne, and W. Effelsberg, “Segmentation and classification of moving video objects,” *Handbook of Video Databases*, CRC Press, vol. 8, pp. 561 – 591, 2002. (Cited on page 65.)
- [94] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62 –66, Jan. 1979. (Cited on pages 66, 74 and 77.)
- [95] ITU-T, “P.910 Subjective video quality assessment methods for multimedia applications,” International Telecommunication Union, available at: <http://www.itu.org/>, Tech. Rep., 2008. (Cited on pages 67, 91 and 92.)
- [96] M. Arvanitidou and T. Sikora, “Motion saliency for spatial pooling of objective video quality metrics,” in *Workshop on Quality of Experience for Multimedia Content Sharing (QoEMCS), proceedings of the European Conference on Interactive TV and Video (EuroiTV)*, Berlin, Jul. 2012. (Cited on page 90.)
- [97] ITU-R, “BT.500-13 Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, available at: <http://www.itu.org/>, Tech. Rep., 2012. (Cited on pages 91 and 115.)
- [98] M. H. Pinson and S. Wolf, “Comparing subjective video quality testing methodologies,” in *Visual Communications and Image Processing*. International Society for Optics and Photonics, 2003, pp. 573–582. (Cited on page 92.)
- [99] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment, ph.II.” VQEG, <http://www.vqeg.org>, Tech. Rep., 2003. (Cited on pages 92, 116 and 118.)
- [100] S. Zielinski, F. Rumsey, and S. Bech, “On some biases encountered in modern audio quality listening tests-a review,” *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008. (Cited on page 92.)
- [101] Z. Wang and A. C. Bovik, “Modern image quality assessment,” *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006. (Cited on pages 92, 93 and 153.)
- [102] H. R. Sheikh, A. C. Bovik, and L. Cormack, “No-reference quality assessment using natural scene statistics: Jpeg2000,” *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1918–1927, 2005. (Cited on page 93.)

- 
- [103] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, vol. 1, 2002, pp. I-477. (Cited on page 93.)
- [104] M. Shahid, A. Rossholm, B. Lövsström, and H.-J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 40, 2014. (Cited on page 93.)
- [105] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Electronic Imaging*. International Society for Optics and Photonics, 2005, pp. 149–159. (Cited on pages 93 and 99.)
- [106] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 202–211, 2009. (Cited on page 93.)
- [107] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004. (Cited on pages 93, 94, 98, 101 and 116.)
- [108] B. Girod, *What's wrong with mean-squared error?* Visual Factors of Electronic Image Communications, MIT Press, 1993. (Cited on page 94.)
- [109] D. Chandler and S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284 –2298, sept. 2007. (Cited on pages 94 and 97.)
- [110] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Elsevier Signal Processing: Image Communication*, vol. 19, no. 1, Jan. 2004. (Cited on pages 95, 100 and 128.)
- [111] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *proceedings of the Asilomar Conference on Signals, Systems and Computers*, vol. 2, nov. 2003, pp. 1398 – 1402 Vol.2. (Cited on pages 95, 116 and 118.)
- [112] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005. (Cited on pages 95 and 116.)
- [113] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004. (Cited on pages 95, 130 and 131.)

- [114] K. Seshadrinathan, , and A. Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010. (Cited on pages 95, 130 and 131.)
- [115] Z. Wang and A. C. Bovik, “Mean squared error: love it or leave it? a new look at signal fidelity measures,” *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009. (Cited on page 96.)
- [116] Z. Wang, A. C. Bovik, and L. Lu, “Why is image quality assessment so difficult?” in *proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2002, pp. IV–3313. (Cited on page 96.)
- [117] B. A. Wandell, *Foundations of vision*. Sinauer Associates, 1995. (Cited on page 97.)
- [118] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov 1998. (Cited on pages 98, 105, 111, 112, 119, 121 and 122.)
- [119] L. Itti and P. F. Baldi, “Bayesian surprise attracts human attention,” in *Advances in neural information processing systems*, 2005, pp. 547–554. (Cited on page 98.)
- [120] W. Osberger, N. Bergmann, and A. Maeder, “An automatic image quality assessment technique incorporating higher level perceptual factors,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, oct 1998, pp. 414 –418 vol.3. (Cited on page 99.)
- [121] U. Engelke, V. X. Nguyen, and H.-J. Zepernick, “Regional attention to structural degradations for perceptual image quality metric design,” in *proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 869 –872. (Cited on pages 99 and 101.)
- [122] U. Engelke and H. Zepernick, “Quality evaluation in wireless imaging using feature-based objective metrics,” in *proceedings of the International Symposium on Wireless Pervasive Computing*, 2007. (Cited on page 99.)
- [123] A. Cavallaro and S. Winkler, “Segmentation driven perceptual quality metrics,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2004. (Cited on page 99.)
- [124] U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick, “Modelling saliency awareness for objective video quality assessment,” in *proceedings of the International Workshop on Quality of Multimedia Experience (QoMEX)*, Ju. 2010, pp. 212–217. (Cited on page 99.)

- 
- [125] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, April 2009. (Cited on page 99.)
- [126] B. Ghanem, E. Resendiz, and N. Ahuja, "Segmentation-based perceptual image quality assessment (spiq), in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, oct. 2008, pp. 393–396. (Cited on pages 99 and 101.)
- [127] X. Gu, G. Qiu, X. Feng, L. Debing, and C. Zhibo, "Region of interest weighted pooling strategy for video quality metric," *Telecommunication Systems*, pp. 1–11, 2010. (Cited on pages 100, 101 and 104.)
- [128] L. Ma, S. Li, and K. Ngan, "Motion trajectory based visual saliency for video quality assessment," in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, sept. 2011, pp. 233–236. (Cited on pages 100, 111, 121 and 122.)
- [129] Y. Wang, T. Jiang, S. Ma, and W. Gao, "Novel spatio-temporal structural information based video quality metric," *IEEE Transactions on Circuits and Systems for Video Technology*, no. 99, p. 1, 2012. (Cited on page 100.)
- [130] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *Journal of the Optical Society of America*, vol. 24, no. 12, pp. B61–B69, Dec 2007. (Cited on pages 100, 106 and 109.)
- [131] S. Rimac-Drlje, M. Vranjes, and D. Zagar, "Influence of temporal pooling method on the objective video quality evaluation," in *proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2009, pp. 1–5. (Cited on page 100.)
- [132] H. de Ridder, "Minkowski-metrics as a combination rule for digital-image-coding impairments," in *SPIE/IS&T Symposium on Electronic Imaging: Science and Technology*, 1992, pp. 16–26. (Cited on pages 100 and 113.)
- [133] D. E. Pearson, "Viewer response to time-varying video quality," in *SPIE Human Vision and Electronic Imaging*, vol. 3299. International Society for Optics and Photonics, 1998, pp. 16–25. (Cited on page 100.)
- [134] K. Lee, J. Park, S. Lee, and A. C. Bovik, "Temporal pooling of video quality estimates using perceptual motion models," in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 2493–2496. (Cited on page 101.)
- [135] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 253–265, 2009. (Cited on page 101.)

- [136] A. Moorthy and A. Bovik, “Efficient video quality assessment along temporal trajectories,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1653–1658, Nov. 2010. (Cited on pages 101, 130 and 131.)
- [137] J. You, J. Korhonen, and A. Perkis, “Attention modeling for video quality assessment: Balancing global quality and local quality,” in *proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2010, pp. 914–919. (Cited on pages 101, 114, 128, 129 and 153.)
- [138] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, “Spatio-temporal quality pooling accounting for transient severe impairments and egomotion,” in *proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2509–2512. (Cited on page 101.)
- [139] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, “To pool or not to pool: A comparison of temporal pooling methods for HTTP adaptive video streaming,” in *proceedings of the IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 52–57. (Cited on page 101.)
- [140] C. Tomasi and T. Kanade, *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon University, 1991. (Cited on page 107.)
- [141] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981. (Cited on page 108.)
- [142] M. A. Cohen and S. Grossberg, “Neural dynamics of brightness perception: Features, boundaries, diffusion, and resonance,” *Perception & psychophysics*, vol. 36, no. 5, pp. 428–456, 1984. (Cited on page 109.)
- [143] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010. (Cited on pages 115 and 147.)