Advanced Methods for Image Information Mining System: Evaluation and Enhancement of User Relevance

von Diplom-Ingenieur
Herbert Andreas Daschiel
aus Traunstein

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

Doktor der Ingenieurwissenschaften
– Dr.-Ing. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Heinz Lemke
Gutachter:	Prof. Dr.-Ing. Olaf Hellwich
Gutachter:	Prof. Dr.-Ing. Mihai Datcu
Gutachter:	Prof. Dr.-Ing. Lothar Gründig

Tag der wissenschaftlichen Aussprache: 15. Juli 2004

Berlin 2004
D 83

To Angelika and Michael

Abstract

Spurred by the needs to provide innovative tools for the management and query of large image databases, techniques for content-based image retrieval have been developed within the last years. As in other research fields, further progress depends on the ability to carefully evaluate both image retrieval and image understanding functions — a field that has not been given much attention yet. The difficulty in validating an image retrieval system is due to the lack of standardized evaluation criteria and the arbitrary selection of datasets.

In this dissertation, we present a concept for the enhancement and evaluation of a knowledge-driven content-based image information mining system. The main application of the system is to provide users with a tool by which the content of remote sensing image archives can be explored without actually browsing them. Therefore, the system models the image data in a hierarchical Bayesian way, where the content information is arranged at multiple levels of different semantic abstraction. The hierarchy consists of two major parts: a computational-intensive off-line data processing, which aims at the extraction and description of the image content in a completely unsupervised and application-free way (level 0 to 3) and a fast, user-specific semantic labeling of cover-types (level 4). Since each semantic label is linked to the image content by using stochastic parametric signal models, the archive can be queried in a probabilistic way.

Although the concept of application-free image content modeling and application-specific labeling has shown its usefulness in various practical applications, it is, however, of limited use for the definition of complex cover-types. Hence, a new level of image content abstraction is introduced: semantic grouping. This method is based on the aggregation of existing labels to high-level semantic concepts (level 5) according to the user's feedback samples. Additionally, this method of image content definition at semantic level can be extended to query remote sensing image archives across sensors and data collections.

This thesis mainly contributes to content-based image retrieval system evaluation and verification. Unlike other methods that express the overall performance of a retrieval system in terms of relevant and irrelevant images in the query results, our approach first decomposes the overall system into its basic components, then evaluates each one using information-theoretic quantities and finally combines the individual measurements to indicate the overall system performance. The advantage

of this strategy is its full adaptability to the hierarchical scheme underlying the mining system. Moreover, the stochastic nature of the system is fully incorporated in information theory. Additionally, analyzing individual system modules and fusing the measurements is easier than analyzing the overall system complexity in one step.

Information theory provides a number of powerful measurements to analyze the technical objective quality of system components, to identify user-related subjective concepts and to verify the information flow during system operation. We outline methods to validate the information content of primitive image features, unsupervised clusters and semantic labels, to analyze the effectiveness of the interactive learning and probabilistic search system modules and to demonstrate how semantic cover-types are preserved for increasing volumes of data and subspace clusters. Subjective evaluation mainly involves the analysis of human-machine interactions, that is, performed actions, timing of actions, communication and information representation aspects and the prediction of the users' interests. In order to verify the information transmission between different levels in the hierarchical scheme, we analyze the mining system from an imperfect communication channel point of view.

Validating the image information mining system requires a particular organization and the development of appropriate tools. Therefore, we trace all performed actions between the users and the system and implemented functions for the statistical analysis of user-tracing parameters. With these tools we get a set of evaluation measurements that reflect the objective quality of the system and compare them with subjective user satisfaction. In order to verify the performance of the obtained measurements, we organized a large-scale system evaluation test with participants from different working fields. The results obtained show both the performance of the system and the relevance of implemented functions for system validation.

In order to verify the overall performance of the system, we selected various datasets with different properties according to certain application scenarios. We analyzed the computational requirements for off-line feature extraction, clustering and catalogue entry generation, the quality of the system to manage large volumes of data, the complexity of the applied datasets, the performance of the graphical man-machine interface and the quality of the defined semantic image content.

Zusammenfassung

Die Notwendigkeit eines effizienten Zugriffs auf große Bildbestände in Datenbanken führte zur Entwicklung von Verfahren der inhaltsbasierten Bildsuche (Image Data Mining). Die Weiterentwicklung dieser Verfahren hängt — so wie in anderen Forschungsgebieten auch — davon ab, ob sich die Methoden der Bildabfrage als auch des Bildverstehens quantitativ bewerten lassen. Bislang wurde der Thematik der Evaluierung von Systemen zur inhaltsorientierten Bildsuche nur sehr wenig Aufmerksamkeit geschenkt. Die Schwierigkeit dieser Systemevaluierung liegt am Mangel standardisierter Kriterien und willkürlicher Bildauswahl.

In dieser Dissertation wird ein Konzept zur Weiterentwicklung und Evaluierung eines wissens- und inhaltsbasierten Bildarchivierungssystems vorgestellt. Die Hauptanwendung des Systems liegt darin, Nutzern ohne die Visualisierung alle Bilder den Zugriff auf den Bildinhalt in Fernerkundungsdatenbanken zu ermöglichen. Um diesen Zugriff effektiv zu gestalten, werden die Bilddaten in einer sog. Bayes'schen Hierarchie modelliert. Der Informationsinhalt wird dann auf mehreren Ebenen mit verschiedener semantischer Abstraktion angeordnet. Die Hierarchie gliedert sich in zwei Hauptteile. Der erste Teil besteht aus einer rechenintensiven off-line Verarbeitung der Bilddaten mit dem Ziel, Bildmerkmale zu extrahieren und den Bildinhalt in einer vollständig unüberwachten (unsupervised) und anwendungsfreien Art abstrakt auf Signalebene zu beschreiben. Der zweite Teil besteht aus der Definition von Bodenbedeckungstypen und der semantischen Namenszuweisung. Da jede Bodenbedeckungsart über stochastische Signalmodelle dem Bildinhalt zugeordnet ist, kann das Bildarchiv "stochastisch" abgefragt werden.

Obwohl das Konzept der anwendungsfreien Modellierung und der anwendungsspezifischen semantischen Beschreibung in vielen praktischen Beispielen zum Erfolg führte, war es nur von begrenztem Nutzen für die Definition von komplexen Bildinhalten. Deshalb wurde eine neue Abstraktionsebene eingeführt: die semantische Gruppierung. Diese Methode basiert auf der Zusammenfassung von vorhandenen Bodenbedeckungstypen in höherrangige semantische Konzepte unter Einbeziehung der Nutzer. Darüberhinaus kann diese Methode, unabhängig von Aufnahmesensor und Datensätzen, für die Beschreibung von Bildinhalten auf semantischem Niveau für die Abfrage von Fernerkundungsdatenbanken verwendet werden.

Der Hauptteil dieser Arbeit beschäftigt sich mit der Evaluierung eines Systems zur inhaltsbasierten Bildsuche. Im Gegensatz zu anderen Verfahren, welche die

Gesamtleistung eines Systems hinsichtlich relevanter und irrelevanter Bilder in den Abfrageergebnissen beschreiben, zerlegt der vorgestellte Ansatz zuerst das Gesamtsystem in seine Einzelkomponenten. Bevor die einzelnen Messwerte zusammengefasst werden, um die Gesamtleistung des Systems zu ermitteln, wird jedes Modul basierend auf informationstheoretischen Messgrössen evaluiert. Der Vorteil dieses Ansatzes liegt darin, dass er vollständig in das hierarchische Konzept des Bildarchivierungssystems implementiert werden kann. Hinzu kommt, dass die stochastische Natur des Systems vollständig in die Informationstheorie eingebunden wird. Denn es ist einfacher, einzelne Systemmodule zu analysieren und die Messgrössen zu kombinieren als die Gesamtkomplexität des Systems in einem Schritt zu beschreiben.

Die Informationstheorie stellt eine Reihe leistungsstarker Messgrössen zur Verfügung, um die technische Qualität von Komponenten des Systems zu überprüfen, nutzerrelevante Aspekte zu verifizieren und den Informationsfluss während des Systembetriebs zu untersuchen. Es werden Verfahren zur Evaluierung von primitiven Bildmerkmalen, unüberwachten Clustern und überwachten Bodenbedeckungstypen mit semantischer Bedeutung vorgestellt. Desweiteren wird die Leistungsfähigkeit der Systemmodule "Interactive Learning" und "Probabilistic Search" beschrieben und gezeigt, inwieweit Bodenbedeckungstypen für zunehmende Datenmengen und Cluster von Unterräumen beibehalten werden. Die subjektive Evaluierung beinhaltet hauptsächlich die Mensch-Maschine Kommunikation, das heißt die ausgeführten Aktionen von Nutzern, den Zeitaufwand für Aktionen, die Aspekte der Kommunikation und Informationsdarstellung und die Vorhersage von Nutzerinteressen. Um die Informationsübertragung zwischen verschiedenen Systemniveaus zu evaluieren, wird das Bildarchivierungssystem aus Sicht eines unvollkommenen Nachrichtenübertragungskanals untersucht.

Die Überprüfung des verwendeten Mining Systems erfordert eine besondere Architektur und die Entwicklung von geeigneten Verfahren. Um dies zu bewerkstelligen, werden alle getätigten Aktionen zwischen Nutzer und System festgehalten und statistische Verfahren zur Untersuchung dieser Parameter entwickelt. Diese Methoden liefern eine Reihe von Messwerten, welche die objektive Qualität des Systems wiederspiegeln. Im Anschluß daran werden sie mit der subjektiven Bewertung der Nutzer verglichen. Zur Überprüfung der Signifikanz der erhaltenen Messwerte wurde ein grossangelegter Systemtest mit Teilnehmern aus verschiedenen Arbeitsgebieten durchgeführt. Die Ergebnisse geben Auskunft über die Leistung des Systems und die Bedeutung der implementierten Module für die Systemevaluierung.

Um die Gesamtleistung des Mining Systems zu evaluieren, wurden, je nach Anwendung, mehrere Datensätze mit verschiedenen Eigenschaften ausgewählt. Folgende Systemanforderungen wurden untersucht: die Extraktion und Kompression von Bildmerkmalen, die Erzeugung der Daten, welche in das System integriert werden, die Qualität des Systems zur Verwaltung großer Datenmengen, die Komplexität der verwendeten Daten, die Merkmale der graphischen Benutzerschnittstelle und die Qualität von definierten Bodenbedeckungstypen.

Contents

List of Figures

List of Tables

1

Introduction

Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

T. S. Eliot [The Rock, 1934]

How to measure the performance of a remote sensing image information mining system? How to organize an overall retrieval system evaluation procedure and what has to be included to obtain reliable and meaningful results? In recent years, many new developments and tools have entered the field of accessing digital images in large databases by their visual content. Most of the content-based image retrieval systems have been equipped with functions to extract relevant visual information from the data, to generate a content-index and to realize the search process in an interactive way based on a graphical human-computer interface. As in any other field, however, further progress depends on the ability to evaluate the image mining and image understanding functions and methods. If someone wants to compare individual functions or the overall performance of a system with others, problems will occur since most image retrieval systems are either not or only very sparsely validated. Evaluation is one of the most neglected fields in content-based image retrieval: there is no common test-bed for testing and verifying and most attempts are restricted to precision/recall graphs and related quantities. Precision/recall graphs, which have originally been developed in textual document retrieval, are less suited for content-based image retrieval since they do not allow us to assess the performance of individual system modules. They only validate the retrieval capability according to relevant and irrelevant images that the search results contain.

In addition to the performance validation of a content-based retrieval system, we further face rather complex remote sensing image data that are characterized by a high diversity of structures and objects. There exist several intelligent systems that provide novel methods to access the contents of this special kind of data. An approach that proved to be useful in several applications models the image data in a hierarchical Bayesian way. First, primitive features and meta-features are extracted from the image data and then clustered in an unsupervised and application-free

way (Datcu et al. 1999). From the clustering we obtain a vocabulary of characteristic signal classes that is valid across all images in the archive. User-related and application-specific semantic cover-type labels can be defined based on this vocabulary. The scheme and the implemented system provide the basis for the evaluation performed. The elements at a certain level in the Bayesian hierarchy are obtained from the elements of the level(s) below in a probabilistic way. That's why this arrangement enables us to determine both the information content of individual levels and the information flow between different levels using information-theoretic measurements.

After this short introduction into content-based image retrieval and evaluation, we will give an overview about techniques to access digital data by their content in Sec. 1.1. Although this short review does not comprehensively reflect the situation in content-based retrieval and evaluation, it gives the reader a first impression of the complexity of this field. In Sec. 1.2, we deal with existing problems and the reason why we verify our retrieval system. In Sec. 1.3, we briefly sketch out the basic evaluation concept that will be explained and applied throughout this dissertation. We conclude this chapter with an overview of the organization of this thesis in Sec. 1.4.

1.1 Content-based Retrieval of Digital Data

A characteristic element of today's 'information society' is the ability to interactively communicate with a variety of new digital media. The performance of a communication system is equivalent to the computational performance that has resulted in fast processors and large data repositories to convert and store the massive volumes of textual, audio and visual information (Neuman 1991). With the introduction of the Internet the digital data are not only characterized by local networking, computation and storage capacity but also by the loss of geographical restrictions. While it is fascinating to immerge in unlimited sources of data, it can at the same time lead to considerable resignation and frustration. Therefore, one challenge is not only to improve the fast access and storage of large data quantities, but also to measure the value of the data by extracting meaningful and relevant information from it.

The state-of-the-art database management systems can only process structured data in alphanumeric form. These data are characterized by the fact that the information is in the database tables, that is the data samples themselves. With the newly established discipline of data mining a number of methods are available for information extraction and manipulation: filtering, sorting, classification, retrieving and summarizing of data content. The information extraction approach changes for unstructured data types such as text, images, video and other multimedia. A common method to describe the content is to annotate metadata or textual descriptions provided by humans. But data like sound or images can have multiple interpre-

tations and since the annotated keywords usually reflect more the preferences and intentions of the person that produced them, text annotation is not a suitable technique for handling large datasets. A more reliable approach is to directly describe the data with regard to content-based attributes.

Text document retrieval

In the time of Internet most users regularly apply search engines like Google or Yahoo as a convenient way to look for specific text documents. To efficiently search and retrieve relevant documents from large databases, most engines use a pre-generated content-index that has automatically been computed (Schäuble 1997). In the case of text retrieval the meaning of the document is given by the content itself. Semantic ambiguities of the index, however, require exact queries to obtain the desired results. Despite exact queries, search engines usually return quite a high number of results, where only one half of the atop 20 top-ranked ones is relevant for the user (Casasola 1998). In order to increase the efficiency of search engines, intensive research efforts have been made to incorporate the user's behaviour in the retrieval process. Hereby, the analysis of human-computer interactions has been at the center of interest.

Image database retrieval

Experience has shown that classical image file text annotation does not meet the requirements that visual content representation and retrieval from large image repositories demand. To avoid text annotation, systems for content-based image retrieval (CBIR) have been implemented. This way users can directly formulate their queries with regard to visual content parameters. A standard procedure for CBIR is accomplished in an interactive way: the user selects image examples, the retrieval system proposes the best matches from the database and the user provides another set of examples. The advantage of an interactive CBIR system is that the user gets involved and can guide the retrieval process by making decisions based on visual similarity. However, a system does not always retrieve images the user is searching for since there is a gap between visual and semantic similarity. As a consequence, key issues of content-based image retrieval are to find visual attributes that perform well for image matching and to learn from the user's interactions in order to optimally incorporate the user's feedback in the retrieval process.

The development of CBIR systems started about 20 years ago. Some representatives of early operational systems are IBM's Query By Image Content (QBIC) (Flickner et al. 1995), Excalibur's RetrievalWare (Dowe 1993), MIT's Photobook (Pentland et al. 1996) and FourEyes (Minka and Picard 1997). The number of experimental and operational systems implemented in computer vision and multimedia groups shows that content-based image retrieval is still quite an

intensive research area. An overview of existing systems and current research directions can be obtained via Internet by the Viper project page (Computer Vision Group 2003), by one of the comprehensive reviewing articles (Smeulders et al. 2000) or books (Veltkamp et al. 2001). With contributions by pattern analysis and machine learning many advances have been made including the extraction of visual image features, the indexing of multidimensional feature vectors and system design (Rui et al. 1999).

While the size and the information content of images has continuously increased, CBIR with its global feature representation has not been satisfactory anymore. As a consequence, region-based image retrieval (RBIR) has been developed. The approach aims at the segmentation of each image into individual regions that are indexed by local characteristic attributes to make possible a more detailed description and interpretation of the content. Systems that offer region-based query functions with an automatically performed segmentation are 'Blobworld' (Carson et al. 1997) and 'Netra' (Ma and Manjunath 1999).

Both CBIR and RBIR are computer-centered approaches and the concepts only partly adapt to the user's needs. Image retrieval systems have been equipped with relevance feedback functions to incorporate the user's feedback during interactive learning and to search images similar to the user's conjecture. MIT's FourEyes successfully implemented user feedback in the retrieval process using a selection and combination of feature grouping. The method refines the user's interaction and enhances the quality of the queries.

1.2 Problem Definition and Motivation

The most frequently applied measurements for system effectiveness evaluation are precision/recall (PR) graphs or quantities like false alarms or visual inspection of the queried top-ranked images. In this frame, a typical practice for retrieval effectiveness verification is just to query the system. The query delivers a set of highest-ranked images that are either relevant or irrelevant. If a retrieved image is relevant according to the query, it is answered by relevance judgements (or ground-truth information), which is by far the weakest point in this validation method. The lack of ground-truth can be overcome in the same way as it is solved in text document retrieval (Harman 1992): visual inspection and keyword annotation. This approach works well for small or medium datasets, but is intractable for large volumes of data with several millions of images that remote sensing archives contain today.

In remote sensing, retrieval effectiveness measurements should not be focused on counting relevant and non-relevant images in the search results. Instead, a more significant assessment is to analyze the precision of the coverage in each individual image. Similar to PR graphs, this approach requires ground-truth cover maps that are hardly available and, if so, only for small and connected test areas. Therefore,

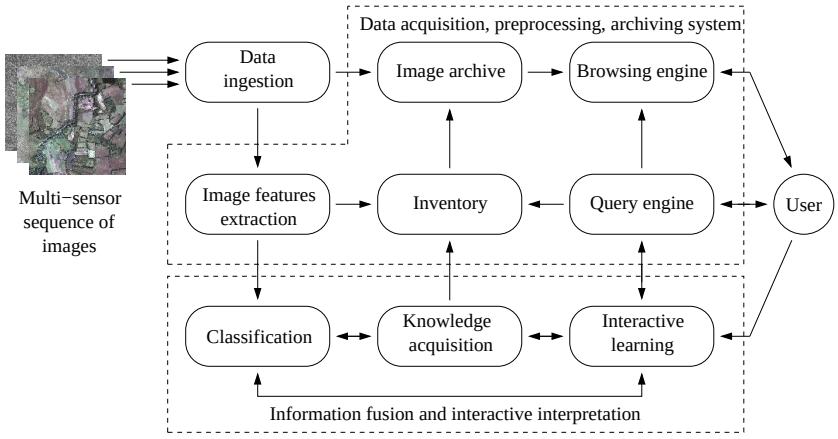


Figure 1.1: Architecture of DLR's image information mining system for interactive learning and probabilistic retrieval of remote sensing image content. The system is composed of a computer and a data intensive off-line part for visual feature extraction and indexing, and a fast on-line interface where the user learns the system based on the generated content-index. Considering the system architecture with different modules and interrelations, an overall evaluation is a complex task and the subject of this thesis.

realizing this retrieval performance test depends on whether ground-truth can be provided or not.

From the technical point of view, the evaluation of a content-based image retrieval system should not remain at the level of precision/recall graphs and related quantities. These measurements only yield information about the system query function and do not include properties that are characteristic of a modern image mining and retrieval system such as: the time to extract visual attributes like color, texture and shape, the capability of these features to represent structures and patterns in the image, the quality of the generated content-index, etc. Since modern retrieval systems are based on a server-client architecture where the human interacts with the computer by a graphical user interface (GUI), communication and information representation methods are a matter of evaluation, too. Meaningful measurements could be the time-span of learning the system and its functions, the time required to achieve goals on benchmark tasks, the error rates and time-retention during the use of the interface (Baeza-Yates and Ribeiro-Neto 1999).

However, the probably strongest performance evaluation is the user's judgment of the retrieval system. But such a performance test is not an easy task since many expert users from different application fields have to be included in the experiments to obtain reliable and statistically correct results. Furthermore, tests involving humans are usually hard to carry out, subjects must be carefully selected and experiments must be well-designed in order not to influence the outcomes and to shift them in

the desired direction. Large-scale tests involving many participants also put high requirements on the organization: a high number of tests has to be performed in a comparatively short time, results have to be recorded and the retrieval system must be kept stable during the experiments.

Altogether, an overall system evaluation concerning objective and subjective issues is a rather complex task. This dissertation aims at the development of methods and algorithms for system validation and system upgrade to achieve an optimal performance. Additionally, we measure the relevance of the proposed methods from the user's point of view in various applications. The complete system evaluation procedure is based on the Image Information Mining (I^2M) system of the German Aerospace Center (DLR) as depicted in (Fig. 1.1). Owing to the complexity of the system, the overall system performance cannot be represented by just one single measurement. The evaluation concept includes specific methods to determine both the technical quality and complexity of the system, e.g. information content of primitive image features and semantic cover-types, and (subjective) user-conjecture. We will not only verify individual system modules but also the man-machine interactions and the information flow during system operation.

1.3 Basic Evaluation Concept and its Positioning

From the application point of view, image information mining is quite close to content-based image retrieval, and thus, the concept of this dissertation for system performance evaluation belongs to this domain. However, in contrast to standard multimedia retrieval systems we deal with remote sensing data that need particular treatment. Whereas in standard multimedia applications the aim is to derive high-level semantic concepts from the image content, e.g. 'playing kids in a park', the content of remote sensing data is characterized by its high diversity of objects and structures at different scales and thus requires a complete description of the scene.

To achieve this goal, a concept for modeling the content of remote sensing images in a signal-oriented way has been developed, implemented and successfully tested. In a first step, an unsupervised content-index is generated from pre-extracted primitive image attributes. Later, this index is linked to the user's interests and allows him to query for relevant images in the archive without actually browsing them. According to this concept for representing the image content at multiple levels of different semantic abstraction, we organize the evaluation procedure in the same hierarchical way; we start the evaluation with an analysis of the quality of the image data and determine the information content of primitive image features and content-index. This analysis is based on techniques from statistics and statistical pattern recognition.

However, for the performance evaluation of the image information mining system, other scientific fields are important. With confusion or error matrices as a standard

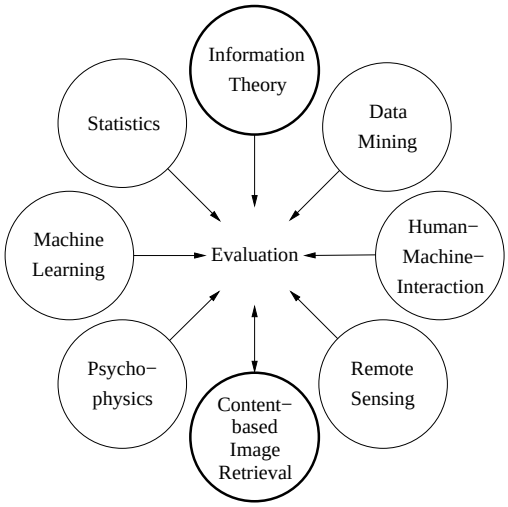


Figure 1.2: Positioning of the thesis' concept for system evaluation and the participation of several scientific fields. Whereas the most important input comes from information theory, the most significant part of the work is content-based image retrieval.

technique in remote sensing, we obtain the accuracy and separability of user-specific semantic cover-types. The quality of the stochastic link between (subjective) semantic labels and (objective) signal classes from the content-index is verified by using information-theoretic quantities. Since a semantic label is trained in several interactions between the operator and the mining system, psychophysical aspects are a matter of evaluation, too. Comprising, various scientific disciplines are involved in the overall system evaluation as summarized in (Fig. 1.2).

The main concern of this thesis is to obtain the quality of the mining system using measurements from estimation and information theory. In the Bayesian hierarchical representation of the image content (see Chap. 4 for details), the observed elements — image data, features and meta-features, signal classes, semantic cover-types and aggregated semantic labels — are regarded as random variables at each level and the process of information extraction is realized by estimating the parameters of the random process. In order to draw conclusions from data (observations at a certain level) to unknown parameters (elements of another level) that have to be determined, we apply stochastic signal models that are given in the mathematical form $p(X|\theta)$. These models express the probability of the data X conditioned on a particular parameter θ. In this model definition, different structures in the data are reflected by different values of the parameter. With Bayes' formula, the probability (likelihood) of the data $p(X|\theta)$ can be converted to the posterior probability $p(\theta|X)$ by using some prior information $p(\theta)$. It thus allows us to determine the parameter θ from the data X a posteriori.

Based on this stochastic modeling, the determination of the information content

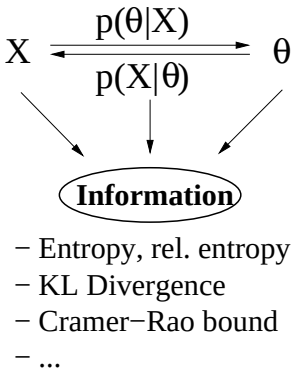


Figure 1.3: Basic principle of extracting and measuring information. From observations X, information in form of a stochastic parameter θ is extracted applying a parametric signal model $p(X|\theta)$. The amount of information or uncertainty contained in single random variables X and θ is measured by entropy, information between two different random variables over the same space by Kullback-Leibler divergence and the accuracy of an estimated parameter vector is reflected by the Cramér-Rao bound.

of observations X, estimated parameter θ and — if we consider X and θ as two discrete random variables indicating the input and output of a communication channel — the association between X and θ is evident as depicted in (Fig. 1.3). Since we assume all quantities to be random variables, we can apply measurements like the entropy $H(p_i)$ to determine the amount of information (or uncertainty) contained in a single probability distribution p_i, e.g. over the data X, Kullback-Leibler divergence $D(p_i; q_i)$ as the 'distance' between two probability distributions p_i and q_i, e.g. prior $p_i = p(\theta)$ and a posteriori $q_i = p(\theta|X)$, and the Cramér-Rao bound $\sigma^2_{\hat{\theta}}$ to indicate the accuracy of the estimate of parameter θ. Of course, information theory does not always provide the most suitable measures for system effectiveness assessment. To evaluate the system in terms of usability, questionnaires are the adequate method. On the other hand, most observations can be expressed as probabilities and information theory provides a number of powerful measurements that reflect the amount of information.

Now in this dissertation, we first explain the outlined principles of measuring information in detail, Chap. 3, and then apply them in the following parts: determining the technical system quality in terms of information content of extracted primitive image features (Sec. 7.1), content-index (Sec. 7.2) and semantic cover-type labels (Sec. 7.3 and 7.4), subjective evaluation aspects according to system operation (Sec. 7.6) and man-machine interaction (Sec. 7.7), and the association between elements of different system levels (Sec. 7.8). Chap. 8 describes the applied concept in a performed overall system validation procedure and Chap. 9 shows the executive summary of the main evaluation results.

1.4 Outline of the Dissertation

In **chapter 1**, we have given a brief introduction to content-based retrieval of digital data. The complex situation in image retrieval system performance validation and its necessity to have further progress has been described. Then, the applied evaluation concept of measuring the information content of stochastic quantities has been pointed out.

In **chapter 2**, we show characteristic techniques and concepts applied in content-based image retrieval and understanding. We discuss basic visual image attributes like color, texture and shape and how these features can be efficiently indexed. We sketch out methods for the fusion of different sources of data to improve the retrieval results and outline how high-level semantic concepts can be derived from low-level features. Since the visual content in terms of objective attributes does not always correspond to the user-related subjective image interpretation, the user's relevance feedback is incorporated in the retrieval process. Unfortunately, the system verification and validation, which is subject of this dissertation, is often a neglected topic in multimedia and remote sensing image retrieval.

The information-theoretic background of measuring the amount of information contained in probabilistic quantities, which is the core of the evaluation principle in this thesis, is outlined in **chapter 3**. We discuss probability, inference, basic measures of information, their main properties and relationships, and how they are applied to evaluate system components as well as the overall system performance.

In **chapter 4**, we explain the hierarchical scheme of modeling the image content using multiple levels of different semantic abstraction. This chapter summarizes the basic theoretical concepts of Bayesian image content representation as it is implemented in the information mining system. In this thesis, we arrange both evaluation methodology and procedure in the same hierarchical organization: image data, features and meta features, content-index, user-specific cover-types and aggregated semantic labels.

In **chapter 5**, we present a realization of the hierarchical modeling of image content: image information mining in remote sensing image archives. We outline methods to describe the content of both optical and radar image data and the generation of a vocabulary of characteristic signal classes that is valid across all images in the archive. Then, we assign user-specific semantic cover-type labels to this objective and application-free description of the image content by using simple Bayesian networks whose parameters are learned in several human-computer interactions. This learning paradigm is implemented as an on-line graphical interface that supports the user and continuously gives him relevance feedback about the learning progress. If a cover-type is trained, its stochastic definition can be used to query for images in the archive with a content similar to the label in a probabilistic way. We conclude this chapter with system configuration issues and practical applications.

The following four chapters present the main outcomes of this thesis:

- In **chapter 6**, we expand the demonstrated concept of modeling the image content at multiple levels of different semantic abstraction. Since the signal-based learning of image content does not allow the discrimination of complex image structures, we aggregate user-related labels in order to obtain higher-level semantic concepts. We discuss the algorithm for interactive learning this new kind of image content abstraction and show how it can be extended to index and query heterogenous collections of images. Finally, we outline the importance of domain ontology for the representation of semantic categories using individual user-specific cover-types.

- Then, in **chapter 7**, we present a methodology for the performance evaluation of specific components of the image information mining system and their interrelations. The methods aim at determining the technical objective quality, the complexity of the system and the subjective user conjecture. We measure the information content of primitive spectral and textural image features, feature space clusters, unsupervised content-index and semantic cover-type labels. Additionally, we evaluate human-machine interactions, the information flow between user and database during system operation and communication/knowledge representation aspects.

- In **chapter 8**, we demonstrate a procedure that combines individual evaluation measures defined in Chap. 7 to indicate the overall system performance. We describe the organization of a large-scale system validation, analyze the relevance and the correlation of the proposed measurements using a Karhunen-Loéve transform and compare the outcome with the user's degree of satisfaction.

- Experiments and results of an extended one-week system verification test are summarized in **chapter 9**. We show properties of the applied datasets, the efficiency of the off-line data ingestion chain, the scalability of the system to add large amounts of data, the complexity of image archives, the performance of the on-line graphical user interface and the quality of the defined semantic image content.

We conclude this dissertation in **chapter 10** with a summary of the main results and an outlook. Thereby, we outline relevance and limitations of the proposed approach and show further methods for system optimization and evaluation.

In **appendix A** we point out details of the Karhunen-Loéve transform and in **appendix B** we summarize variables and abbreviations used in this dissertation.

Experimental setup

Throughout this thesis, we show many examples and visualizations that reflect the performance of the implemented image information mining system. It can be ac-

cessed by everyone via Internet [1]. The methods and algorithms we apply in the evaluation procedure are not visible in the on-line version of the system, they are linked to the system via a tracing module that stores all performed human-machine interactions in the DBMS. Based on this information, a set of parameters is computed for each inserted user-specific semantic cover-type that constitutes the objective part of the evaluation protocol. In order to analyze the significance of these measurements and their correlations, we compute the Karhunen-Loéve transform and compare the outcomes with the subjective part of the evaluation protocol. The results are subject of a new mining function that after each system operation indicates to the user the learning progress. In the experimental version of the mining system coded in C++ and IDL [2], the new relevance feedback function is currently under investigation.

Publications

Some major parts of this dissertation have already been published in several conference proceedings and in two reviewed articles. The basic concept of the image information mining system has first been published in (Schröder et al. 2000) as interactive learning and probabilistic retrieval in remote sensing image archives. Then followed the first part of the knowledge-driven image information mining system article (Datcu et al. 2003), this time with emphasis on information representation, transmission, multi-mission data mining and system operation aspects. The system evaluation methodology, the procedure and the experimental results have mainly been presented in the second part of the knowledge-driven image information mining system article (Daschiel et al. 2003). Minor parts of this article were previously published in (Daschiel and Datcu 2002b) and a detailed description of the dyadic k-means clustering algorithm as briefly presented in (Datcu et al. 2003) is given in (Daschiel and Datcu 2002a). The evaluation of the information flow and communication channel view of the system was published in (Daschiel and Datcu 2003c). An extension of the evaluation of the mining system in terms of man-machine interaction, graphical user interface and information representation and communication aspects will perhaps be published in (Daschiel and Datcu 2003b). An article dealing with the enhancement of the mining system, namely the aggregation of user-specific cover-types labels, indexing of heterogenous collections of images and learning semantic categories, is currently under review as (Daschiel and Datcu 2003a).

[1] `http://www.acsys.it:8080/kim`

[2] Interactive data language

Part I

Preliminaries

2

Review in Content-based Retrieval of Image Data

The situation in content-based retrieval of image data is characterized by many systems and algorithms but only little attention has been paid to system verification and evaluation so far. No common benchmark test environment has been developed and established. Therefore, research groups use their own image collections, and performance effectiveness measurements can arbitrarily be selected to obtain good results. As a consequence, it is hard to compare image retrieval methods, select those image features that best represent the content, automatically evaluate the generated content-index and verify the user relevance (Smith 1998). Despite this lack of standardization, we will provide a detailed review of previous work in content-based retrieval of digital images. Here, the focus is on describing the applied techniques, positioning them in terms of performance evaluation, finally to arrive at a generic concept for content-based image retrieval systems.

In Sec. 2.1, we give an overview of methods used to describe the image content by primitive features like color, texture and shape. Additionally, we show the performance of features for content-based retrieval applications. How the produced large volumes of primitive image features can be reduced and indexed to improve the speed performance of a system is explained in Sec. 2.2. Since no individual feature can capture all structures and objects of a scene, a common way to increase the retrieval performance is to combine features as outlined in Sec. 2.3. An aim of content-based image retrieval — the derivation of high-level semantic concepts from low-level features — and the incorporation of the user's feedback in the query process in order to return images similar to the user conjecture is dealt with in Sec. 2.4 and Sec. 2.5, respectively. After Sec. 2.6, a section dealing with different aspects of image retrieval system evaluation and its functions, we describe the situation in remote sensing image retrieval from archives in Sec. 2.7. In Sec. 2.8, we will present a generic concept for content-based image retrieval systems and its importance for evaluation before we conclude this chapter in Sec. 2.9.

	system	features	evaluation method	reference
	QBIC	C, T, S	target images	(Niblack et al. 1993)
	Photobook	C, T, S		(Pentland et al. 1996)
	MARS	C, T, S	retrieval efficiency	(Ortega et al. 1997)
I.	PicHunter	C	target testing	(Cox et al. 2000)
	FourEyes	T	learning time	(Minka and Picard 1997)
	Netra	C, T, S	retrieval performance	(Ma and Manjunath 1999)
II.	Blobworld	C, T, S	precision-recall graphs	(Carson et al. 1999)

Table 2.1: Color (C), texture (T) and shape (S) image parameters applied in content-based image retrieval systems. The first category (I.) supports the search using global image attributes while the second category (II.) provides query mechanisms based on local features that have been computed for individual segments. The retrieval performance of features have been validated using different approaches.

2.1 Image Content Descriptors

How to characterize the content of digital images? Which descriptors best represent structures and objects in an image? To answer these questions, characteristics of the image data and the application to be designed have to be considered. In content-based image retrieval a large number of different descriptors have been applied. Simple color histograms were used in the early years and more complicated texture and shape representations in recent years. Now we give a general survey of applied techniques in image retrieval to describe the visual image content by color, texture and shape (Fig. 2.1).

Color

One of the first approaches to describing the scene content by color was the use of color histograms (Swain and Ballard 1991) and of color moments (Stricker and Orengo 1995). Color moments are a more robust version of color histograms since they combine the entire histogram information by low-order moments like mean and variance. The advantage of color histograms is that they can be efficiently computed. However, they lack spatial relationships with the effect that images with very different appearances may have the same histogram. An image with many distributed green pixels may have the same histogram as an image with just one green object, for instance. To overcome this obstacle, color coherence vectors (Pass et al. 1996) have been developed that aim at the distinction whether image pixels belong to large areas with uniform color or not. Based on this knowledge, each histogram bin is divided into two: one for coherent and one for incoherent pixels. With the incorporated spatial information, color coherence vectors demonstrated better query results than color histograms. A refinement of spatial color correlations from the previous

approach is shown in (Huang et al. 1997). The authors computed color correlograms from an image where they analyzed the spatial correlations between pairs of colors with a changing distance. In image retrieval experiments, it was shown that this method yields better retrieval results than color histograms and color coherence vectors (Ma and Zhang 1998). Recently, color invariant features (Gevers and Smeulders 2000) for image indexing and retrieval have been proposed. With the applied color models very high retrieval accuracies could be obtained since the models are independent of the illumination and the geometry of objects.

Texture

A feature that has proved to be rather important to characterize the image content is texture. This can be seen by the large number of papers and books dealing with it (Tuceryan and Jain 1998) (Reed 1993). Texture is used because it is not limited to single pixel values or certain color correlations, but considers spatial relationships in an extended neighbourhood, too. Therefore, spatial descriptors are important to describe complicated and extended image structures. Next, we summarize basic approaches for content description by texture that have been successfully implemented and tested in image retrieval.

In image analysis, simple statistical measurements can be directly derived from the histogram, such as mean, variance or higher order moments. These features can be fast computed but they fail to discriminate individual image structures. Features calculated from second-order statistics (grey-level co-occurrence matrix) were first proposed by (Haralick et al. 1973) and were useful in various experiments. From the complete set of 14 different features, a texture performance evaluation (du Buf et al. 1990) showed that contrast is the most significant feature and should be preferred. To search images by their contents from databases, the variance feature from the co-occurrence matrix was applied and the retrieval performance was compared with the performance of other features (Aksoy and Haralick 1998b). The suitability of second-order statistics to separate texture in remote sensing images was tested by (Schröder and Dimai 1998). The outcomes proved that these features are weaker than Gabor filters or Gibbs-Markov random fields. Although co-occurrence matrices perform insufficiently, they have often been applied in image retrieval since they can quickly be computed and therefore offer a convenient benchmark.

Another method of deriving features for browsing and querying large image databases is to apply signal processing and filtering approaches. In (Manjunath and Ma 1996), 2D Gabor filters have been suggested to extract spatial information by using the mean and standard deviation of the filter response. To obtain structures at multiple resolutions, 4 scales and 6 different orientations within each scale were used. An experimental validation was performed based on the Brodatz texture database where the quality of extracted Gabor features was compared with the effectiveness of

pyramid-structured wavelet transform (Daubechies 1990), tree-structured wavelet transform (Chang and Kuo 1993) and the multiresolution simultaneous autoregressive model (Mao and Jain 1992). Experiments indicate that Gabor features yield the best retrieval performance by means of the percentage of querying correct patterns. Based on the results of Manjunath et al., Dimai (Dimai 1999b) proposed a rotation invariant texture descriptor using scale and orientation of tunable Gabor filters. The retrieval effectiveness of this feature was compared with Manjunath et al.'s non-rotation invariant descriptor and a rotation invariant texture descriptor based on Fourier coefficients (Tan 1998). The efficiency of this approach for content-based retrieval was experimentally tested with two different collections, one consisting of 1,000 and the other of 5,000 heterogenous images. Search precision indicates that the invariant feature provides better results than the other ones.

Motivated by a study of Rao and Lohse (Rao and Lohse 1993) that found out that the three most important dimensions of human perception are periodicity, directionality and complexity, Wold decomposition features (Picard and Lui 1994) were developed and implemented in MIT's Texture Photobook. The relevance of the proposed texture model is due to its robustness to image transformations and local inhomogenities that can occur in natural textures. In image retrieval experiments, the quality of the Wold decomposition in modeling Brodatz textures has been evaluated (Liu and Picard 1996) by comparing the performance effectiveness with shift-invariant principal component analysis (Picard and Kabir 1993), tree-structured wavelet transform, multiresolution simultaneous autoregressive and Tamura's (Tamura et al. 1978) modeling. In the experiments, the Wold model performed slightly more accurate than Mao et al.'s simultaneous autoregressive model as far as the average recognition rate is concerned. The approach in CANDID (Kelly and Cannon 1994) uses Laws' convolution kernels to extract textural features for each pixel from 152 pulmonary CT medical imagery and then models the features by weighted Gaussian distributions.

Texture models that have rarely been used in content-based image retrieval are Markov random fields (MRFs). In (Gimel'farb and Jain 1996), MRFs demonstrated to be important for content-based retrieval where the models have been applied to query images from the Brodatz texture database. The structure of multiple pairwise pixel interactions of MRFs is compared with grey-level difference histograms to match the query image to the database content. The validation experiment, however, was performed on the assumption that there are only homogeneous textured images in the database. Under these constraints, a retrieval accuracy of about 90% could be obtained. A model close to MRFs is the multiresolution simultaneous autoregressive model that has previously been mentioned. That this model is to be prefered is due to its capability to characterize texture by transformation invariant features within a varying neighbourhood size.

In this section, we presented a number of major approaches to characterizing texture that have been applied in image analysis and content-based retrieval. The

ones that have not been considered are fractals (Mandelbrot 1982), 3D texture (Leung and Malik 1999), syntactic tree representation (Fu 1982) and Voronoi tesselation (Tuceryan and Jain 1990). Fractals use self-similarity across scales to model natural surfaces and texture. To solve the problems that may arise with 3D effects, e.g. shadow and specularity, particular methods have been developed. In (Leung and Malik 1999), the authors model the texture by a vocabulary of prototype tiny surface patches with associated local geometric and photometric properties. With the constructed vocabulary of 3D textons, 3D texture can be recognized from multiple viewpoints as shown in a classification assessment example. Grammatical and tesselation techniques try to represent geometric structures in images, e.g. line segments or closed boundaries.

Shape

A feature that has attracted much attention in the content-based image retrieval community is shape. However, characterizing image content by shape has proved to be rather difficult in image analysis (Mumford 1987) and acceptable results could only be obtained under certain constraints. Unfortunately, these problems remain in content-based image retrieval applications. An overview of shape analysis and matching techniques is given in (Veltkamp and Hagedoorn 1999).

As one of the first retrieval systems, QBIC (Niblack et al. 1993) integrated shape to query images by their content. The shape features combine the heuristic shape features area, circularity, eccentricity, major axis orientation and a set of algebraic moment invariants. In order to express each shape as a binary image, all shapes must be non-occluded planar. The partitioning of images into significant regions plays an important role in content-based image retrieval. Owing to the high diversity of images in databases, segmentation algorithms should be robust, unsupervised, independent of human interactions and applicable to a wide range of applications. Starting with these requirements, Dimai presented a method to extract salient regions involving local and area based information (Dimai 1999c). The region growing segmentation algorithm makes use of a single edge evidence map that has been obtained from edge evidences for several features and scales.

Motivated that a comparison of shape similarities between two objects can be understood as analyzing the objects' deformations, MIT implemented the 'Shape Photobook'. Therefore, instead of using image correlations, Pentland et al. modeled the physical 'interconnectedness' of shape which resulted in the calculation of a so-called stiffness matrix. This matrix represents the relation of each object point to every other point. Similar to MIT's 'Appearance Photobook', the deformation relative to some base or average shape is given by the calculated eigenvectors of the stiffness matrix. If the eigenvector shape description has been determined, shapes can easily be compared just by analysing the amplitudes of the eigenvectors. In (Sclaroff 1997), an advanced approach of 'Shape Photobook' is presented

to query images in databases using strain energy from prototypes to characterize shape categories. Based on the objects' deformations and their decomposition by a Karhunen-Loéve transform (KLT), the shapes in the database can be ordered by means of inelastic deformations. Thus, instead of comparing a selected shape with all ingested shapes in the database, the search is reduced to a small number of representative prototypes. Another benefit of the proposed method is that through the performed decomposition, it is invariant to rotation, translation and scaling. Another approach that makes use of eigenshape decomposition for image retrieval is proposed by (Günsel and Tekalp 1998). The authors' implementation first models objects by selected boundary and/or contour points as a shape description with different levels of details and then ranks the objects/images in the database using shape-similarity in the eigenshape space. Again, the method is translation, rotation and scale invariant.

The image/object retrieval concept in (Latecki and Lakämpfer 2000) assumes that contours of objects are influenced by noise and segmentation errors. In order to avoid distortions and to preserve the appearance of the original contours at the same time, the shapes are approximated by a curve evolution. Processing the curve evolution does not depend on control parameters, and retrieval performance relevance is shown by experimental results.

Del Bimbo and Pala developed an elastic deformation of user sketches to solve the problem of retrieving images by shape similarity (Del Bimbo and Pala 1997). Their approach of shape similarity matching — obtained by optimizing the elastic sketch — is said to be close to human perception similarity and proves to be robust to distorted shapes. In their approach, the sketch is warped during the optimization process in order to adjust itself to the objects' shapes in the images and this information, together with the deformation energies, is used to query images in the database.

In the image retrieval and analysis system MARS, a modified Fourier descriptor (Rui et al. 1998b) was applied to support user queries based on shape. Rui et al. demonstrated the invariance of their method both in terms of geometric transformations and noise. They also compared its robustness and computational complexity with two other Fourier descriptors.

Since image segmentation is still a weak point of many shape extraction algorithms, (Nastar 1997) proposed a technique for content description and retrieval using just the image shape spectrum. His approach is invariant to geometric transformations and shows robustness to noise and occlusion. Instead of deriving shape features from segmentation results, the author considers the image shape as the local shape of the intensity surface of the image. After defining an image shape index as the quantitative measure of image shape, the image shape spectrum is constructed as the histogram of the index over the whole image. Nastar achieved promising results and demonstrated the retrieval performance of his method by applying various databases with human faces, vasaline bottles and medicine packs.

2.2 Content Indexing

In the last section, we presented methods of describing image content by color, texture and shape. This process of visual feature extraction usually produces large amounts of data that cannot be managed in practise, particularly, if local features are used instead of global or aggregated ones to provide subimage query capabilities. In both global and local approaches, features are vectors that represent interesting points at image or region level. If a user performs a query to the system, the system compares the query with all images/regions in the database by directly applying the pre-computed feature vectors. In order to find images/regions in the database that correspond to the query, each feature vector from the query has to be compared with all stored feature vectors in the database. This method of querying images in terms of their visual content is rather expensive, especially if the number of images is large and the feature vectors' dimension high. In literature, this phenomenon is called "curse of dimensionality".

Reducing the computational demands and making content-based image retrieval scalable to large data quantities, efficient multidimensional indexing techniques have been developed. One promising approach is to first reduce the dimensionality and then apply an efficient method for multidimensional indexing to avoid the time-consuming calculation of Euclidean similarity measures. In retrieval applications, the dimension of extracted feature vectors is normally quite high, however, the number of individual dimensions with discrimination performance (embedded dimension) is much lower. A common method of reducing the dimensionality of the feature space is to apply the Karhunen-Loéve transform (Duda et al. 2001). In (Faloutsos and Lin 1995), the authors proposed a fast algorithm for approximating the KLT and showed in experiments that the dimension of feature vectors can be substantially reduced without considerably affecting retrieval performance. One challenge in content-based image retrieval is to have a system that dynamically updates the transformed visual attributes when new images are ingested in the database. This function is supplied by the low-rank singular value decomposition incremental update method (Chandrasekaran et al. 1997) that performed a fast and stable computation of the Karhunen-Loéve transform. Having computed the KLT, it is advisable to control the obtained results in order to avoid losing information by a reduction below the embedded dimension. Motivated by the failure of the KLT to explore nonlinear structures in high-dimensional datasets under certain conditions, a locality preserving projection (LPP) method has been proposed (He 2002). He applied the algorithm to original datasets of 435 dimensions and compared the outcomes of KLT and LPP in terms of retrieval accuracy. In all experiments, LPP delivered better results than KLT.

After reducing the feature vectors to the embedded dimension, a further decrease of the computational complexity can be achieved by indexing the reduced but still high-dimensional feature vectors. At this point we want to point out that

indexing can be performed both at level of raw and reduced (embedded dimension) feature vectors. Multidimensional indexing started in the seventies with quad-tree and k-d tree methods originally developed for traditional database systems. Recent developments have yielded the tree adaptation method (Tagare 1997), for instance. One of the first systems that applied multidimensional indexing to enhance the speed of the query process was QBIC. Both color and texture features were indexed using R-trees and the 18-dimensional shape feature vector was indexed by R-trees after performing a Karhunen-Loéve transform. The best method for image retrieval applications to select from a set of tree indexing techniques was proposed by (Ng and Sedighian 1998). The authors begin with a dimensionality reduction followed by a validation test of different indexing methods and finally customize the most successful indexing technique. Despite some advances, most of the tree indexing approaches failed to fulfill the requirements and characteristics of visual content indexing for content-based image retrieval (Rui et al. 1999).

A more promising category of content-indexing techniques is to cluster the whole set of feature vectors into similar groups to obtain another set of characteristic features. Clustering, which is similar to a vector quantization process (McLean 1993), projects the raw or reduced feature vectors into another space of quantized feature vectors. After clustering, each feature vector is substituted by its assignment to one of the clusters. An overview of existing clustering methods applicable to image retrieval is given in (Jain and Dubes 1988) and (Arabie et al. 1996). The advantage of generating a "visual vocabulary" of characteristic groups by clustering is due to its lower computational costs for calculating visual similarity. In comparison to tree-based indexing approaches, clustering is not limited to Euclidean metric similarity measures. However, there are disadvantages, too. First, the question of how to select the number of clusters appropriately arises since this influences the retrieval performance. Although there exist some methods to deal with the detection of the optimal number of clusters, e.g. the Bayesian classification algorithm AutoClass (Cheeseman and Stutz 1995), they all require tremendous computational costs. Thus, they are only applicable for data mining purposes under certain conditions . Other disadvantages of clustering are that the process of partitioning points in the feature space into similar groups is rather expensive and has to be dynamically updated every time new images enter the database. A solution to the latter problem was proposed by (Charikar et al. 1997). Charikar et al. implemented and tested an algorithm that incorporates the incremental and dynamic update of the clusters and manages high-dimensional datasets. This approach was further developed in terms of query acceleration and refinement (Rui et al. 1997a). Content-based image retrieval systems that make use of clustering for content-index generation are MIT's FourEyes and DLR's I^{2}M. The former applies a hierarchical clustering algorithm to quantize visual attributes and I^{2}M uses a dyadic k-means algorithm. A variation of constructing a vocabulary of characteristic content-index by clustering is gridding. Although this approach is slightly less accurate than clustering, it proved to be fast

and robust for heterogenous image datasets (Lorenz 1996).

Next to clustering, image content-indexing using self-organizing feature maps (Kohonen 1989) [1] turned out to be effective for image retrieval. Self-organizing feature maps (SOMs) aim at the representation of all points in the high-dimensional feature space by points in a 2-dimensional grid or target space. This way distance and proximity relationships are preserved as much as possible. The goal of preserving the neighbourhood can be realized if each sample (node) in the input (feature) layer is related to each target (node) in the map grid with assigned weights. All weights associated to a particular node in the 2-dimensional map define the reference vector of this node. Then, image similarity measures are computed by comparing all the reference vectors with one input feature vector (query) and determining the node that matches best. To accelerate the search, Zhang and Zhong applied hierarchical self-organization maps to construct an index tree (Zhang and Zhong 1995). With this tree image similarities supplied by a nearest neighbor search can be computed. The authors' experimental results using the Brodatz texture database showed that self-organizing maps can increase retrieval performance. Another reason for the popularity of SOMs is the capability to efficiently combine different features, e.g. color and texture, and to improve retrieval performance as demonstrated in the following section.

2.3 Combination of Features

The last decade in content-based image retrieval with its numerous applications and systems has shown that no single visual descriptor contains sufficient discriminatory information to achieve acceptable search results. Thus, various attempts have been made to increase the accuracy of image queries by integrating the retrieval results based on individual features. This combination of multiple measures (also known as data fusion) can be performed either at raw feature vector level or at content-index level.

The aggregation of different feature vectors in a global vector has often resulted in high dimensional feature representations (Aksoy and Haralick 1998a). However, a pure concatenation can both decrease and improve the performance.

An approach that goes one step further was to combine individual features by weighting. In (Bach et al. 1996), a system is proposed that enables the user to select the weights assigned with color, color layout, texture and shape according to his interests. Another retrieval scheme that integrates color and shape by weighting was shown by Jain and Vailaya (Jain and Vailaya 1996). The accuracy, stability and speed of that system was evaluated. The authors found out that a combination of color and shape yielded a retrieval effectiveness of 99% in terms of images being

[1] Sometimes, self-organizing feature maps are called topologically ordered maps or Kohonen self-organizing feature maps.

queried within the top-ranked. Both algorithms mentioned lead to an enhancement in retrieval accuracy. However, the results depend on the weight factors which have to be set to values fixed a priori. A possibility to avoid this shortcoming is to determine the weights regarding the performance of individual feature-based retrievals as exemplified in (Liu and Picard 1996). The authors fused the ranks of different texture models using weights derived from measures of the texture periodicity. Berman and Shapiro argue that retrieval systems like QBIC and Virage just offer the ability to query with a weighted combination of features and do not provide queries like "match on colors, unless the texture and shape are both very close". To give the user an extended and more flexible vocabulary of similarity distance combinations, the authors suggest a set of operations, including addition, weighting, min and max (Berman and Shapiro 1999).

Feature combination for content-based image retrieval can be seen as a specific high-level classification problem to partition images in the database into semantic categories using primitive image attributes. Based on this classification, the features' discriminatory performance can be measured based on intra-class and inter-class distances. In (Vailaya et al. 1998), the authors applied a nearest neighbor classifier to first separate the entire dataset into city-landscape using 5 different features. As the edge direction coherence vector was identified as the strongest feature, it was used to successively divide 'city' and 'landscape' into 11 subclasses such as 'beach', 'mountain' and 'towers'. Finally, multiple 2-class classifiers were combined into a single hierarchical classifier to improve the performance. Vailaya et al. also combined binary Bayesian classifiers into a single hierarchical classifier as shown in (Vailaya et al. 2001). The major disadvantage of their approach is that each image is assumed to belong to exactly one of the semantic classes. Therefore, an extension to heterogenous image databases is of limited use.

In recent years, the pattern recognition and content-based image retrieval community has dealt with neural networks for feature selection and combination. The aim is to reduce processing time by choosing a subset from a collection of image descriptors while still preserving the entire discriminatory performance of the feature set. In (Haering and da Vitoria Lobo 1999), a back-propagation neural network is applied to classify deciduous trees in images with a subset of features. Hearing et al. analyzed the relevance of 51 image parameters obtained from seven different feature extraction methods: co-occurrence matrix, Gabor filters, fractal dimension, steerable filters, Fourier transform, entropy and colors. The best selection of the 13 image descriptors was found to combine features from each model and is almost as powerful as the complete set. Worth mentioning is the 75% time-reduction for the feature extraction process. In order to achieve a thorough evaluation, they compared their methods with linear, quadratic and eigenanalysis methods both for feature subset definition and classification. Neural networks have also been used for handwriting recognition. Oh et al. proposed two algorithms to first validate the class-separation of features and then to combine various attributes to obtain a new

vector with an improved discrimination (Oh et al. 1999) for each class.

Efficient methods in combining heterogenous features are Bayesian networks. An important fact is that they describe the data in a natural way, that means, they are based on conditional probabilities that are estimated from imperfect data. Furthermore, expert knowledge can be included through a priori probabilities. Bayesian networks are applied in many working fields, e.g. filtering junk e-mails, pattern classification and speech recognition. Inside the Bayesian framework, naive Bayesian networks (or naive Bayes classifier) have attracted much attention due to their low computational complexity. In (Schröder et al. 2000), naive Bayesian networks are used to link the user's interests, that is, the interpretation of remote sensing covertypes, to the image content. The conditional probabilities in the network started with uniform priors that were iteratively updated based on the user's feedback. After computing the posterior probabilities in the classifier, a complete image archive was searched for relevant images. Another application of Bayesian networks was presented in (Kumar and Desai 1996). Kumar and Desai segmented an aerial image and extracted features from it. Then they approximated the conditional probabilities for identifiable objects in the image using the histograms of features.

2.4 Semantics

Querying images by their visual content involves comparing the query image (or region/object) with all other images (or regions/objects) contained in the database. In contrast to text-based retrieval systems with relational databases where images are found by exact matches using SQL language, content-based query systems utilize similarity measures. Images are ordered according to the similarity distance to the target image and the highest-ranked are displayed as search results. Similarity measures normally include geometric or probabilistic calculations and can be performed at different levels of abstraction. At the first level, the image (raw data) level, one can use the correlation coefficient or Euclidean distance (Castelli et al. 1998). However, since both approaches require an approximated location at pixel level and the computation is rather expensive, they are only of limited benefit for image retrieval. Even advanced matching techniques, which use multiresolution image representations (Li and Chen 1996) or which are computed in the Fourier domain (Stone and Li 1997), still require a considerable amount of computation. The next higher abstraction level is characterized by extracted features that represent the appearance of the image content, like color, texture and shape, for example. Similarity calculation at this level demands the comparison of multidimensional feature vectors with the query feature vector. To avoid determining geometric distances in a multidimensional space, probabilities instead of geometric distances are often used. A method of facilitating similarity computing is to first reduce the dimension of feature vectors using clustering algorithms or neural networks (see Sec. 2.2) and then apply simi-

larities to the reduced data. Here, problems arise. If a user searches an archive for images that contain certain structures or objects, the database search relies on visual descriptors extracted from data. Due to discrepancies between visual and semantic similarity, problems can occur. This phenomenon is called "semantic gap" and can result in situations where the set of returned images only partly responds to the user's query, or even worse, has nothing to do with it at all (Smeulders et al. 2000). Because of these shortcomings, research tries to link higher-level semantics to data-driven features in different ways.

To semantically describe regions or objects in an image, one requires visual content descriptors in combination with a methodology. In the remote sensing domain, for instance, a common technique to associate ground cover-types with semantic labels is to provide training samples and then perform a supervised classification. To access images in art galleries, they are annotated by keywords or captions. Although these approaches reduce content-based access to text-based retrieval, they are both time- and cost-intensive. The expenses reduce the flexibility of image databases and automatically result in authorized systems. Additionally, semantic labeling by keywords seldom includes details, is subjective (annotator and user normally are different persons) and finally does not solve the problem of image information retrieval (Colombo et al. 1999).

However, there are systems that provide semantic-based retrieval functions even in a restricted context. The "SceneryAnalyzer" by Song and Chang focuses on the extraction of semantics from scenery images (Song and Zhang 2003). The authors extracted low-level features from images and modeled them to get high-level features. After classifying and clustering them, each cluster is associated with certain semantics. According to this approach, all images can be automatically annotated with category keywords, e.g. background, wave, sky, etc. For this type of images with its low content complexity and well-separated objects, Song and Zhang achieved good results. However, their approach is hardly applicable to other datasets. In (Castelli et al. 1998), a progressive framework has been developed that enables the user to specify the content of remote sensing imagery at three different levels of abstraction: image, feature and semantic. A typical query starts with a reduction of the search space by specifying metadata restrictions. Then, after defining and searching objects at image or feature level, semantic objects are produced by an automatic classification. To test the accuracy of the system, Castelli et al. defined several benchmark queries, each given by different metadata constraints. Then, the authors let each search run against all of the images in the database that satisfy the constraints. Another example system that enables users to search for semantics in images is presented in (Schröder et al. 2000). First, users can interactively train semantic cover-types based on an unsupervised hierarchical representation of the image content. Later all pre-defined labels can be used to search the archive for relevant data. This scheme was successfully tested for remote sensing images and is currently under development.

2.5 Relevance Feedback

So far, we have considered content-based image retrieval as a computer-centered approach close to computer vision and pattern recognition. In the early years of image retrieval, systems were characterized by the performance of fully automated queries and the attempt to select a single best visual attribute. Computer centric systems assumed that expressing high-level concepts (semantic image interpretation) by low-level features is easy for users. This mapping may work for certain cases but proved to fail in practice. Another reason for this failure is the subjectivity of human perception (Picard 1996) that occurs at different levels. Persons may select color or/and texture to describe the image content. While one of them may associate 'cities' with large buildings and crowded roads, the other person thinks about beautiful parks and green vegetation around the blocks. As opposed to the automated approach and motivated by its restrictions, modern image information retrieval systems work interactively and include the user in the retrieval loop. This technique uses the synergy of computer and human and has been labelled 'relevance feedback'. It aims at accelerating and specifying the search performance to optimally adapt to the user's requirements.

Relevance feedback has been originally developed as a re-formulation method for text document retrieval. The main idea is to present a number of queried documents to the user, let him evaluate them and mark the relevant ones (Pao and Lee 1989). Through this verification, important keywords or expressions (terms) that are annotated to the identified documents are selected. Their importance is enhanced and used in a next query iteration to come closer to relevant documents. Techniques for relevance feedback are the addition of new terms from selected documents (term selection) and the modification of term weights according to the user's feedback (term re-weighting) (Baeza-Yates and Ribeiro-Neto 1999). In comparison to post-processing strategies, relevance feedback has the advantages that the entire search process is partitioned into several smaller steps, the user can concentrate on marking documents as relevant or irrelevant and all the interactions are under control of the system.

Although relevance feedback methods resulted in an increasing performance for text-based retrieval systems, they cannot be applied without any changes to image retrieval. As stated in previous sections, manual keyword annotation is impossible for large image archives and failed to cover the entire image content. Furthermore, content-based image retrieval requires the interplay between user and system to a much higher degree than in traditional text-document search to guide the query in the desired direction (Smeulders et al. 2000). However, early relevance feedback methods attempted to link the image retrieval model to the term re-weighting model from text-based retrieval.

In MARS (Rui et al. 1997b), image feature vectors are converted to weighted-term vectors in order to utilize well established text retrieval methods. For this

conversion, Rui et al. proposed two approaches. The first one includes component analysis and inverse collection importance analysis to obtain both the relative importance of components within a feature vector and the importance of components across feature vectors over the whole image database. The weights are implicitly contained in the vector. The other method appends the weights outside the vector by a Gaussian normalization of each component and then updates the weight after each interaction. In experimental results the authors measured the retrieval precision during several iterations and demonstrated the efficiency of both approaches. Another retrieval system that incorporates the vector space relevance feedback model is PicToSeek (Gevers and Smeulders 1999). It retrieves images according to the user's given feature weights by finding those database images closest to the query image in terms of weighting. But PicToSeek goes one step further than MARS. With the help of positive and negative image samples it also learns from the user's feedback which features are the most important ones. The drawbacks of MARS and PicToSeek are that both require the specification of appropriate weights for the relevance feedback query formulation. Unfortunately, the original image query is preserved during all iterations, too.

Having a single specific feature and a fixed similarity measure, retrieval systems usually deliver poor search results since they do not adapt to the user's subjectivity and needs. Due to this limitation, relevance feedback is used to choose between various features and/or similarities to obtain the best retrieval performance. Systems like MetaSeek (Benitez et al. 1998) compare the rankings of database images based on different features or similarities with the user-related ranking of relevant images.

To address the difficulties faced by computer-centered approaches, a popular method for relevance feedback is to assign weights to image attributes and update them according to the user's positive and negative feedback. In (Aksoy et al. 2000), Aksoy et al. proposed a weighted distance approach where the weight factors are derived as the quotient between the standard deviations of feature values, once for the entire image collection and once for images marked as relevant by the user. Using this feedback model, the weights are independently and incrementally updated and applied to change the features' influence from iteration to iteration. The authors tested the approach on a collection of about 10,000 images and successfully improved the retrieval performance at 19% after the first iteration. Rui et al. addressed the problem of subjectivity in content-based image retrieval by a multimedia object model where each image was represented at different levels and updated by weights (Rui et al. 1998a). Each image was modeled at image raw data level, feature level and representation level for a specific feature. The overall similarity between the query image (object) and other database images (objects) is the result of linear combinations of their lower-level similarities with associated weights. After returning a set of highest ranked images, the user marks each one as 'highly-relevant', 'relevant', 'no-opinion', 'non-relevant' or 'highly non-relevant'. According to the user's feedback, the system independently updates the weights at each level so that

the query image corresponds better to the user's search category. The proposed relevance feedback algorithm was tested on two different image collections both in terms of search efficiency and effectiveness and delivered better results than the vector space model.

Visual features are context-dependent, often noisy and, consequently, no single model is able to satisfactorily link the user's interests to the image content. Thus, Minka and Picard (Minka and Picard 1997) suggested interactive learning using a "society of models" for accessing image databases. Instead of selecting and combining features in high-dimensional spaces, they applied data grouping with self organizing maps guided by an interactive learning process with positive and negative examples from the user. They found out that in contrast to conventional learning algorithms their approach with feedback samples is more helpful to the user and the performed training can be usefully integrated for continuous learning.

In (Cox et al. 2000), a Bayesian relevance feedback algorithm is presented to accelerate the search in large image collections. The learning paradigm models the user's behaviour in a probabilistic framework using several man–machine interactions. Having a model how a user would react to a certain target image, the PicHunter system infers the probability of the target image that the user wants on the basis of the entire history of performed actions. The probability for each image being the target was computed and updated after each iteration. In experimental results Cox et al. demonstrated that PicHunter increases the search speed 10 times in comparison to randomly selected images (Cox et al. 1996). Search efficiency is measured by the average number of iterations to locate the target image.

Another Bayesian relevance feedback learning algorithm that incorporates the user's positive and negative feed-back samples was presented in (Vasconcelos and Lippman 2000). In order to explicitly choose regions or pixels that are relevant for the search, Schröder et al. demonstrated a Bayesian learning algorithm based on the user's positive and negative feedback (Schröder et al. 2000). From their examples, probabilities are updated and with a simple Bayesian classifier a map is depicted that gives the users an intuitive feedback and helps to understand the training results. A recent application of Bayesian relevance feedback treats positive and negative feedback samples differently, extracts a feature subspace and progressively updates it using a principal component analysis (Su et al. 2003). The outlined approach provides fast query capabilities, requires only limited memory and increases the system query performance significantly.

So far we have described approaches that take relevance feedback at feature level into account but have not considered the semantics of images yet. As stated in the previous section, representing the image content by low-level features is often less important than the actual semantic content as compared to annotated keywords in text-based retrieval. Therefore, efforts have been made to include high-level semantics in relevance feedback for content-based image retrieval. In (Lu et al. 2000), both features and semantics are incorporated into a relevance feedback algorithm.

The links between database images and keywords are realized in a semantic network where the associated weights are updated after each user interaction. If no semantic information is available, content-based image retrieval is reduced to conventional low-level feature retrieval. Lu et al. showed an increasing retrieval performance with the integrated semantic knowledge on a large collection of images. However, the performance depends on the selected keywords and the initialized weights. Another method combines low-level features and semantics for relevance feedback using a semantic correlation matrix (Lee et al. 1999). Altogether, including semantics in relevance feedback is only possible under certain conditions and further research is necessary to provide robust solutions.

2.6 Evaluation of Image Retrieval Systems

In this chapter we gave an overview of content-based image retrieval, explained basic functions, recent advances and remaining problems. As in other research areas, further progress and the successful and useful application of image retrieval and understanding depends on the ability to evaluate retrieval methods and results.

Objective evaluation aspects

The beginnings of image database query evaluation came from text document retrieval and focused on precision-recall and derived measurements. Suppose a system returns A images for a given query, B images are relevant in the answer set and S is the total number of relevant images in the repository. Then, recall is defined as the fraction of relevant images which has been returned, $R = B/S$, and precision is given as the fraction of retrieved images which are relevant, $P = B/A$. Although precision-recall (PR) are the standard evaluation measurements in text document retrieval (Harman 1992), they are only of limited use for image collections (Müller et al. 2001). First, PR assumes ground-truth information to partition the complete archive into relevant and non-relevant images according to the user's query. This requires the semantic interpretation of images which is much more difficult to determine than in text-based retrieval. Another shortcoming of PR is that the order of the ranking of returned data is completely disregarded. To overcome this negligence, alternative measures are proposed, e.g. harmonic mean of recall and precision (Baeza-Yates and Ribeiro-Neto 1999), normalized rank sum (Stricker and Dimai 1997) and rank-difference trend analysis (Dimai 1999a). However, despite the limitations of PR measures, they are often used in research literature, especially for small and well-known collections of images (Ma and Zhang 1998).

With the growing complexity of content-based image retrieval — systems are composed of modules for feature extraction, content-indexing, data storage, user interaction and knowledge acquisition/representation — it is necessary to evaluate

system parts individually as well as their mutual dependencies (Smeulders et al. 2000). To achieve this goal, one first has to define particular methodologies (criteria) and appropriate measurements. Evaluation criteria may include user relevance, system stability, visualization of results, scalability to huge amounts of data, time for searching the archive, usability, etc. The second category measures the assessment of certain criteria, e.g. the required time to search the archive and to depict results. Most of the methods cannot be objectively assessed, by time for example, since the system operator is an inevitable part in the evaluation process. User-related assessment ranges from labeling relevant and irrelevant images in the data repository to make possible PR measures to questionnaires. However, subjective concepts are time-consuming, demand a strict organization to perform many experiments with many users and are hardly reproducible. Thus, evaluation is an unattended research area in content-based image retrieval.

Despite these limitations, a few approaches exist that go beyond PR and related measures in image retrieval evaluation. In (Jermyn et al. 2002), Jermyn et al. addressed the discussion of high-level evaluation methodologies, showed where they are suitable and where they are not, and analyzed image segmentation algorithms based on them (Shaffrey et al. 2002). The approach started with a database consisting of different levels of abstraction: image space, index space and semantic (query) space. Exchanging information can be interpreted as a mapping between the spaces themselves or as a mapping using probability measure spaces for those spaces. Then, so-called 'knowledge-scenarios' are built for system evaluation purposes that try to algorithmically express either spaces or mappings between them.

Subjective evaluation aspects

In order to get the degree of users' satisfaction in the Epic image retrieval system, a user-centered, task-oriented and comparative validation was undertaken (Jose et al. 1998). Jose et al. recruited 8 people as system operators with no a priori knowledge and confronted them with two version of Epic: one allows only spatial queries and the other textual queries. From a series of questionnaires using semantic differentials between search sessions, values indicating acceptability or user satisfaction could be derived. Final results demonstrate that the spatial-querying system performs significantly better than the other.

To explore and evaluate the PicHunter image retrieval system, Papathomas et al. made various psychophysical performance tests (Papathomas et al. 1998). In a first experiment they studied the importance of semantic information for the query, earlier user input and similarity measurements. Each experiment was organized in a way that the user had to search for a target image in the database. Experimental results demonstrate that the best performance could be reached with semantics in the queries. A second series of tests compared the time to find a target image with PicHunter and by random selection using a generated baseline for the dataset. The

outputs did not meet the expectations.

2.7 Retrieval of Remotely Sensed Images

In this section, we show existing techniques to retrieve remotely sensed images from archives and how the influence of content-based image retrieval has led to major advances.

Over the last decades imaging satellite sensors have acquired huge volumes of data. Optical, SAR and other sensors have delivered several millions of scenes that have been systematically collected, processed and stored. For instance, the DLR ground station at Oberpfaffenhofen, Germany, receives about 100 GBytes/day of imagery that results in an accumulation of 10^4 GBytes in the repository. The state-of-the-art systems for accessing remote sensing data and images in particular, make use of meta information, e.g. the DLR EOWEB satellite information service or the image retrieval system of Spot Image (Tab. 2.2). The applied meta information depends on the acquired data and contains coordinates of the satellite during the image acquisition and the covered ground. Additionally, radiometric and spectral properties of the sensor and the date of image take are stored. With this kind of information it is possible to perform a query such as: "show me all images that cover Oberpfaffenhofen and have been acquired by the Landsat TM sensor in year 2002". This information allows only constrained queries and, consequently, only little of the image content is actually used (Datcu et al. 2002). In the future, the access to image archives will even become more difficult due to the enormous data quantities acquired by a new generation of high-resolution satellite sensors. As a consequence, new technologies are needed to easily and selectively access the information content of image archives and to increase the actual exploitation of satellite observations (Datcu and Seidel 1999).

For content-based browsing and retrieval of remote sensing images, special properties of the data have to be considered: characteristic image size from 6,000 × 6,000 up to 24,000 × 24,000 pixels, resolution of less than 1m up to 1000m and a content with a high diversity of natural elements and man-made structures. Due to this complexity, image descriptors have to be carefully selected. In remote sensing, one of the well-established features is spectral. This descriptor demonstrated its benefit in many applications such as land use classification, for example, but cannot separate complicated image structures. Hence, spectral information was often used in combination with texture and yielded good results. The upcoming generation of high to very high resolution satellite sensors will produce image data that needs different content modeling, not just by spectral and texture. Geometrical (structural) features are better suited to describe completed objects in images.

Having a certain application in mind, a common method in remote sensing to query images by their content is to segment each image in a number of a priori known

	URL	system/organization
	http://www.spotimage.fr	Spot Image
	http://eoweb.dlr.de:8080	DLR/DFD
	http://www.nrsa.gov.in	NRSA India
I.	http://edcsns17.cr.usgs.gov/EarthExplorer	USGS geo data Explorer
	http://www.ssc.nasa.gov/∼sirs	NASA/SSC
II.	http://earth.esa.int/services	ESA
	http://www.alexandria.ucsb.edu/adl	Alexandria Digital Library
	http://www.vision.ee.ethz.ch/∼rsia/	ETHZ/DLR
III.	http://www.acsys.it:8080/kim	**DLR**

Table 2.2: Remote sensing image retrieval systems. The first group (I.) makes possible database queries with meta information such as sensor type and date of acquisition while the second group (II.) supports queries based on special content information obtained from case studies. The third group (III.) offers content-based image retrieval and thus makes the actual image content accessible for the user. The evaluation we address throughout this thesis is based on DLR's image information mining system (I^2M).

cover-types. Then, thematic maps are computed where the map indices correspond to the segmented classes of cover-types and the following queries are related to this pre-extracted information. Systems based on this concept have performed well because the content representation is only focused on a segmentation with exactly the given cover-types (Bertoia and Ramsay 1998). Otherwise, problems may arise whether the system is extended to different cover-types or to datasets from another sensor.

Apart from the traditional and application-specific methods of retrieving remotely sensed images, a few research groups and companies apply content-based query functions in their systems. In (Bretschneider et al. 2002), the authors present a system that offers image retrieval by metadata as well as content-based query functions. For the latter, an unsupervised classification was computed using a modified k-means clustering algorithm to extract the entire multispectral information. With just multispectral information, the authors obtained retrieval accuracies of about 80% for urban, farm land and forest cover-types. However, the results were obtained for selected queries and datasets. Barros et al. (Barros et al. 1995) first clustered a Landsat TM image based only on the spectral image values and then tried to query images using the spectral information of each region. For content-based search and clustering of remotely sensed imagery, Marchisio et al. applied multiple features in their GeoBrowse system (Marchisio and Cornelison 1999). The system offers the composition of image processing and mining functions with the statistical analysis software S-PLUS. Methods for image content representation by

features comprise segmentation, multispectral analysis, non-parametric regression and texture analysis. Spatial image content is described using statistical methods (grey-level co-occurrence matrix) and wavelets. In (Ma and Manjunath 1998), Ma and Manjunath demonstrated Netra, a system for browsing, searching and retrieving of remotely sensed images. The authors started partitioning database images into equally-sized blocks and then extracted Gabor texture features for each block. In experimental results, a retrieval performance of about 90% was achieved in terms of average percentage of retrieving similar patterns. Netra has been developed within the Alexandria Digital Library (ADL) project. Together with its continuation Alexandria Digital Earth Prototype (ADEPT), it aims at the use of the digital earth metaphor for organizing, using and presenting information at all levels of spatial and temporal resolution. Perhaps the most enhanced system for content-based image retrieval is DLR's Image Information mining system (Datcu et al. 1999) and its update KIM (Datcu et al. 2003). Starting with an unsupervised content-index generated from pre-extracted spectral and textural features, a user can train semantic cover-types of his interest by data fusion and search archives for relevant images. Currently, this concept is under further development and tested with various datasets. The evaluation methods that are presented in this thesis are based on the concept of this system.

2.8 Generic Concept

In this chapter, we presented methods that are implemented in experimental and commercial content-based image retrieval systems. Systems like QBIC, FourEyes and I^2M, for instance, are equipped with different functions to extract properties of image structures and objects, features are reduced, compressed and indexed to obtain computationally manageable data quantities, different sources of information are fused and relevance feedback functions are applied to return images similar to the user conjecture. Although content-based image retrieval systems are rather heterogenous according to the applied techniques, they can be grouped in a more general diagram as depicted in (Fig. 2.1). In order to assess the overall effectiveness of content-based image retrieval systems and to be independent of the system specificity, a certain evaluation methodology that is adaptable to the generic concept is required.

The standard method for image retrieval system evaluation using precision/recall measures allows only to analyze the query results in terms of relevant and irrelevant images. Although this approach allows to assess the retrieval performance of a system, it fails to reflect the quality of individual features, the effectiveness of indexing techniques and does not include subjective human factors at all. Consequently, this approach is not suited to performimg a detailed system evaluation procedure.

Since the information content of the image data at different levels of semantic

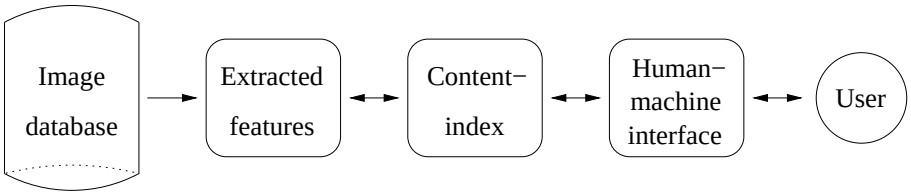


Figure 2.1: Generic concept for content-based image retrieval systems. Primitive visual features are extracted from images in the database and indexed. Based on the indexed features, the database can be queried for relevant images and the top-ranked ones are displayed as query results. Although this description presents a one-way data flow from the archive to the user, some systems are equipped with relevance feedback to include higher-level user concepts in the retrieval process.

abstraction in the generic scheme is implicitly contained in the applied models for feature extraction, content indexing and relevance feedback, it is necessary to provide an evaluation methodology that takes this fact into account. Information and estimation theory provides a number of powerful measures that can be applied for the verification of probabilistic retrieval systems: with mutual information we can determine the amount of information a random variable contains about another one, Kullback-Leibler divergence measures the 'distance' between two different probability distributions of the same random variable and entropy quantizes the amount of uncertainty of a single random variable. If the applied stochastic models are in parametric form with unknown quantities that have to be estimated from a limited number of observations, determining the accuracy of the estimates is a matter of evaluation, too.

With entropy, Kullback-Leibler divergence, mutual information and the accuracy of estimated parameters we have at hand all relevant quantities we need to analyze content-based image retrieval systems in a detailed way. Before we apply the proposed information-theoretic measures on the I^2M system in Chap. 7 and 8, we will first define them and their properties in the following chapter.

2.9 Conclusions

In this chapter, we have discussed the following items:

- Content-based image retrieval deals with the query of large image databases using visual attributes like color, texture and shape. To accelerate the speed performance of systems and to retrieve images similar to the users' concepts, various techniques for feature indexing, information fusion and relevance feedback have been implemented.

- An important topic in image retrieval that has not been given much attention

so far is system verification and validation. The applied approaches mainly stay on the level of precision/recall measures and thus are not suited for validating interactive retrieval systems in which the user is an inherent part.

- Methods for content-based image retrieval that have originally been developed to manage and explore the growing amounts of multimedia datasets have also been successfully applied in remote sensing. Whereas the standard technique for retrieving remote sensing data is limited in that it just allows the query based on metadata, content-based approaches are more helpful for users as they provide access to the actual information content of the image archive.

- Owing to the diversity of algorithms implemented in image retrieval systems and to provide an evaluation methodology that does not depend on the system specificity, a generic concept for content-based image retrieval was demonstrated. Based on this concept and the fact that information is in the applied signal models, the evaluation is equivalent to quantize the information content of the parameters and observations in each individual model. In the following, suitable measurements are described for validation and applied in subsequent chapters.

3

Probability, Bayesian Inference and Information Theory

In the preceding chapter, we gave a detailed overview of the state-of-the-art concepts and techniques in content-based image retrieval for low-level feature extraction, content-indexing, data fusion and learning higher-level semantics. Different criteria exist to classify the applied methods. One can group them according to efficiency (how much data can be managed), level of detail (global vs. local image interpretation) or model. For all steps of information representation one can apply different models that can be divided into two major categories: deterministic and stochastic. The first group of models is to be favoured if the data in question consist only of a systematic and noise-free component. In case that the data are the result of two components, a systematic and a random process, probabilistic models are usually preferred to avoid unpredictable results.

The image information mining system on which our evaluation procedure is focused fully incorporates the probabilistic approach where the image content is modeled in a hierarchical Bayesian way using several levels of different semantic abstraction. In the system, elements at each level — image data, features, content-index, individual and aggregated semantic cover-type labels — are considered as random samples. They are obtained in a step of Bayesian inference from one or more levels below using different stochastic models. In order to perform a validation that is adapted to the probabilistic nature of the system, a particular methodology and measurements are needed. A domain that provides suitable measures based on probabilistic quantities is information theory. With entropy we can determine the amount of information contained in the distribution of a single random variable, mutual information describes the amount of information that one random variable gives about another and Kullback-Leibler divergence reflects the discrimination between two different probability distributions. In addition to entropy and its related quantities, Fisher information represents the accuracy of a parameter estimated from a limited number of observations. Before we explain these measurements in more detail in Sec. 3.3, we will first give a short introduction to probability in Sec. 3.1 and to basic principles of Bayesian inference in Sec. 3.2. In Sec. 3.4, we conclude

this chapter with a summary of the main results and contributions.

3.1 Probability

As far as the notation of 'probability' is concerned, scientists do not share the same view. We refer the interested reader to (Cox 1946), (Bernoulli 1713) and (Bayes 1763) for a deeper insight into probability and its definitions. The most well-known view is certainly the 'frequentist' one. It defines the probability $Pr(X|H)$ as the relative frequency or occurrence of the event X conditioned on some assumptions or causes H. Laplace and Bernoulli considered probability from another point of view. They defined the probability $Pr(X|H)$ as the quotient of the range measurement of X and H and the range measure of H solely. Opposed to these definitions is the approach of Bayes who considered $Pr(X|H)$ as the measure (or degree) of certainty (or belief) that X follows the cause H. The Bayesian view is sometimes also called 'Subjective' since it depends on the selection of prior knowledge about the degree of belief in the cause H.

In the following, we do not intend to stimulate a long discussion about objective and subjective aspects of probability, instead, we consider probability as an objective measure of certainty that is sometimes combined with some prior information and therefore may seem subjective.

Consistency postulates

To ensure logical consistency for calculations that include probabilistic measures, the following four postulates have to be fulfilled (Papoulis 1984):

1. Positivity

$$Pr(X|H) \geq 0 \tag{3.1}$$

2. Certain event

$$Pr(X|X) = 1 \tag{3.2}$$

3. Sum rule

$$Pr(X|H) + Pr(\neg X|H) = 1 \tag{3.3}$$

4. Product rule

$$Pr(X, Y|H) = Pr(X|Y, H)\, Pr(Y|H) \tag{3.4}$$

Here, X and Y indicate two different events as results or outcomes of random experiments, $\neg X$ means not event X and some conditions or hypotheses are denoted by H. With the four postulates we have the basic algebra of probabilistic mathematics at hand from which all the following equations can be derived.

Bayes' formula

From the product rule, Eq. 3.4, we can directly infer the important Bayesian formula

$$Pr(H|X) = \frac{Pr(X|H)\,Pr(H)}{Pr(X)} \ , \tag{3.5}$$

that shows that the probability $Pr(X|H)$ of the observed data X, e.g. the pixels of an image, given some hypothesis H, can be inverted to $Pr(H|X)$ including some prior information. In Eq. 3.5, $Pr(X|H)$ is called the likelihood of H with respect to X, which indicates that the hypothesis H for which $Pr(X|H)$ is large is more 'likely' to be the true hypothesis. The quantity on the right hand side of the likelihood term, $Pr(H)$, is called the prior probability and reflects the prior belief in the hypothesis H. In order to determine the posterior probability $Pr(H|X)$, the most important factors are the likelihood and the prior probability; the predictive probability or evidence, $Pr(X)$, only acts as a normalization constant and guarantees that the posterior probability adds up to one. The evidence is given by the likelihoods and priors of all hypothesis H_i as

$$Pr(X) = \sum_i Pr(X, H_i) = \sum_i Pr(X|H_i)\,Pr(H_i) \tag{3.6}$$

and indicates that small variations in the hypothesis of only one likelihood or prior may lead to significant changes in the posterior. After explaining the different terms in Bayes' rule, we can summarize the rule in an informal expression by saying that

$$posterior = \frac{likelihood \times prior}{evidence} \ . \tag{3.7}$$

The importance of Bayes' theorem for image processing and content-based retrieval, respectively, is due to strong stochastic signal models that can either be used to extract information from the data or to incorporate knowledge as prior information. In this dissertation, we will use stochastic models at various levels of image information representation and in different contexts. We will apply stochastic texture models on optical image data to capture primitive image features from the data and to reconstruct the content of speckled radar images by incorporating prior information in form of a texture model. At a higher level, stochastic signal models connect the semantic image interpretation to image feature models, e.g. 'mountain' is connected to spectral and texture content-index. Elements at the highest level of semantic abstraction, e.g. 'a river in a hilly terrain next to a city' are again associated with stochastic models to individual semantics, such as 'river', 'hilly terrain' and 'city' in this example.

Probability and probability density function

Assuming that a random variable X has an arbitrarily large number of N possible outcomes with $N \to \infty$, $Pr(X|H)$ is no longer a discrete probability but rather a

continuous probability density function (pdf)

$$p(X|H) = \lim_{\Delta x \to 0} \frac{Pr(x \leq X \leq x + \Delta x)}{\Delta x} \ . \tag{3.8}$$

If the random variable X is inside the interval x and $x+\Delta x$, we obtain the probability for X as

$$Pr(X|H) = \int_{x}^{x+\Delta x} p(X|H)\, dX \ . \tag{3.9}$$

Since we use the probability distribution over discrete states of a random variable in later chapters and in order to guarantee consistency between the discrete and the continuous state, we denote hereafter everything related to probabilities by $p(\cdot)$.

Elimination of nuisance parameters

A reason for the popularity of Bayesian formalism is the possibility to eliminate parameters that are of no interest for the data analysis. This has been labelled the "removal of nuisance parameters". Consider Eq. 3.4 with parameters X and Y and some hypothesis H. If we want to remove an unwanted parameter, e.g. Y, we can just do this by the integration

$$p(X|H) = \int p(X, Y|H)\, dY \tag{3.10}$$

over the complete space of Y. However, this marginalization results in a loss of information since all knowledge contained in the parameter Y is lost.

3.2 Bayesian Inference

In the previous section we pointed out how to infer the posterior probability with Bayes' formula from the likelihood and some additional prior information. The difficulty in modeling complex data, particularly to describe the content of images, is to find a suitable model for the observed data X. Usually these models are expressed in their parametric form $p(X|\boldsymbol{\theta}, M)$ and are uniquely determined by the values of the parameter vector $\boldsymbol{\theta}$ assigned to a certain model M. If it is obvious which model is concerned, the symbol M is often neglected. For instance, we can model data to be Gaussian, $p(X|\boldsymbol{\theta}) \ \sim \mathcal{N}(\mu, \sigma^2)$, with the parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)$. The probability $p(X|\boldsymbol{\theta})$ of the data X conditioned on the model parameter vector $\boldsymbol{\theta}$ is sometimes called the likelihood of $\boldsymbol{\theta}$.

3.2.1 Parameter Estimation

Information extraction by model parameter estimation, also called first level of Bayesian inference, deals with inferring the values of the parameter vector $\boldsymbol{\theta}$ from

the observations X that are considered as random variables, e.g. pixels of an image. There are two categories of estimation problems. The first category is when the quantity to be estimated — our parameter vector $\boldsymbol{\theta}$ — is deterministic but unknown. The second category is when the quantity to be estimated is a random variable. In the following, we describe both estimation techniques before we briefly demonstrate Bayesian model selection in the next section.

Maximum likelihood estimation

Suppose that a dataset X consists of N samples $x_1, x_2, \ldots x_N$ and let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ denote a parameter vector with l components. To find values for the elements of the parameter vector $\boldsymbol{\theta}$ that make the given samples of the dataset X the most likely ones, we have to maximize the likelihood $p(X|\boldsymbol{\theta})$ as (Duda et al. 2001)

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta}) \ . \tag{3.11}$$

If we consider an infinite large number of possible states instead of discrete values for a random variable, we go from discrete probabilities to a continuous probability density function.

Maximum a-posteriori estimation

In contrast to the ML estimator, the maximum a-posteriori (MAP) estimator aims at the maximization of the posterior probability $p(\boldsymbol{\theta}|X) \propto p(X|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ and yields

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|X) \\ &= \arg\max_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \ , \end{aligned} \tag{3.12}$$

where $p(\boldsymbol{\theta})$ denotes the prior probability of different parameter values. Comparing both estimators from Eq. 3.12 and 3.11, we see that the ML estimator is a MAP estimator with a uniform prior.

Comparison of ML and MAP estimation

In the case of infinite training samples, $N \to \infty$, and a selected prior distribution that does not affect the outcome, maximum likelihood and maximum a-posteriori estimates deliver identical results. However, since the number of data samples is restricted in practical situations, one has to ask in which cases the two estimators yield different solutions and which technique is preferable. For this decision, a number of criteria exist. First, from the computational point of view, ML techniques are often preferable since they are based on first and second order derivatives whereas MAP approaches can result in a time-consuming multidimensional integration. For

interpretability reasons, maximum likelihood estimates are calculated from one single model. Thus, they are easier to understand than maximum a-posteriori results as the weighted average of two models that can also have different functional forms.

The difficulty in maximum a-posteriori estimation is the choice of a suitable prior distribution which is often criticized. If the prior and its parameters are not correct, the estimation convergence may be rather slow or the results may even be shifted. On the other hand, if the prior is carefully chosen, it can significantly improve the convergence. To achieve this goal, a common way is not to take a fixed prior but rather to derive it from the underlying data.

3.2.2 Model Selection

In the previous section, we outlined how information is extracted from data in the Bayesian way. We estimated the parameters of stochastic models incorporating some prior knowledge. As noted, Bayesian formalism can also be applied to select appropriate models from a set of candidate models, known as second level of Bayesian inference. Let's consider a set of models $M_1, M_2, \ldots, M_T$ that represent the data X. Then, we can compare the different models by calculating the posterior probability for each model M given the data X and obtain with Bayes' rule, Eq 3.5,

$$p(M|X) = \frac{p(X|M)\,p(M)}{p(X)} \quad . \tag{3.13}$$

The term $p(M)$ indicates the prior of the model M, $p(X|M)$ the likelihood of the data X given the model and $p(X)$ the prior predictive. If all models are assumed to be equally likely, that means no one is preferred, the prior is given by the uniform distribution $p(M) = \frac{1}{T}$. Analogous to Eq. 3.12, where we neglected $p(X)$, we can express the posterior probability of the model through the likelihood times prior of the model as

$$p(M|X) \propto p(X|M)\,p(M) \quad . \tag{3.14}$$

Notice, that the term $p(X|M)$ has already occurred in Bayes' formula as the prior predictive or evidence and was not taken into account at that point since it only acted as a normalization constant. However, at this point the determination of the evidence is the core in comparing models and has to be obtained via the integration

$$p(X|M) = \int p(X|\boldsymbol{\theta}, M)\,p(\boldsymbol{\theta}|M)\,d\boldsymbol{\theta} \tag{3.15}$$

over the complete parameter space of $\boldsymbol{\theta}$. The calculation of the evidence, the integration of the product likelihood times prior, is non-trivial and can hardly be performed in practical applications. Particularly in the context of multidimensional datasets and rather complicated probability distributions, the integration in Eq. 3.15 over all the parameters in the model is not manageable. Therefore, one solution is to approximate the evidence as the maximum of $p(X|\boldsymbol{\theta}, M)\,p(\boldsymbol{\theta}|M)$ multiplied by its width (MacKay 1991).

3.3 Measures of Information

In this section we outline the basic concept of information and give some basic properties that are essential to guarantee consistency. The best known measures of information are Shannon's entropy, relative entropy, mutual information and Fisher's definition of information (Blahut 1987). Some of the concepts which are involved in the notation of information and which we present later are statistical entropy, uncertainty, coding, questionnaires, statistical independence, probabilistic distance and discrimination ability.

3.3.1 Entropy

Shannon suggested the concept of entropy (Shannon 1948) as the measure of the average amount of information (or uncertainty) of a single random variable. Let X be a discrete random variable with states $x_1, x_2, \ldots, x_n$ and probability distribution $p(x)$. Then, the entropy of X is given as

$$H(X) = -\sum_x p(x) \log p(x) \ , \tag{3.16}$$

where the logarithm may be either to base 2 or base e. When the logarithm is to base 2, the average information is measured in bits and if the logarithm is to base e, $H(X)$ is measured in nats. In the limit that one probability $p(x_i)$ vanishes, the contribution to entropy $H(X)$ is defined as $\lim_{p(x_i)\to 0} p(x_i) \log p(x_i) = 0$.

Properties of entropy

As when working with probabilities, entropy has to fulfill the following basic properties to guarantee consistency during all calculations (Blahut 1987):

1. Continuity:
   Small changes in the probability distribution $p(x)$ do not produce large changes in the average information $H(X)$.

2. Positivity:
   $H(X) \geq 0$ and $H(X) = 0$ if and only if all the $p(x_i)$ except one are equal to zero.

3. Extremal property:
   For a given J, $H(X) \leq \log J$ and $H(X) = \log J$ if and only if all occurrences of X are equal to $1/J$. This property shows that the entropy reaches is maximum if all states of X are equally likely.

4. Entropy is a concave function of X.

Joint and conditional entropy

Suppose we have another random variable Y. Then, similar to Eq. 3.16, we can define the average information in the joint space (X,Y) as

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) \ . \tag{3.17}$$

We can further define the conditional entropy that is the entropy of a random variable given another one as

$$H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y) \ . \tag{3.18}$$

The definition of joint and conditional entropy is completed by the condition that the entropy of a pair of random variables can be expressed as the entropy of one in addition to the conditional entropy of the other one:

$$H(X,Y) = H(X) + H(Y|X) \ . \tag{3.19}$$

3.3.2 Kullback-Leibler Divergence

Kullback-Leibler (KL) divergence (or relative entropy or Kullback-Leibler distance) is a generalization of Shannon's measure of information. The divergence is a function of two probability distributions $p(x)$ and $q(x)$ that potentially characterize the same random variable X:

$$D(p;q) = \sum_{x} p(x) \ \log \frac{p(x)}{q(x)} \ . \tag{3.20}$$

In this definition, the convention is used that $0 \log \frac{0}{q(x)} = 0$ and $p(x) \log \frac{p(x)}{0} \rightarrow \infty$. The KL-divergence is nothing else but Shannon's measure of uncertainty for a random variable X if $q(x)$ is a uniform probability distribution. Thus, Shannon's entropy can be interpreted as the amount of information in a model $q(x)$ of X compared to the maximum incertitude model — the uniform distribution. The uniform distribution is the one with maximum entropy.

3.3.3 Mutual Information

Mutual information is the third measure of information next to entropy and Kullback-Leibler divergence. In the same way as we wrote the conditional entropy, we can express the mutual information

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= \sum_{x,y} p(x,y) \ \log \frac{p(x,y)}{p(x)\,p(y)} \end{aligned} \tag{3.21}$$

as the reduction in uncertainty of X due to the knowledge of another random variable Y. For instance, if X is observed under rather noisy conditions, then X and Y will be statistically independent and $I(X,Y) = 0$.

Data processing inequality

So far we have considered the correlation between two random variables X and Y modeled by the causality

$$X \rightarrow Y \ . \tag{3.22}$$

The causality between X and Y is also said to form a Markov chain in the order $X \rightarrow Y$ if the conditional distribution of Y only depends on X. If we consider the random variables U, X, Y and V with the more general causality

$$U \rightarrow X \rightarrow Y \rightarrow V \ , \tag{3.23}$$

described by the joint probability distribution

$$p(U,X,Y,V) = p(U)\, p(X|U)\, p(Y|X)\, p(V|Y) \tag{3.24}$$

we obtain the following inequality (data processing inequality):

$$I(U;V) \leq I(X;Y) \ . \tag{3.25}$$

This equation states that data processing never increases information, or in other words, no clever manipulation of the data can improve the inferences that can be made from the data.

To make Eq. 3.24 more evident, we can interprete the model as the causality between different levels of semantic abstraction in the hierarchical image content representation implemented in our system: image data (U), extracted primitive image attributes (X), content-index (Y) and finally the semantic interpretation of the image content by the user (V).

3.3.4 Fisher Information and Cramér-Rao Inequality

In Sec. 3.2, we dealt with the estimation of a parameter vector $\boldsymbol{\theta}$ from a limited set of observations X. For convenience, we consider a scalar parameter θ instead of a parameter vector $\boldsymbol{\theta}$ in the following. Since the elements of X are assigned with some uncertainties — they are random samples — also the estimate $\hat{\theta}$ as a function of X is affected by an error we aim to determine.

Quality of an estimator

For many applications, it is often necessary to measure the accuracy of the estimation result to evaluate the belief in $\hat{\theta}$. If a poor estimation is the outcome, one can increase the number of observations to attempt to obtain better results, for instance. To describe the accuracy of estimated parameters, estimation theory provides several criteria that define good estimates (VAN TREES 1968). The quality of an estimator $\hat{\theta}$ is judged if

1. it yields on the average the true values of the unknown parameters. Mathematically, if

$$E\{\hat{\theta}\} = \theta \tag{3.26}$$

   the estimator is called unbiased

2. and the conditional variance $\sigma^2_{\hat{\theta}}$ satisfies

$$\sigma^2_{\hat{\theta}} = E\left\{\hat{\theta} - E\{\hat{\theta}\}\right\}^2 \quad . \tag{3.27}$$

Intuitively, one hopes to select an estimator $\hat{\theta}$ so that it is unbiased, $E\{\hat{\theta}\} = \theta$, and the conditional variance $\sigma^2_{\hat{\theta}}$ is as small as possible. If both criteria are fulfilled for an estimator, we call it the minimum-variance unbiased estimator and can write

$$\theta = \hat{\theta} \pm \sigma_{\hat{\theta}} \quad . \tag{3.28}$$

The minimum-variance unbiased estimator is sometimes substituted if it does not exist or others are preferred. Variance is not the only quality measurement, however, variance contains most information about the estimator and is therefore of utmost importance for the accuracy assessment.

Cramér-Rao bound

Placing a lower bound on the variance of an unbiased estimator has demonstrated to be useful in practice. In the best case, the estimator is the minimum-variance unbiased estimator. Thereby, the estimator reaches the bound for all values of the estimated parameters. At worst, the bound allows to make a decision for or against an estimator from a set of estimators. The core in measuring the accuracy of an estimator is thus to find a method to determine the bound. Although there exist several of these bounds (MCAULAY and HOFSTETTER 1971) (ZIV and ZAKAI 1969), the Cramér-Rao bound (CRB) is the most convenient one. For any unbiased estimator, the CR bound is the lower bound on the conditional variance and can be expressed as (KAY 1993)

$$\sigma^2_{\hat{\theta}} \geq \frac{1}{E\left\{\left(\frac{\partial \log p(X|\theta)}{\partial \theta}\right)\right\}^2} \quad , \tag{3.29}$$

where $p(X|\theta)$ is the probability (or pdf in the continuous case) of a data set X given a scalar parameter θ. The CR bound can also be expressed in equivalent form in terms of the second order derivative and yields

$$\sigma_{\hat{\theta}}^2 \geq \frac{1}{-E\left\{\dfrac{\partial^2 \log p(X|\theta)}{\partial\theta^2}\right\}} \quad . \tag{3.30}$$

If an estimator is unbiased and attains the Cramér-Rao bound, it is said to be efficient in that it efficiently makes use of the data. The denominator of Eq. 3.30 is referred to as the Fisher information $I(\theta)$ for the data X (Fisher 1925).

In practice one is sometimes confronted with the estimation of a parameter that is a function of a more fundamental parameter. If we denote a transformed parameter with $\theta' = f(\theta)$, then the Cramér-Rao bound is given as (Kay 1993)

$$\sigma_{\hat{\theta}'}^2 \geq \frac{(\partial f/\partial\theta)^2}{-E\left\{\dfrac{\partial^2 \log p(X|\theta)}{\partial\theta^2}\right\}} \quad . \tag{3.31}$$

However, if f is a non-linear transformation, the efficiency of an estimator is destroyed; it is only maintained for linear transformations.

The lower bound on an estimate as presented in Eq. 3.29 and 3.30 makes only use of the likelihood $p(X|\theta)$. Thus, the Cramér-Rao bound is related to the determination of an unknown and deterministic parameter as outlined in Sec. 3.2 for maximum-likelihood parameter estimation. If the parameter is stochastic and the prior information about the parameter is contained in the probability $p(\theta)$, then the bound on the maximum a-posteriori estimate can be found to be (van Trees 1968)

$$\sigma_{\hat{\theta}}^2 \geq \frac{1}{-E\left\{\dfrac{\partial^2 \log p(X|\theta)}{\partial\theta^2}\right\} - E\left\{\dfrac{\partial^2 \log p(\theta)}{\partial\theta^2}\right\}} \quad . \tag{3.32}$$

Fisher information matrix

Sometimes one does not want to determine the value of a single parameter, but rather to estimate a vector parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$. If we assume an unbiased estimator $\hat{\boldsymbol{\theta}}$, $E\{\hat{\boldsymbol{\theta}}\} = \boldsymbol{\theta}$, we can use the Cramér-Rao bound to place a bound on the variance of each element of $\boldsymbol{\theta}$. Then, the covariance matrix of any unbiased estimator satisfies

$$\sigma_{\hat{\boldsymbol{\theta}}}^2 \geq \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \tag{3.33}$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ denotes the Fisher information matrix. This matrix is defined by

$$[\boldsymbol{I}(\boldsymbol{\theta})]_{ij} = -E\left\{\frac{\partial^2 \log p(X|\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j}\right\} \quad . \tag{3.34}$$

with θ_i and θ_j as two different elements of the parameter vector $\boldsymbol{\theta}$. The Fisher information matrix for the MAP estimate of a parameter vector is defined analogously to Eq. 3.32.

Analogy of Fisher information and entropy

We outlined that Fisher's measure of information is a measure of the quality of estimating a parameter of a distribution. Entropy defined the amount of information contained in the distribution of a single random variable. Now, we briefly focus on the relationships between these two fundamental measures of information.

If X is a continuous random variable with finite variance and corresponding probability density function $p(x)$ and Z is an independent Gaussian random variable with $Z \sim \mathcal{N}(0,1)$, the differential entropy $H(Y)$ [1] of the mixture $Y = X + \sqrt{t}\,Z$ is given as (Cover and Thomas 1991)

$$\frac{1}{2} I(X + \sqrt{t}\,Z) = \frac{\partial}{\partial t} H(X + \sqrt{t}\,Z) \tag{3.35}$$

and in the limit $t \to 0$

$$\frac{1}{2} I(X) = \frac{\partial}{\partial t} H(X + \sqrt{t}\,Z) \mid_{t=0} \quad . \tag{3.36}$$

This means that Fisher's information $I(\cdot)$ can be expressed in terms of entropy $H(\cdot)$ and vice versa (de Bruijn's identity). Including Eq. 3.35, we arrive at a level where we can express all demonstrated information-theoretic measures by entropy: Kullback-Leibler divergence can be seen as Shannon's entropy of a certain model compared to the uniform distribution, mutual information as the difference between entropy and conditional entropy and finally Fisher's measure of information as differential entropy. Consequently, entropy can be seen as the basic measure of information which other measures can be ascribed to.

3.3.5 Combination of Information-theoretic Measures

How to assess the quality and performance of a complex system like the image information mining system? In previous sections, we outlined basic measurement categories to determine both the accuracy and the information content of a single quantity and the association between two or more quantities. With the proposed

[1] In contrast to the definition of entropy for a discrete random variable, differential entropy is defined as

$$H(Y) = -\int_{-\infty}^{\infty} p(y)\, \log p(y)\, dy \quad .$$

where $p(y)$ denotes the probability density function of Y.

measurements we have a set of tools at hand to specify the performance of individual system parts and their interactions. However, no single measurement is suitable to represent the overall system performance in one condensed representation. Each quantity is just assigned to one particular module of the system or to the interconnection between two or more of them. On the other hand, information-theoretic measures indicate the complexity of individual system parts. Thus, a particular methodology is required to make use of the described information-theoretic measurements and to combine them to arrive at a quantity that reflects the overall system performance.

The approach we apply to derive a quantity that reflects the overall system performance is as follows. First, we apply entropy, Kullback-Leibler divergence, mutual information and Fisher's information to obtain a measure for the quality of individual system modules, e.g. the complexity of human-computer operations using entropy or the quality of the link between user-defined semantic labels and content-index using Kullback-Leibler divergence. Then, we analyze the set of acquired measures using a Karhunen-Loéve transform. This transform allows us to verify both the relevance of our measurements and the correlation between them. After this analysis, we arrive at an ordered sequence of principal components where just the first principal component enables us to reflect the overall system performance. The verification of this approach to fuse various sources of information measurements is performed by a comparison with the user's degree of satisfaction as described in Chap. 8.

3.3.6 Other Measures of Information

The best established measures of information are the demonstrated information-theoretic quantities. However, there are several others that are applied to quantize information we briefly describe in this section.

In addition to Shannon's measure of information there exist several other definitions of entropy: Renyi entropy, entropy of a certain degree, quadratic entropy, R-norm entropy and effective entropy (Jumarie 1990). By definition, these entropies look rather similar to Shannon's entropy and make use of a probability distribution, too. However, they have other properties. Since Shannon's measure of uncertainty is the most simple one of all possible entropies, we confine ourselves to it in this thesis.

A standard technique to measure information in remote sensing is to compute the classification error in form of a confusion or error matrix. This matrix gives the quality of the applied classifier to correctly assign image pixels to certain classes by a comparison of the computed classification map and ground-truth data. After normalizing the error matrix, one arrives at a probabilistic representation of the classification accuracy. Again, a final measure of information is reflected by entropy.

Questionnaires (Lehtonen and Pahkinen 2003) are a very popular and inex-

pensive way to gather information about data or the performance of a system. Often they are even the only feasible way to reach a large number of reviewers that is large enough to infer statistically significant results. Although questionnaires are rather adaptable in what to measure ranging from the overall performance of a system to individual components, they are not suitable to measure all kinds of data. What has to be included in the analysis is the environment in which the questionnaire takes place, such as qualitative or quantitative questions. For the system evaluation we use the second category of questionnaire. System operators learn semantic covertypes of their interests, search the archive for these labels and evaluate the overall performance by marking the degree of satisfaction. In a final step of evaluation, we compare the users' degree of satisfaction with the derived objective measurements during system operation.

3.4 Conclusions

In this chapter, we have discussed the following items:

- The different terms of 'probability' and the Bayesian way of data interpretation were shown. We demonstrated how information in form of model parameters can be extracted from data by incorporating prior knowledge. However, the right choice of such a prior is crucial and often a matter of criticism. In order to verify which model has to be preferred, we shortly pointed out Bayesian model selection.

- Based on random variables and parametric signal models, a number of measures to quantize information exist. With Shannon's entropy, Kullback-Leibler divergence and Fisher's measure of information we described well-established information-theoretic quantities that are appropriate for system validation. Since the main focus of this thesis is not the verification of individual system components but rather the entire system evaluation, we combine several measures of information in order to obtain the overall system performance.

Part II

System Concept

4

Hierarchical Bayesian Image Information Representation

Before we focus on the image information mining system and its evaluation in the following chapters, we will primarily demonstrate the basic theoretical principles of hierarchical Bayesian image content representation on which the system is based. This scheme was first applied in the domain of content-based retrieval (Schröder et al. 2000) to arrange the content of remote sensing images at different levels of semantic abstraction and extended by (Daschiel and Datcu 2003a) for defining higher-level semantic concepts. From the computational point of view, the hierarchy can be partitioned into two major parts: a computational intensive off-line part which aims at the extraction and description of the image content in a completely unsupervised, application-free way (levels 0 to 3), and the fast, user-specific definition and aggregation of semantic cover-types (levels 4 and 5). The process of information extraction in the hierarchical scheme is organized similarly to MIT's FourEyes (Minka and Picard 1997) that consists of unsupervised grouping and supervised learning. However, this time it is organized in a way that information at a certain level is determined from one or more levels below in a step of Bayesian inference.

In the following section, Sec. 4.1, we show the hierarchical modeling of image content starting with image data at the lowest level and finish at higher-level semantic concepts. In this context we outline the flow of information between system levels using various steps of stochastic inference. In Sec. 4.2, we point out how the elements at each level are affected and result in a decrease of information flow. We demonstrate methods to evaluate these influences and refer to later parts of the thesis dealing with it. Comprising, this chapter gives an introduction to basic theoretical concepts underlying the mining system. It also shows both the complexity of the mining system and the difficulty to perform an overall system effectiveness evaluation.

4.1 Hierarchical Bayesian Image Content Modeling

The concept of information representation on different hierarchical levels of semantic abstraction is based on a 6 level Bayesian learning model as illustrated in (Fig. 4.1). We will now proceed to describe how the image information at a certain level is determined from level(s) below in a step of Bayesian inference.

From level 0 to level 1: Primitive feature extraction using parametric signal models

In a first step, we extract spatial, spectral and geometrical features $\boldsymbol{\theta}$ (level 1) from the image data D (level 0) using different stochastic signal models M. These models are expressed in their parametric form $p(D|\boldsymbol{\theta}, M)$ and assign the probability to a realization of the data D for a particular parameter vector $\boldsymbol{\theta}$. At the core of the information extraction process is the estimation of the parameter vector $\boldsymbol{\theta}$ given the data D. In our concept of Bayesian modeling, this process is realized as the maximum a-posteriori estimate of the parameter vector as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D, M) \ . \tag{4.1}$$

The results of this information extraction procedure indicate the elements on level 1 in our hierarchical scheme. Of course, which structures and objects can be captured and the accuracy of the estimation process depend both on applied model and image data. In Sec. 5.1, we show several models to extract primitive features such as spectral, textural and geometrical attributes from optical as well as radar image data.

From level 1 to level 2: Meta features

In the same way as the Bayesian technique was applied to estimate image parameters from the data, it can be used to find the most evident model. This approach is opposed to finding the most complex model that always describes the data best. The evidence of the model, e.g. the probability of the model given the data, can be obtained as

$$p(M|D) = \frac{p(D|M)\, p(M)}{p(D)} \ , \tag{4.2}$$

where the probability of the data D can be obtained via the integration

$$p(D|M) = \int p(D|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M) d\boldsymbol{\theta} \tag{4.3}$$

over the complete parameter space of $\boldsymbol{\theta}$. In Eq. 4.3, the penalty factor (Occam's razor) is implicitly contained as the width in $\boldsymbol{\theta}$ between the likelihood $p(D|\boldsymbol{\theta}, M)$ and

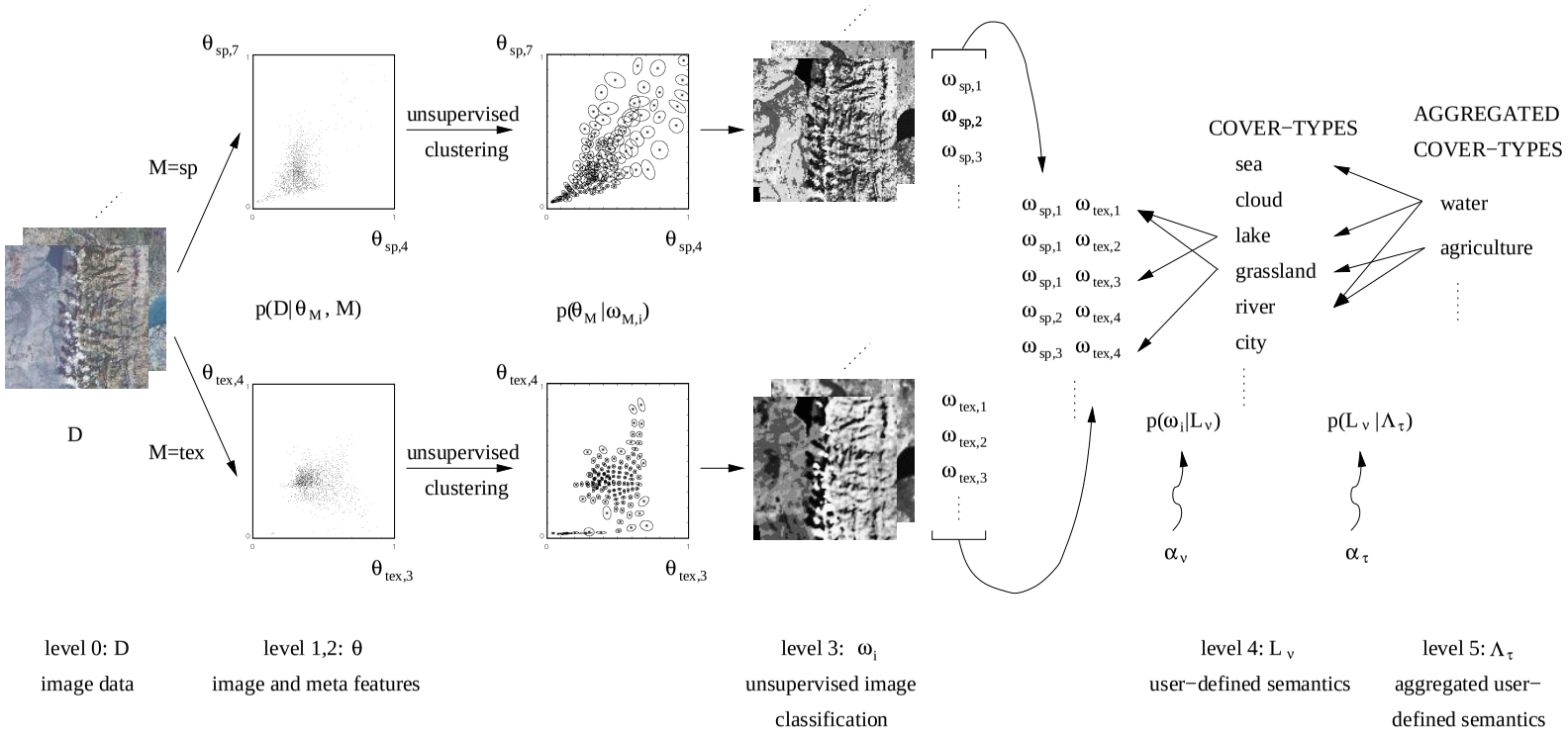


Figure 4.1: Hierarchical modeling of image content and user semantics. First, primitive image features $\boldsymbol{\theta}$ and meta-features are extracted from image data D, based on different parametric signal models M, e.g. spectral (sp) and texture (tex). Through an unsupervised clustering of the features, we obtain a vocabulary of signal classes ω_i for each model M. Using Bayesian networks, user's interests, that is, the semantic interpretation of cover-type labels L_ν, are linked to the signal classes ω_i. In a final step of stochastic modeling, semantic labels can be grouped (aggregated) into higher-level semantic concepts Λ_τ by probabilities $p(L_\nu|\Lambda_\tau)$.

the prior distribution $p(\boldsymbol{\theta}|M)$. While the estimated image parameters are related to the image content, the evidence keeps information about the model and its parameters to describe the image content. These features of features are called 'meta features' and indicate the next level (level 2) in our hierarchical representation.

From level 1/2 to level 3: Unsupervised image classification

Feature extraction from level 1 and 2 produces large volumes of data that cannot be managed in practice. Thus, we have to reduce and compress the estimated parameters to obtain manageable data quantities. Clustering, which is similar to a quantization process, reduces the accuracy of the system, but justifies its practical use due to a large data reduction. In order to reject existing structures in the different feature spaces of the data and to avoid the time-consuming calculation of similarity functions, the unsupervised clustering is performed across all images as illustrated in (Fig. 4.1).

From the clustered data we obtain a vocabulary of signal classes (level 3) that reflect characteristic structures in the different feature spaces, e.g. significant spectral structures of multispectral images. We perform the global unsupervised clustering using a dyadic k-means algorithm (Daschiel and Datcu 2002a) with a predefined number of clusters. Our clustering method is a modified and accelerated version of the well-known k-means algorithm implemented to manage and process large amounts of data. Even if it is less accurate than other clustering methods, e.g. clustering by melting (Wong 1993) or Bayesian classification (Cheeseman and Stutz 1995), it justifies its practical use due to a significantly reduced processing time.

Common to all clustering algorithms is that they take a model for the clusters ω_i into account. In our approach, we can express the cluster model in its parametric form as

$$p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_j \epsilon\, \omega_i, V, T) \ , \tag{4.4}$$

where the jth data point of estimated parameter vector $\boldsymbol{\theta}_j$ is associated by probabilities to the ith cluster ω_i in the feature space. V indicates the parameters for each cluster, mean and variance, and T the mathematical model (inclusive the number of clusters).

After grouping all points in the feature space, each individual point $\boldsymbol{\theta}_j$ from the estimated parameters is clearly associated to one of the clusters ω_i. Mathematically expressed, clustering yields the posterior probability

$$p(\omega_i|\boldsymbol{\theta}_j) \tag{4.5}$$

of signal classes ω_i given the estimated parameter vectors $\boldsymbol{\theta}_j$. In a step of Bayesian inference we can make the connection between data D and signal classes ω_i as

$$p(\omega_i|D) = \int p(\omega_i|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} \ . \tag{4.6}$$

Since solving this integral is rather expensive — for a detailed explanation see (SCHRÖDER-BRZOSNIOWSKY 2000) — we make the approximation

$$p(\omega_i|D) \approx p(\omega_i|\boldsymbol{\theta}) \ . \tag{4.7}$$

In order to decrease the computational demands and to make the approach applicable to large data quantities, this approximation is assumed in all the following steps.

From level 3 to level 4: Semantic labeling

The first three levels in our hierarchical representation describe the data D in a completely unsupervised and application-free way. Since no single signal model may capture the whole information contained in image data, several models are normally applied on level 1. Based on the objective characterization of the image data on level 3, $p(\omega_i|D)$, we can link the subjective user interests in form of semantic cover-types L_ν (level 4) to signal classes ω_i by probabilities $p(\omega_i|L_\nu)$ as

$$p(L_\nu|D) = \sum_i p(L_\nu|\omega_i)\, p(\omega_i|D) \ . \tag{4.8}$$

In the current system, the probabilistic link $p(\omega_i|L_\nu)$ is derived from the user's positive and negative fed-back samples using a vector of hyper-parameters $\boldsymbol{\alpha}_\nu$ (Fig. 4.1). Each time the user interacts with the system and provides new training samples, the vector of hyper-parameters and, as a consequence, the probabilistic link are updated. In Sec. 5.4, we show the detailed learning procedure of how probabilities are derived from the user's training examples.

From level 4 to level 5: Aggregation of semantic labels

The elements on level 4 constitute the users' interests in form of semantic interpretations of the image content based on the characteristic vocabulary of signal classes on level 3. With the demonstrated learning paradigm users can define specific semantic cover-type labels, e.g. 'lake', 'city' or 'mountains'. However, it does not allow the definition of complex semantics such as 'a house next to a river' or 'trees around a lake in a park'. Therefore, we formulate an additional step of Bayesian inference in our hierarchical image content modeling and aggregate user-specific semantic labels L_ν to higher-level semantic concepts Λ_τ (level 5).

In the same way as we made the inference from semantic cover-type labels L_ν to image data D in the last step, we can now link higher-level semantics to image data as

$$p(\Lambda_\tau|D) = \sum_\nu p(\Lambda_\tau|L_\nu)\, p(L_\nu|D) \ . \tag{4.9}$$

Before we can learn the probabilities $p(L_\nu|\Lambda_\tau)$ from the user's feedback, we have to link semantic cover-types L_ν to image data D

$$p(L_\nu|D) = \sum_i p(L_\nu|\omega_i)\, p(\omega_i|D) \quad . \tag{4.10}$$

Note that this link is calculated according to Eq. 4.6 between level 0 and level 3, however, this time between image data D and all semantic labels L_ν in the database inventory. Since the computation of $p(L_\nu|D)$ in Eq. 4.10 over all labels in the DBMS requires some efforts, which runs counter to our aim to achieve a fast on-line system operation, we derive the link between image data D and semantic labels L_ν via a supervised maximum a-posteriori (Bayesian) classification. The user's positive training samples for L_ν serve as input training data for the classification algorithm (Sec. 6.2).

The probabilities $p(L_\nu|\Lambda_\tau)$ are the core in Eq. 4.9 that are derived from the user's positive and negative learning feedback. Again, we use a vector of hyper-parameters $\boldsymbol{\alpha}_\tau$ to describe the stochastic link.

4.2 Validation Issues of the Hierarchical Scheme

The previous section demonstrated a scheme for hierarchical modeling of image content based on several levels of different semantic abstraction. In this section, we use these levels again and point out how their elements can be affected and lead to a verification of the information flow. In detail, we aim at determining the information content of the elements at each level in order to derive accuracy measurements indicating the information flow from image raw data up to higher-level semantic concepts.

The process of information extraction starts at level 0 where primitive image parameters $\boldsymbol{\theta}$ are estimated from data D using different stochastic signal models M (Fig. 4.1). Since the quality of the acquired image data influences the estimation process, we start with analyzing the data quality in a first step. The method we apply to measure the radiometric image 'quality' is to determine the degree of noise in the data signal. With this analysis, we can avoid obtaining unpredictable results in the further image content representation: clustering and semantic labeling.

Having estimated the model parameter vector $\boldsymbol{\theta}$ that indicates the elements on level 1, we determine the accuracy of the elements of $\boldsymbol{\theta}$ using information-theoretic measures. Of central importance in this context is the Fisher information matrix, which reflects the bounded variance of the elements of the parameter vector.

On level 3, the 'clouds' of estimated feature parameter vectors $\boldsymbol{\theta}_j$ in a multi-dimensional feature space are substituted by clusters ω_i as their condensed representations. The performance of the clustering process and the information content of clusters, respectively, depends on the ability to capture characteristic groups in the feature spaces of the different signal models. Since the question as to identifying

level	variable	model	evaluation measure	reference
5	Λ_τ	$p(L_\nu\|\Lambda_\tau)$	retrieval accuracy	Sec. 6.2, 6.3
4	L_ν	$p(\omega_i\|L_\nu)$	supervised classification and retrieval accuracy, man-machine interaction	Sec. 7.3, 7.4 7.5, 7.6, 7.7
3	ω_i	$p(\boldsymbol{\theta}\|\omega_i)$	scatter matrices, Bayes' probability of error, density estimation, divergence, unsupervised classification accuracy	Sec. 7.2
1	$\boldsymbol{\theta}$	$p(D\|\boldsymbol{\theta}, M)$	Fisher information matrix	Sec. 7.1
0	D		noise variance estimation	Sec. 7.1

Table 4.1: Validation measures for the hierarchical scheme of image content modeling and representation. The references point to sections which deal with the analysis of the elements at specific levels. The elements at levels 0 to 3 are verified in a complete objective way whereas cover-types (level 4) and aggregated labels (level 5) are influence by human-factors, too.

'characteristic' groups cannot be answered without time-consuming feature space analysis, a common approach is to first cluster the feature space and then measure separability and isolation of the obtained clusters to evaluate the results. Methods that are often used in this context are scatter matrices, Bayes' probability of error, non-parametric feature space density estimation and divergence. They all try to identify the optimal partition of points in the feature space by a restricted number of clusters. We want to point out that a high-performance clustering is the assumption to access information details at image level and to successfully link the user's interests (specific and aggregated semantics) to image data.

The evaluation of the elements at the next hierarchical level, level 4, is mostly related to classification selectivity, cover-type separation and strength of the stochastic link between signal classes ω_i and cover-type labels L_ν. Classification, separation and stochastic link involve the user and its interaction with the system and thus do not make a purely objective evaluation possible.

Verifying aggregated semantic labels is even more complicated than validating individual labels on level 4 since the cover-types are already subjective. The difficulty is that elements on level 4 may be insufficiently defined or the semantic does not reflect the actual meaning, e.g. a cover-type 'forest' is ingested as 'grassland'. Thus, we confine the evaluation of elements at level 5 to the identification of target and misclassified images in the probabilistic search results. In (Tab. 4.1), we summarize validation measurements for individual system levels and refer to later parts dealing with them.

Before we describe the evaluation concept and measurements in more detail in Chap. 7, we will first outline the image information mining system and its enhancement in Chap. 5 and 6, respectively.

4.3 Conclusions

In this chapter, we have discussed the following items:

- We demonstrated how to arrange the information content of image data at multiple levels of different semantic abstraction. Therefore, we applied a hierarchical scheme where the elements at a certain level are obtained from levels below in a step of Bayesian inference. The main application of this scheme is to first describe the image content in an unsupervised and application-free way and then to link user-specific interest to this content-index.

- The basic hierarchial concept is extended to a new level of image content abstraction: semantic aggregation. At this level, existing user-specific labels are grouped to higher-level semantic concepts that can be used for database retrieval.

- Based on the hierarchical modeling of image content from image data up to higher-level semantics, we proposed an evaluation methodology that is adapted to this scheme. Since the information is contained in the applied stochastic signal models and the parameters to be determined, evaluating the image information mining system means quantizing the information content of observations and model parameters.

5

Information Mining in Remote Sensing Image Archives

The last decade has been marked by important research efforts to develop content-based image retrieval concepts and systems. Images in an archive are searched by their visual similarities with respect to color, texture and shape. However, CBIR concepts have been computer-centered approaches — the concepts hardly allowing any adaptivity to users' needs. Thus, image retrieval systems have been equipped with relevance feedback functions (Cox et al. 1996). The systems are designed to search images similar to the users' conjecture. Another interesting approach is based on a learning algorithm to select and combine feature grouping and to allow users to give positive and negative examples (Minka and Picard 1997). Both concepts are first approaches to include the user in the search loop, they are information mining concepts (Zhang et al. 2001). They are also methods in the trend of designing human-centric systems.

In addition to the operational state-of-the-art archive and database systems, we have implemented a concept for image information mining that supports the man-machine interaction via the Internet and adaptively incorporates application-specific interests. The system concept for exploring the information content of remote sensing images and understanding the observed scenes was first implemented and successfully tested in the Multi-Mission Demonstrator (MMDEMO) (Schröder et al. 2000). A further upgrade was made by (Datcu et al. 2003) that resulted in a prototype of a Knowledge-driven Information Mining (KIM) system. Before we focus on the system enhancement and evaluation of the image understanding and mining functions in the following chapters, we will explain the concepts underlying image information mining. We follow the order of hierarchical image information modeling as presented in the previous chapter.

We begin this chapter with the extraction of primitive image parameters in Sec. 5.1. We mainly focus on Gibbs random field texture models to capture structures, objects and scattering properties in optical as well as in radar scenes. In the next section, Sec. 5.2, we show how a vocabulary of characteristic signal classes that is valid across all images in the archive is computed from the estimated features.

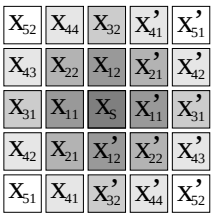


Figure 5.1: Graphical definition of a Gibbs-Markov random field up to order 5. Neighbouring pixels x_{ij} and $x_{ij}{}'$ are in interaction with a central pixel x_s.

Sec. 5.3 and 5.4, respectively, deal with semantic labeling of user-specific cover-type labels by interactive learning based on the generated unsupervised content-index. Then, in Sec. 5.5, we point out how the entire archive can be searched in a probabilistic way for images that contain structures similar to a defined cover-type. After Sec. 5.6, a section dedicated to system description, configuration and information transmission aspects, we conclude this chapter with an illustration of the system with several practical applications.

5.1 Primitive Image Parameter Extraction

Automatic interpretation of remote sensing images and the growing interest for image information mining and query by image content from large remote sensing image archives rely on the ability and robustness of information extraction from the observed data. We focus on the modern Bayesian way of thinking and introduce a pragmatic approach to extract structural information from remote sensing images by selecting those prior models which best explain the structures within an image. On the lowest level, the image data D, we apply stochastic models to capture spatial, spectral and geometric structures in the image. These models are given as parametric data models $p(D|\boldsymbol{\theta}, M)$ and assign the probability to a given realization of the data D for a particular value of the parameter vector $\boldsymbol{\theta}$.

5.1.1 Optical Images

To apply parametric stochastic models (Datcu et al. 1998) in order to extract primitive image features, the data is understood as a realization of a stochastic process. The Gibbs-Markov random field (GRF) family of stochastic models assumes that the statistics of the grey level of a pixel in the image depends only on the grey levels of the pixels belonging to a neighbourhood with a restricted dimension (Fig. 5.1). The probability of the grey level of the pixel x_s is given by

$$p(x_s|\partial x_s, \boldsymbol{\theta}) = \frac{1}{Z_s} \exp\left(-H(x_s|\partial x_s, \boldsymbol{\theta})\right) \quad , \tag{5.1}$$

where Z_s acts as a normalization factor given by the sum over all the possible states for the pixel x_s. Assumptions have to be made for the functional form of the energy function H. In this approach we use an auto-binomial model with its energy function

$$H(x_s|\partial x_s, \boldsymbol{\theta}) = -\log \begin{pmatrix} G \\ x_s \end{pmatrix} - x_s \cdot \eta \tag{5.2}$$

and

$$\eta = a + \sum_{ij} \theta_{ij} \frac{x_{ij} + x'_{ij}}{G} \tag{5.3}$$

as the joint influence of all neighbours weighted by the elements of the parameter vector $\boldsymbol{\theta}$. Each element θ_{ij} of the parameter vector describes the interaction between the pixel x_s and the pair x_{ij}, $x_{ij}{}'$, while the parameter a represents a sort of auto-interaction. G indicates the maximum grey value, e.g. 255 for a 8 bit image.

A fitting of the model on the image is performed in order to obtain the best fitting parameters. For the estimation a conditional least-squares estimator (Lele and Ord 1986) is obtained by

$$\hat{\boldsymbol{\theta}}_{CLS} = \arg\min_{\boldsymbol{\theta}} \sum_s \left(x_s - \sum_{x_s=0}^{G} x_s \, p(x_s|\partial x_s, \boldsymbol{\theta}) \right)^2 \quad . \tag{5.4}$$

The evidence of the model, e.g. the probability of the model given the data, can be calculated by

$$p(M|D) = \frac{p(D|M)\, p(M)}{p(D)} \quad , \tag{5.5}$$

where the probability of the data D can be obtained via the integration

$$p(D|M) = \int p(D|\boldsymbol{\theta}, M)\, p(\boldsymbol{\theta}|M)\, d\boldsymbol{\theta} \quad . \tag{5.6}$$

From the estimated parameters, we derive several features to describe the image content: the norm of the estimated parameters $|\hat{\boldsymbol{\theta}}|$ as the strength of the texture, the estimate of the variance $\hat{\sigma}^2_M$ as the difference between signal and model energy (Ruanaidh and Fitzgerald 1996), the evidence of the model M, Eq. 5.5, and the local mean of the estimation kernel (Fig. 5.2).

To describe the image content by using the spectral properties, we do not have to explicitly estimate the parameter vector $\boldsymbol{\theta}$. Instead, we can directly assign the individual spectral channels (after a normalization) to elements of the vector $\boldsymbol{\theta}$, e.g. the 6 spectral channels in the visible spectrum of Landsat TM result in a six-dimensional vector $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_6\}$.

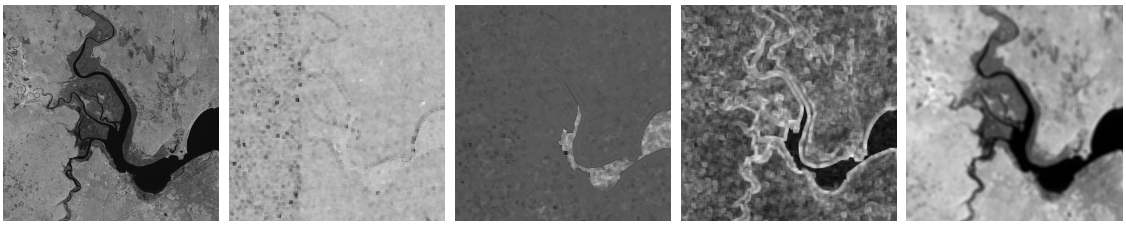


Figure 5.2: Extracted textural features from a Landsat TM image. Images from left to right. 4th band of Landsat TM as data D, norm $|\hat{\boldsymbol{\theta}}|$ of the estimated texture parameters, variance $\hat{\sigma}^2_M$, evidence $p(M|D)$ and local mean of the estimation kernel.

5.1.2 SAR Images

In the system, there exist co-registered optical and SAR images. To include radar information in the retrieval process for an entire exploitation of the image archive and to enable the mining of multi-sensor data for sensor qualification, we have to extract content-based image parameters from SAR data, too.

The information extraction is achieved as a model-based Bayesian approach (Datcu et al. 1998) (Walessa and Datcu 2000). The system models and reconstructs an estimated backscatter image that is free of speckle noise, while still completely preserving its most important structural information. Furthermore, the system evaluates the parameters $\boldsymbol{\theta}$ that describe the scene structures, e.g. textures, edges and strong targets. The information extraction is a space variant process describing precisely the scene non-stationarity.

Since the system takes both the statistics of the noisy and the noise-free data in a Bayesian framework into account, the choice of an appropriate model for the estimated backscatter image plays an important role, and affects the obtained results directly. In order to filter out speckle, the Bayesian formula

$$p(x|y,\boldsymbol{\theta}) = \frac{p(y|x)\,p(x|\boldsymbol{\theta})}{p(y|\boldsymbol{\theta})} \tag{5.7}$$

is used, where we try to estimate the noise-free image which best explains the noisy observation assuming some prior information. By x we describe a noise-free pixel of the image, y indicates a pixel of the noisy observation, e.g. the ERS1 image, and by $\boldsymbol{\theta}$ we characterize the parameters of the applied model.

The Bayes' equation, Eq. 5.7, allows the formulation of the information extraction problem as a maximum a posteriori (MAP) estimation:

$$\hat{x}_{MAP} = \arg\max_{x}\; p(x|y,\boldsymbol{\theta}) \tag{5.8}$$

and

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}}\; p(\boldsymbol{\theta}|y)\ . \tag{5.9}$$

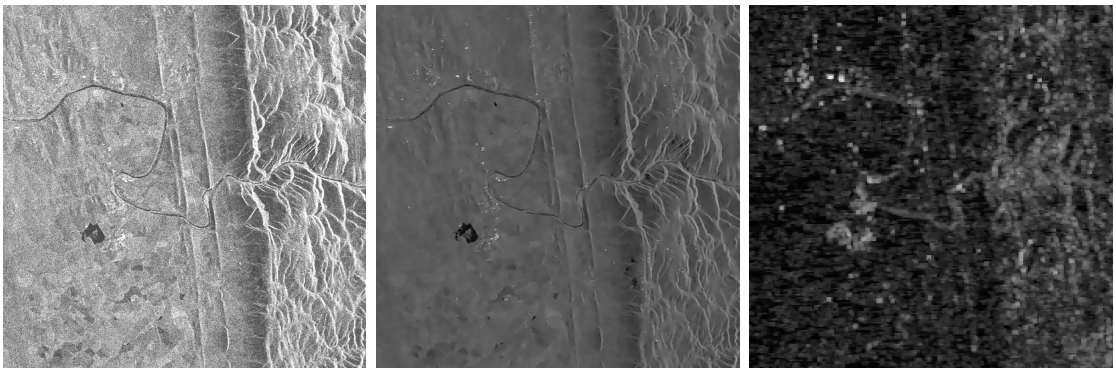


Figure 5.3: Information extraction from radar images. Images from left to right. Original ERS1 intensity image covering Mozambique with a high diversity of structures, such as mountains, rivers and flat terrain, the model-based despeckled image and norm $|\hat{\boldsymbol{\theta}}|$ of the estimated texture parameter vector.

Analytically computed maximum a posteriori estimates of the cross-section are generated from the filter. Subsequently, they are employed to produce parameters for $p(x|y,\boldsymbol{\theta})$ by iterative maximization of the evidence (Walessa and Datcu 2000). Expectation maximization is used to estimate the non-stationary texture parameters that provide the highest evidence value. The estimated model parameters express the characteristics of the texture and the strength of geometrical structures in the data.

The model used as a prior is the Gauss-Markov random field (GMRF) texture model (Datcu et al. 1998) (Walessa and Datcu 2000)

$$p(x_s|\partial x_s,\sigma^2,\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left(-\frac{(x_s-\sum_{ij}\theta_{ij}(x_{ij}+x_{ij}'))^2}{2\sigma^2}\right) \tag{5.10}$$

specified by σ^2 and the parameter vector $\boldsymbol{\theta}=(\theta_{11},\theta_{12},\theta_{21},\ldots)$. The latter is defined on a neighborhood of cliques centered on the generic pixel x_s so that the scalar parameters are symmetric around the central element. The main strength of the Gauss-Markov model lies in its ability to model structures in a wide set of images while still allowing analytical tractability. The likelihood used in the Bayes equation, Eq. 5.7, is the gamma distribution

$$p(y|x) = 2\left(\frac{y}{x}\right)^{2L-1}\ \frac{L^L}{x\Gamma(L)}\exp\left(-L\left(\frac{y}{x}\right)^2\right) \tag{5.11}$$

with L the number of looks of the data and $\Gamma(\cdot)$ the gamma function. From the estimated parameters, we take the model-based filtered intensity image and the norm of the model parameter $|\hat{\boldsymbol{\theta}}|$ as exemplified in (Fig. 5.3).

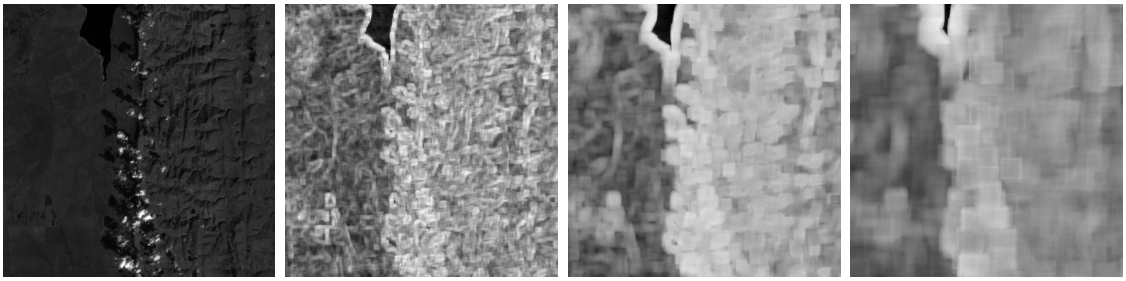


Figure 5.4: Spatial information extraction from Landsat TM at three different scales. Images from left to right. Band 4 of a Landsat TM image from which texture features in form of evidence, Eq. 5.5, are extracted at 30m, 60m and 120m, respectively. Captured large-scale textures show a more compact representation at a scale of 120m while small local variations are quite well extracted from data at original resolution.

5.1.3 Information Extraction at Multiple Scales

We showed that parametric data models are suitable to characterize spatial information in images by its parameter vector $\boldsymbol{\theta}$. Capturing high complex textures that have features at different scales, particularly large-scale structures such as mountains or rivers, requires high order models. With an increasing neighbourhood size, the number of parameters grows and leads to an averaging effect of different parameters. This results in a limited discrimination power of the extracted texture features and impairs the interpretation.

The approach we follow for a quasi-complete description of all texture structures is to generate a multiresolution image pyramid where the original image is located at the lowest layer and the reduced resolution representations of the image at higher layers (Schröder et al. 1998). If the same Gibbs random field texture model is applied with a limited neighbourhood size at different layers, information for different structures is extracted. Thus, we can characterize large-extended spatial information by a restricted model order (Mao and Jain 1992). Another, very remarkable reason for extracting texture information at different scales is the speed increase of the extraction process. The amount of data is decreased by a factor of 1/4 from one scale to the next scale in our dyadic pyramid. In (Datcu et al. 2003), it was demonstrated that the loss of information due to scaling at final semantic labeling is minimal.

An example for the multi-scale approach on Landsat TM image data is shown. These images are characterized by a high diversity of structures at very different scales, e.g. cities, cultivated land and geological structures. The appearing difficulty is to robust extract long-range textures with a periodicity much larger than the Gibbs kernel. A solution is to produce scaled versions of the image to bring the

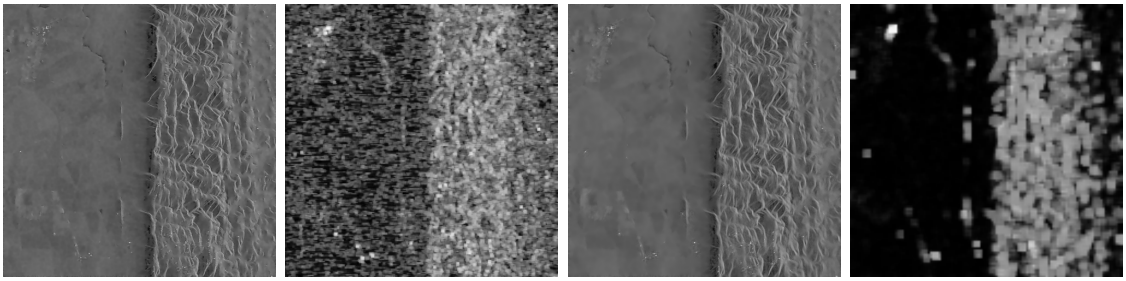


Figure 5.5: Spatial information extraction from an ERS1 image at two different scales. Images from left to right. Enhanced model-based despeckled (EMBD) image at resolution of 60m and the norm $|\hat{\boldsymbol{\theta}}|$ of the estimated model parameter vector. EMBD filtered image at 120m resolution and the norm $|\hat{\boldsymbol{\theta}}|$. With a decreasing resolution, the separability of strong targets increases due to the increased signal-to-noise ratio.

correlation length of large-extended textures to the Gibbs kernel. We perform the scaling at different levels of the Landsat image, 30m, 60m and 120m, by low-pass filtering the image data and downsampling (Fig. 5.4). In connection with SAR images (Fig 5.5) the multi-resolution is a multilook as well. Thus, at lower resolution, the accuracy of the estimated information is more precise due to the better signal-to-noise ratio (Goodman 1975).

5.2 Unsupervised Clustering and Catalogue Entry Generation

In the previous section, we pointed out how the content of optical and radar images can be described by parametric data models. Since the feature extraction produces large volumes of data that cannot be managed in practice, estimated image parameters must be compressed and reduced. Clustering, which is similar to a quantization process, reduces the accuracy of the system, but justifies its practical use due to a large data reduction. In order to reflect existing structures in the different feature spaces of the data and to avoid the time-consuming calculation of similarity functions (Jacobs et al. 1998), unsupervised clustering is performed across all images in the archive (see Fig. 4.1).

In this section, we present the dyadic k-means algorithm, a modified and enhanced version of the traditional k-means clustering tool. Furthermore, we discuss the reasons for and against the dyadic application of k-means for the unsupervised clustering of large remote sensing image datasets. Before we point out advantages and constraints of unsupervised across image classification, we briefly show how the

content-index — a characteristic "vocabulary" of signal classes — is derived from the clustering results.

5.2.1 Cluster Modeling

Assuming the jth data point from the complete set of estimated parameter vectors (elements at levels 1 and 2) is denoted by $\boldsymbol{\theta}_j$, we can associate $\boldsymbol{\theta}_j$ to the ith cluster ω_i in the feature space by the probability density

$$p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_j \epsilon\, \omega_i, V, T) \ , \tag{5.12}$$

where T indicates a particular classification model and V the parameters inside this model. In our dyadic k-means clustering approach, the classification model T is given by the dyadic method itself and V involves the number of pre-defined clusters. Each cluster is modeled as a Gaussian representation with mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\Sigma}_i$ that lead to the following parametric class model

$$p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_j \epsilon\, \omega_i, V, T) = \frac{1}{(2\pi)^{\frac{d}{2}}\ |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}_j - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}_j - \boldsymbol{\mu}_i)\right) \tag{5.13}$$

with d as the dimension of the feature space.

Having a model for the feature points belonging to one of the clusters, we are interested in the maximum a-posteriori model for clustering. With the help of Bayesian inference we can express the probability of a cluster ω_i given a data point $\boldsymbol{\theta}_j$ by

$$p(\omega_i|\boldsymbol{\theta}_j) = \frac{p(\boldsymbol{\theta}_j|\omega_i)\, p(\omega_i)}{\sum_{i=1}^{r} p(\boldsymbol{\theta}_j|\omega_i)\, p(\omega_i)} \ , \tag{5.14}$$

where $p(\omega_i)$ indicates the prior probability of the ith cluster and r the total number of clusters. Note, that the normalization in the denominator is performed over the whole cluster space.

5.2.2 Dyadic k-means Clustering Algorithm

The global unsupervised clustering using a dyadic k-means algorithm (Daschiel and Datcu 2002a) substitutes the "clouds" of primitive features by parametric data models $p(\omega_i|\boldsymbol{\theta}_j)$ as stated in the previous section. From Eq. 5.13 and 5.14 it is obvious that the probability $p(\omega_i|\boldsymbol{\theta}_j)$ is large when the Mahalanobis distance $(\boldsymbol{\theta}_j - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}_j - \boldsymbol{\mu}_i)$ is small.

If we compute the squared Euclidean distance $\|\boldsymbol{\theta}_j - \boldsymbol{\mu}_i\|^2$ instead of the Mahalanobis distance, we find the mean vector $\boldsymbol{\mu}_m$ closest to sample point $\boldsymbol{\theta}_j$ and can approximate $p(\omega_i|\boldsymbol{\theta}_j)$ as (Duda et al. 2001)

$$p(\omega_i|\boldsymbol{\theta}_j) \approx \begin{cases} 1 & \text{if } i = m \\ 0 & \text{otherwise.} \end{cases} \tag{5.15}$$

Input: number of $r = 2^q$ clusters (q = 1, 2, ...), dataset consisting of n points (samples)

Output: 2^q clusters which minimize the cost function E

Method:

1) initialize: select 2 clusters ω_1 and ω_2
2) re-compute cluster centers $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$
3) **for** $i = 1, \ldots, q-1$
4) split each cluster $\omega_1, \ldots, \omega_{2^i}$ into 2 new ones with centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{2^{i+1}}$
5) classify the samples of clusters $\omega_1, \ldots, \omega_{2^i}$ separately to one of the two new (split) ones
6) **until** no significant change in E or in $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{2^{i+1}}$ **and** number of clusters $< 2^q$

Figure 5.6: Pseudo-code of the dyadic k-means clustering algorithm.

Then, the cluster centers $\boldsymbol{\mu}_i$ $(i = 1, \ldots, r)$ can be found by the iterative application of

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} p(\omega_i|\boldsymbol{\theta}_j)\boldsymbol{\theta}_j}{\sum_{j=1}^{n} p(\omega_i|\boldsymbol{\theta}_j)} \ , \tag{5.16}$$

where n denotes the whole number of data samples in the feature space.

In (Fig. 5.6), we show the pseudo-code of the clustering algorithm. Dyadic k-means starts with the initialization of the complete dataset – associating all points $\boldsymbol{\theta}_j$ to one of the two clusters ω_1 and ω_2 with centers $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. To achieve an optimization of the squared-error cost function $E = \|\boldsymbol{\theta}_j - \boldsymbol{\mu}_i\|^2$, the initial state is updated like in k-means (Duda et al. 2001). If convergence is reached, each of the clusters ω_1 and ω_2 is separately divided into two new ones (step 4 in Fig 5.6). E is optimized for the new configuration in a way that points from one cluster can only fall into one of the split clusters. This process is repeated until E or the cluster membership does not change significantly anymore. Both splitting each cluster center into two new ones and optimizing this configuration is one level in dyadic k-means. This procedure is repeated until the defined number of r clusters is reached.

The reason for modifying k-means is to obtain an enhanced and accelerated tool for clustering huge datasets. The most time-consuming computation in k-means is the distance calculation in the d-dimensional feature space because for each point it must be done to all cluster centers. The time required for I/O is less important in comparison to the distance computation but is linearly increasing with the amount of data. Whereas the computational complexity for k-means is $O(rn)$ with r clusters and n data samples, it is $O(n \log r)$ for the dyadic algorithm. Especially for a large number of clusters our algorithm has proved to be very efficient.

From the processing point of view it is consequent to speed up the algo-

rithm as much as possible, particularly for a data-intensive task like data mining (ALSABTI et al. 1998). Since the power of a single processor is limited to a few GHz, it is challenging to run a clustering algorithm on a multi-processor machine. In the past, k-means was mapped to a hybrid processor (GOKHALE et al. 2001) but the results of this experiment did not meet the expectations. With the used configuration only a speed-up of 15% could be reached. With a special hardware configuration, however, an acceleration of factor 10 is said to be possible. For the dyadic k-means, the iterative application of k-means makes the direct efficient utilization of a single instruction multiple data (SIMD) architecture possible. The parallelization of the algorithm is essential and useful for the processing of large datasets, in particular.

Despite the potential and the quality of the dyadic k-means clustering tool, some problems remain. First of all, the I/O costs can affect the processing time, especially for image mining related applications with huge amounts of data. Only a few algorithms can overcome the problem of I/O, e.g. (ZHANG et al. 1996), but are always associated with a decrease of accuracy or other restrictions. Since our algorithm is only slightly more inaccurate than k-means, the disadvantage is balanced by the decrease of computation speed. Whereas our clustering algorithm works well in almost uniformly distributed data samples, problems occur if too many outliers are in the database. The algorithm places too many cluster centers to outliers because of the binary splitting of the clusters.

5.2.3 Coding Classes and Catalogue Entry Generation

From the results of unsupervised feature classification, we derive a set of signal classes that describe characteristic groups of points in the parametric spaces of different models. This "vocabulary" of signal classes is valid across all images, ensured by the global across image classification. The elements ω_i of this "vocabulary" are given by the cluster-membership of all image points to one of the clusters. For each image, this results in as many classification maps as the number of models that are used. From these maps, we calculate the probabilities $p(\omega_i|I_\zeta)$ of the ith class given a certain image I_ζ. These probabilities are separately computed for each signal model. We obtain the probabilities by calculating the histogram of the occurrence of signal classes ω_i in an image I_ζ. The elements of the normalized histograms, the probabilities $p(\omega_i|I_\zeta)$, are stored in a relational database system together with the classification maps. The latter are stored as binary large objects (BLOBs). Additionally, Quick-Looks (QL) and their thumbnails as BLOBs in JPEG-format, meta-information, such as sensor type, time of acquisition, geographical information, etc., are inserted.

5.2.4 Advantages and Constraints of Unsupervised Across Image Clustering

The applied concept of global image feature clustering has several advantages and limitations we will briefly point out.

First and foremost, the unsupervised clustering of image features achieves a large data reduction and represents the multidimensional signal samples in a more compact way. With the derived vocabulary of characteristic signal classes ω_i that is valid across all images, like a codebook in coding theory, for example, the applied method can be seen as a vector quantization of image features. In comparison to per image classification, across image classification does not require the time-intensive calculation of similarity functions (Jacobs et al. 1998).

Apart from the large data compression factor of unsupervised clustering, there are several drawbacks and limitations we should not forget to mention. In terms of clustering massive datasets, the most apparent questions are (1) how much information details can be represented by the global image feature clusters, (2) how many clusters are necessary and (3) how to proceed if the archive is expanded with new data. The question about information details becomes critical if the amount of data is high and the number of points of relevant classes is small. If the archive consists of 1000 images covering rural land and one image covers a city, for instance, it can happen that feature samples representing structures in the single image are not separated in the global feature space. However, this problem can be easily coped with an increase of the clusters. But one question still remains: how to choose the initial number of clusters? This limitation appears frequently in unsupervised clustering applications and requires a special optimization procedure as implemented in the AutoClass algorithm (Cheeseman and Stutz 1995), for example. For the efficient partitioning of massive datasets, however, algorithms like AutoClass are not appropriate due to their high computational complexity. An easy but very heuristic solution is to fix the number of clusters to a "reasonably" high quantity depending on the size of the image archive. In dyadic k-means clustering, we usually chose 128 clusters as a compromise between detail representation and computation time. An awkward effect across image feature clustering is that the entire computation has to be repeated if new data enter the archive. To overcome this limitation, an incremental clustering algorithm is required that just has to cluster the new data.

Even if across image classification is faced with the mentioned difficulties, it proved to be an efficient method to manage large amounts of data. Especially the separation between off-line unsupervised image feature clustering and on-line supervised learning (as it will be described in the following sections) makes across image classification an efficient method and justifies its practical use.

5.3 User-specific Semantic Labeling

The first 3 levels of our hierarchical modeling describe the image data D at level 0 in a completely unsupervised and application-free way (Fig. 4.1). Based on this objective representation, we can now link subjective user interests L_ν (level 4) to the signal classes ω_i by probabilities $p(\omega_i|L_\nu)$. For a robust characterization of user-specific semantics L_ν, several signal models (level 3) have to be applied.

Then, we link the elements of the joint space of signal classes to the user's interests. The stochastic link can be achieved with different models for $p(\omega_i|L_\nu)$, but only if we suppose a full statistic independence written as

$$p(\omega_{jk\ldots}|L_\nu) = p(\omega_j|L_\nu) \cdot p(\omega_k|L_\nu) \cdot \ldots \ , \tag{5.17}$$

a fast computation is possible. In the following, we will restrict ourselves to a statistic independence for $p(\omega_{jk\ldots}|L)$ with two models j and k.

With the results of unsupervised classification (level 3), we obtain the posterior probabilities $p(\omega_{jk}|D)$ for the signal classes ω_{jk} given the data D. With these results and the assumption that the signal characteristics of the semantic label L_ν are fully represented by ω_{jk}, we can calculate the posterior probability as

$$p(L_\nu|D) = \sum_{jk} p(L_\nu|\omega_{jk})\, p(\omega_{jk}|D) \ . \tag{5.18}$$

With Bayes' formula, Eq. 5.18 can further be expressed as

$$p(L_\nu|D) = p(L_\nu) \sum_{jk} \frac{p(\omega_{jk}|L_\nu)\, p(\omega_{jk}|D)}{p(\omega_{jk})} \ , \tag{5.19}$$

where $p(L_\nu)$ indicates the prior probability of semantic labels L_ν and $p(\omega_{jk}) = \sum_\nu p(\omega_{jk}|L_\nu)\, p(L_\nu)$ the prior of signal classes ω_{jk}. Since the posterior probability can be calculated for each image pixel, we can visualize $p(L_\nu|D)$. The spatial visualization of $p(L_\nu|D)$ is named in the following as "posterior map". This map gives the system operator a feedback of how strong and accurate the cover-type label has been already defined.

5.4 Interactive Learning

In order to make the inference from the image data D (level 0) to the cover-type labels L_ν (level 4), the system first has to learn the probabilistic link $p(\omega_i|L_\nu)$ based on user-supplied training samples. As mentioned in the last section, Eq. 5.17, we assume conditional independence for the signal classes ω_{jk} as a combination of 2 features. In the following we denote the classes by ω_i. We perform the probabilistic learning with a simple Bayesian network (Heckermann et al. 1994). Assume we

have a set of user-supplied training data T expressed by $\{N_1, \ldots, N_r\}$ with N_i being the occurrence of ω_i in T. Then, the vector of N_i has a multinomial distribution since ω_i is a variable with r states (BERNARDO and SMITH 2001), if we consider the parameter vector $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_r\}$ as a model for the set of probabilities

$$p(\omega_i|L, \boldsymbol{\phi}) = \phi_i \ . \tag{5.20}$$

Now, we change our discussion from determining the probabilities of the signal classes ω_i to the parameter vector $\boldsymbol{\phi}$. For a newly defined label we start with a constant initial prior distribution

$$p(\boldsymbol{\phi}) = \Gamma(r) = (r-1)\,! \ , \tag{5.21}$$

where r indicates the number of signal classes ω_i and $\Gamma(\cdot)$ the gamma function. With our observed training set T and its instances N_i, we obtain the posterior probability

$$\begin{aligned} p(\boldsymbol{\phi}|T) &= \frac{\Gamma(r+N)}{\prod_i \Gamma(1+N_i)} \prod_i \phi_i^{N_i} \\ &= \mathrm{Dir}(\boldsymbol{\phi}|1+N_1, \ldots, 1+N_r) \\ &= \mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha}) \ , \end{aligned} \tag{5.22}$$

with the total sum of training samples $N = \sum_i N_i$, the Dirichlet function $\mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha})$ and the hyper-parameters

$$\alpha_i = 1 + N_i \ . \tag{5.23}$$

If we observe another training set T' that is considered to be independent on T, we obtain by

$$\begin{aligned} p(\boldsymbol{\phi}|T', T) &= \frac{p(T'|\boldsymbol{\phi}, T)\, p(\boldsymbol{\phi}|T)}{p(T', T)} \\ &= \mathrm{Dir}(\boldsymbol{\phi}|\alpha_1 + N_1', \ldots, \alpha_r + N_r') \ , \end{aligned} \tag{5.24}$$

an additional update of the hyper-parameters by adding the number of times ω_i' occurs in the training dataset T':

$$\alpha_i' = \alpha_i + N_i' \ . \tag{5.25}$$

The initial state of the hyper-parameters is given by

$$\boldsymbol{\alpha}_0 = \{1, \ldots, 1\} \tag{5.26}$$

and a new set of training samples updates the hyper-parameters as given in Eq. 5.25.

Having a definition of the hyper-parameters $\boldsymbol{\alpha}$ by some training sets T, we can finally calculate the probabilities as expectation over all possible values of $\boldsymbol{\phi}$ as

$$p(\omega_i|L, T) = \frac{\alpha_i}{\sum_i \alpha_i} \ . \tag{5.27}$$

The fast computation of the probabilistic link $p(\omega_i|L_\nu)$, Eq. 5.27, and the updating after observing new training data, Eq. 5.25, make the hyper-parameters $\boldsymbol{\alpha}$ a very advantageous tool to describe the stochastic link between objective signal classes and subjective user semantics.

In order to allow high precision training specified on full resolution images, an on-line training interface has been developed (Fig. 5.7).

A human trainer can define an arbitrary number of (pairwise disjunct) cover-types L_ν and $\neg L_\nu$ (e.g. 'lake' and 'not lake') on a set of images in full resolution. After selecting a combination of signal classes of feature models, the trainer can ask for the posterior map of a particular cover-type label or an assessment of the selected features classes. Since the image content has already been extracted up to level 3, only the probabilistic link has to be re-calculated and the response is pretty fast. This allows an iterative refinement of the training regions and "simultaneous" observation of the consequences for the posterior probabilities.

5.5 Probabilistic Search

After training a semantic cover-type, we want to search the entire data repository for relevant images. We can calculate the posterior probability of L_ν given a particular image I_ζ as

$$p(L_\nu|I_\zeta) = \sum_i p(L_\nu|\omega_i)\, p(\omega_i|I_\zeta) \tag{5.28}$$

in an analogous manner as we calculated the posterior probability of L_ν given a particular data D, Eq. 5.19. The posterior probability is a measure of how probable an image "is of" a particular cover-type.

To provide a more practical measure for image retrieval, we compute the "coverage"

$$C = \sum_i p(\omega_i|I_\zeta) \cdot \text{Heaviside}(p(L_\nu|\omega_i) - p_{\text{th}}) \tag{5.29}$$

which specifies the approximate percentage of the image that definitely contains the desired cover-type. The degree of "definitely" is determined via the threshold p_{th}.

Since the distribution of $p(\omega_i|L_\nu)$ — resulting from limited training data — is known in detail, we can specify both the probability of a label in a particular image and the expected degree of variation. We do this by calculating the expected variance of the posterior

$$\delta^2 p(L_\nu|I_\zeta) = \sum_i \delta^2 p(L_\nu|\omega_i) p(\omega_i|I_\zeta) \tag{5.30}$$

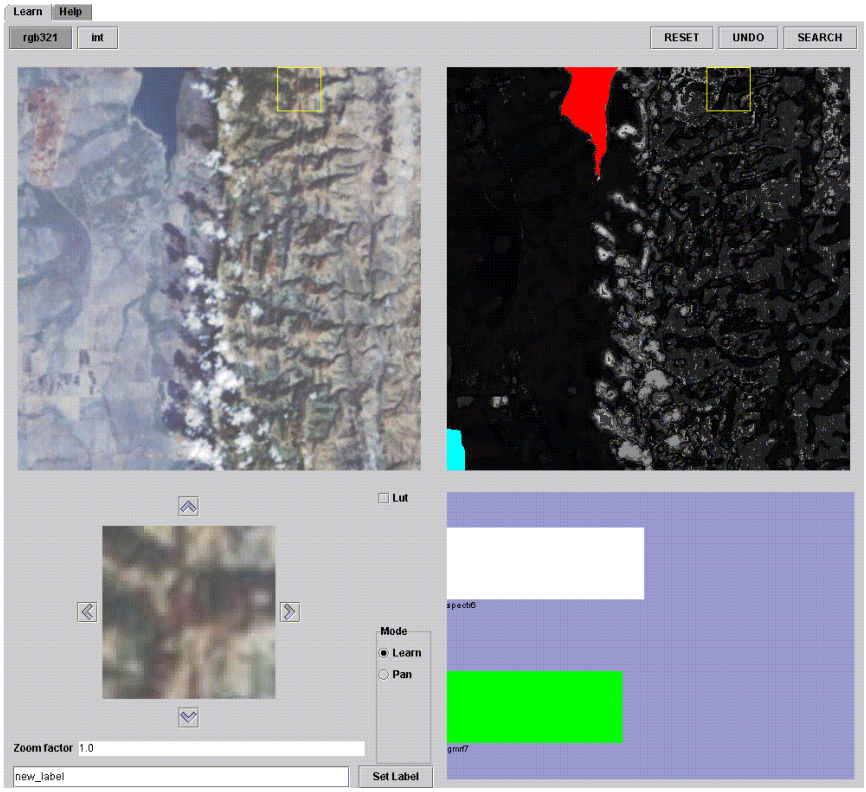


Figure 5.7: Graphical user interface (GUI) in I^{2}M for interactively learning the contents (water) of remote sensing images. The system operator can specify his interests by giving positive and negative training samples, either directly into the original image (left top), the zoom window (left bellow) or the posterior map (right top). After each mouse click, the hyper-parameters $\boldsymbol{\alpha}_\nu$ and the probabilities $p(\omega_i|L_\nu)$ of the stochastic link are updated and the posterior map on the right side is re-computed and re-displayed. As a further quality measurement, we display the divergence, Eq. 7.20, between positive and negative training samples for individual signal models (right bellow). If the user is satisfied with the posterior map, he can search the entire archive for relevant images by clicking on the 'SEARCH' button.

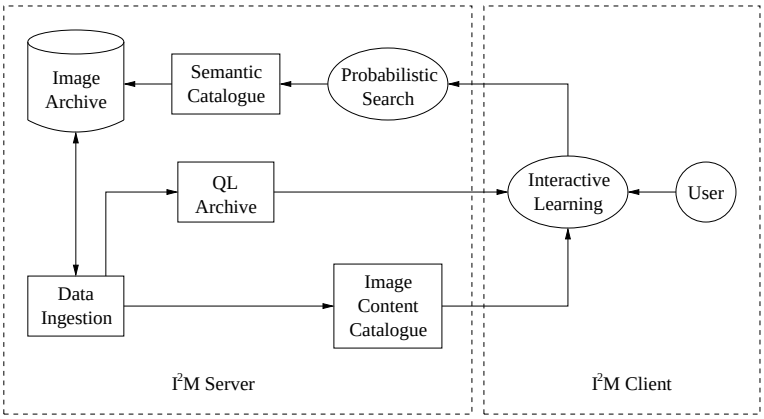


Figure 5.8: Client-server architecture of the image information mining system.

with the δ^2-symbol denoting the variance. As a measure of how well L_ν is separated from $\neg L_\nu$ in a particular image I_ζ we use the separability

$$S(L_\nu|I_\zeta) = \frac{\delta^2 p(L_\nu|I_\zeta)}{p(L_\nu|I_\zeta)(1-p(L_\nu|I_\zeta))} \ , \tag{5.31}$$

which is the variance in units of the maximal possible variance. The smaller $S(\cdot)$, the "better" we call the separability. The separability measurement is very useful for further learning, since retrieved images with low separability are related to performed positive training samples and images with high separability are connected to negative training. The user can either decide to enforce the positive training because of bad query results for low separability or enforce negative training due to bad search results for high separability.

5.6 System Description and Configuration

In this section we describe the mining system from the technical point of view with its main components as illustrated in (Fig. 5.8). To access the system, a user has to register first by choosing a user id and a password. After successful login, a personal welcome page is displayed. The user can decide to perform some administration to start the interactive learning process. The latter requires the selection of a combination of up to 4 signal models. In the mining system, the information extracted from one single sensor as well as the information from multiple sensors can be used for interactive training and probabilistic search. Having selected a certain model combination, the user has to pick out a starting image from a gallery of randomly chosen images. If the gallery does not contain an image of the user's favour, he can choose another set of random images.

Once clicked on an image, the interactive learning process begins as depicted in (Fig. 5.9). In a first step, the following objects are downloaded by the interactive

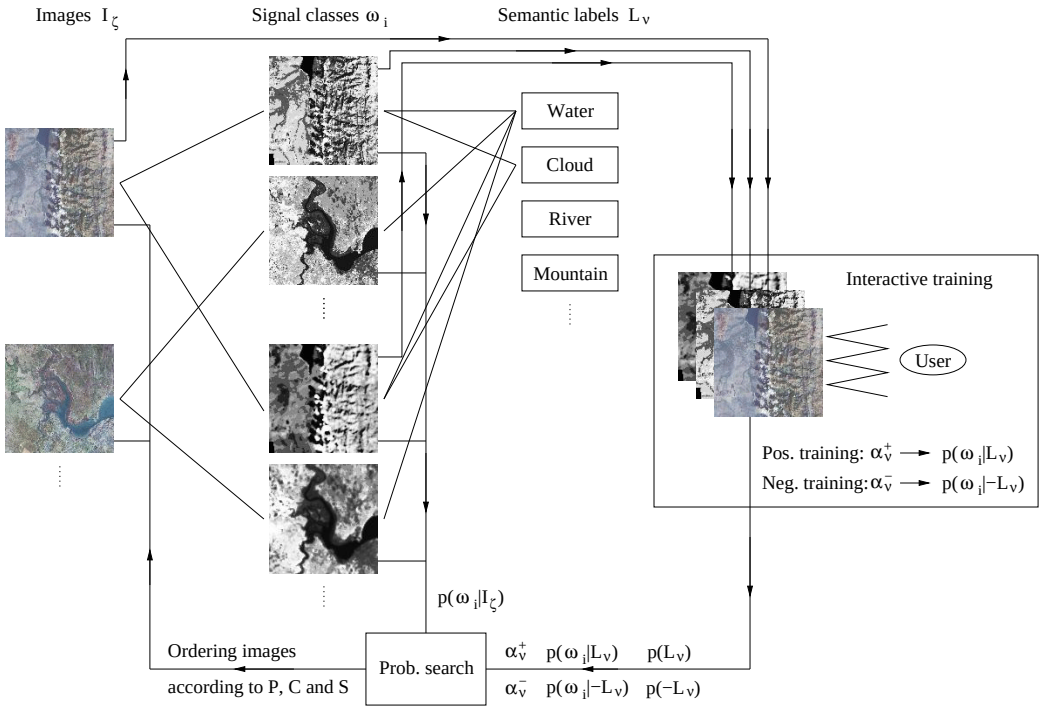


Figure 5.9: Data flow during interactive learning and probabilistic search. After identifying the user and choosing a combination of signal models the learning applet is downloaded from the server to the client browser. The system operator can continuously train a specific label of his interest by giving positive and negative training samples. After each training (mouse click), the hyper-parameters $\boldsymbol{\alpha}_\nu$ are updated and the re-displayed posterior map indicates the current state of the label. If the user is satisfied with the trained label, he can query the entire image archive for the defined cover-type. The system delivers the top-ranked images according to coverage (C), posterior probability (P) and separability (S). For further label definition, the user can select another image from the search results.

learning applet: the QL image in JPEG-format and the corresponding classification maps (image content catalogue) for the selected signal models in raw binary format. When the downloading is finished (after a few seconds), the user can start the definition of a semantic cover-type of his interest by giving positive and negative samples using the left/right mouse button. After each click, the hyper-parameters $\boldsymbol{\alpha}_\nu$, the likelihoods $p(\omega_i|L_\nu)$ and the posterior map are updated. The latter permanently gives the user an intuitive feedback about the quality of the learning process by marking regions corresponding to the cover-type with red colour. If the current label definition is satisfactory, the system operator can query the entire archive for images containing similar structures or objects. For the computation of the probabilistic search measurements on the server's site only the hyper-parameters

with the derived posterior probabilities $p(L_\nu|\omega_i)$ and the probabilities $p(\omega_i|I_\zeta)$ of the generated and inserted catalogue entries are necessary. At this time, the label is persistently stored in the data base. The definition of the label is given by its name, the used image from training, the selected signal models, the hyper-parameters and the resulting (queried) images.

The user can continue the learning process until he is satisfied with the query result. In order to improve the definition of the semantic label, the system operator clicks on another image in the resulting image set and continues to feed in positive and negative examples. Everytime the user selects an image from the query gallery, the QL and the assigned signal model classification maps are transmitted via the world wide web. We want to point out that the cover-type learning using several images is important to obtain a well-defined semantic label. We call this "iterative incremental learning". Each time the user queries the image archive, the semantic label definition in the database is updated.

A tool worth to be mentioned in I^{2}M is the tracking module that stores each human-machine interaction in the database. Based on the stored information, the system computes a number of statistical and information theoretical measures that indicate the goodness of the learning process. These measures give the user a further feedback about the learning progress (lower-left part in Fig. 5.7). The traced and stored human-computer interactions are the central element for the overall system evaluation. In Chap 7, we will demonstrate methods to measure the complexity of human-system operations, to identify target structures, to determine the convergence of the learning process and to predict in which semantic cover-type(s) the user might be interested in. After defining these methods, we will analyze in Chap. 8 their relevance in a large-scale system performance evaluation by comparing objective measures with subjective user satisfaction.

5.7 Practical Applications

The applied concept of unsupervised indexing of image content and the user-specific semantic labeling of cover-types have been extensively tested based on various remote sensing datasets (Tab. 5.1). In the performed experiments, the image data range from monochromatic high-resolution (Ikonos) to hyperspectral (Daedalus ATM) data and from medium-resolution SAR (ERS1) to high-resolution polarimetric (E-SAR) image data. The fusion of different signal models from one sensor as well as the fusion of multi-sensor image data is applied for interactive learning and probabilistic retrieval. With it, we want to demonstrate the power of I^{2}M for data-independent image mining applications.

In the following, we show examples of labeling user-defined semantics and query results from the image archive. We start with the analysis of a cover-type 'mountain' that was trained with different combinations of signal models as shown in (Fig. 5.10).

Sensor	Landsat TM	ERS1	Ikonos	Landsat TM
Coverage	Mozambique	Mozambique	Mozambique	Nepal
No. of scenes	14	32	9	1
No. of images	438	438	207	144
Channels, size	$6, 2000^2$	$1, 2000^2$	$1, 2000^2$	$8, 500^2$
Resolution	25m, geo./co.	25m, geo./co.	1m, geo.	25m, geo.
Signal models	spectral, GRF at 3 scales	EMBD/ GMRF at 2 scales	spectral, GRF at 3 scales	spectral, GRF at 2 scales

Sensor	Landsat TM	Ikonos	Ikonos	Daedalus	E-SAR
Coverage	Switzerland	Germany	Germany	Kosovo	Kosovo
No. of scenes	4	1	1	1	1
No. of images	184	43	43	12	12
Channels, size	$7, 1024^2$	$1, 500^2$	$4, 500^2$	$12, 1000^2$	$4, 1000^2$
Resolution	25m, geo.	1m, geo./co.	4m, geo./co.	1m, geo./co.	1m, geo./co.
Signal models	spectral, GRF at 5 scales	spectral, GRF at 3 scales	spectral	spectral, GRF at 3 scales	EMBD/ GMRF at 2 scales

GRF = Gibbs random field (auto-binomial model)
GMRF = Gauss-Markov random field model
EMBD = Enhanced model-based despeckling
geo. = geocoded
geo./co. = geocoded and co-registered
pol. = polarimetric

Table 5.1: Ingested datasets in I^2M and the applied signal models.

The selected combination of signal models influences both the level of compactness and detail of the semantic label. The retrieved images for the defined cover-type 'mountain' are given in (Fig. 5.11). By default, only the highest 6 top-ranked images are delivered for probability, coverage and separability, but the user can ask for more results.

User-specific interactive learning with information from multiple sensors can be used for sensor qualification and further exploration of the image dataset. For this, the interactive training with high-resolution image data is exemplified in (Fig. 5.12). In a final application we show the classification and retrieval of the label 'water' from co-registered high-resolution hyperspectral and polarimetric radar data (Fig. 5.13). The applied signal models are spectral from the hyperspectral data and from E-SAR the despeckled SAR backscatter (L-band, scale 2m), the despeckled SAR backscatter (L-band, scale 2m) and the norm of the SAR texture vector (L-band, scale 4m). With an increasing number of signal models the number of structural details grows.

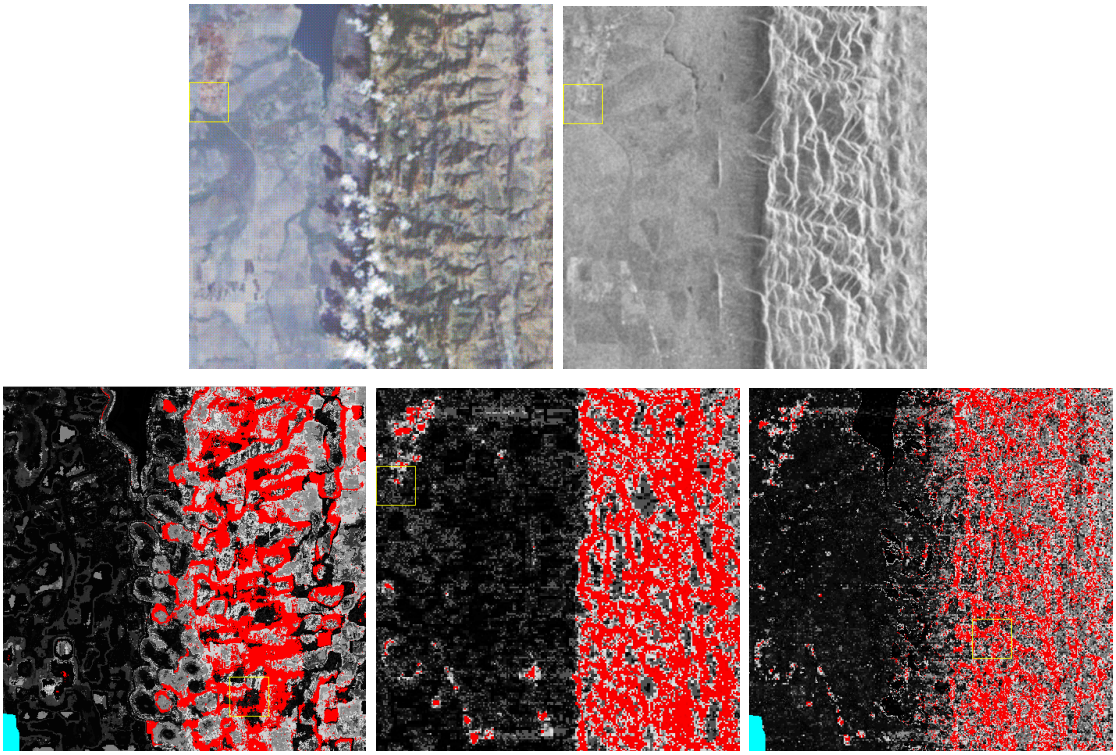


Figure 5.10: Interactive training of mountainous areas with different combinations of signal models. Regions belonging to the semantic label are marked with red colour. Images from left to right: 1st row: QL from Landsat TM and ERS1 image data. 2nd row: Trained semantic label with spectral and texture information from Landsat TM, the obtained results with only texture features from ERS1 at different scales and the defined cover-type with across sensor model combination of spectral (Landsat) and texture (ERS1).

5.8 Conclusions

In this chapter, we have discussed the following items:

- We addressed the concept underlying the I^{2}M system of unsupervised image content modeling and supervised semantic labeling, and demonstrated its performance in several practical applications. The system started with the extraction of primitive features from different kinds of remote sensing image data by using stochastic parametric signal models. Particularly Gibbs-Markov random fields showed to be useful to describe spatial characteristics in images.

- Having extracted visual parameters, the system clusters them in an unsupervised way, derives a characteristic vocabulary for the different feature models, and generates the catalogue entries for the DBMS. With the described dyadic

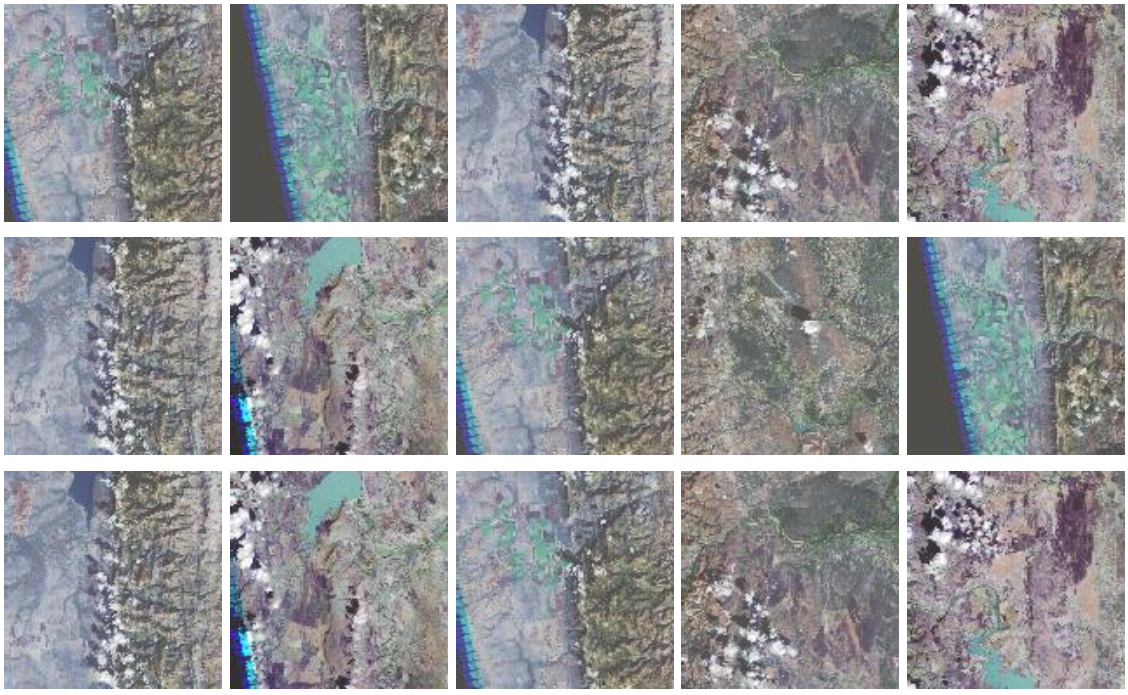


Figure 5.11: Results of probabilistic search for the trained semantic label 'mountain' in Fig. 5.10, 2nd row, left. The queried images are ranked according to coverage (1st row), Eq. 5.29, posterior probability (2nd row), Eq. 5.28 and separability (3rd row), Eq. 5.31. The user can continue training the cover-type by selecting one of the retrieved images.

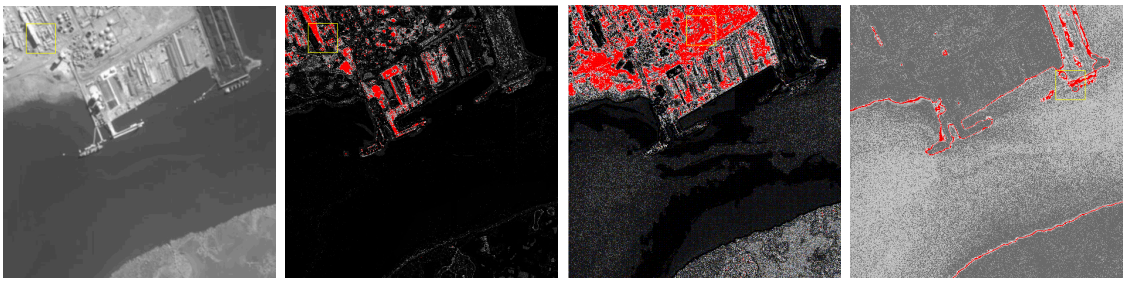


Figure 5.12: Interactive training of semantic labels in an Ikonos image with spectral and texture information. Images from left to right. Quick look of a panchromatic Ikonos image with a resolution of 1m and trained semantic labels 'industrial area', 'grassland' and 'coastline'. Consider that the defined semantic labels were obtained with just a few training samples.

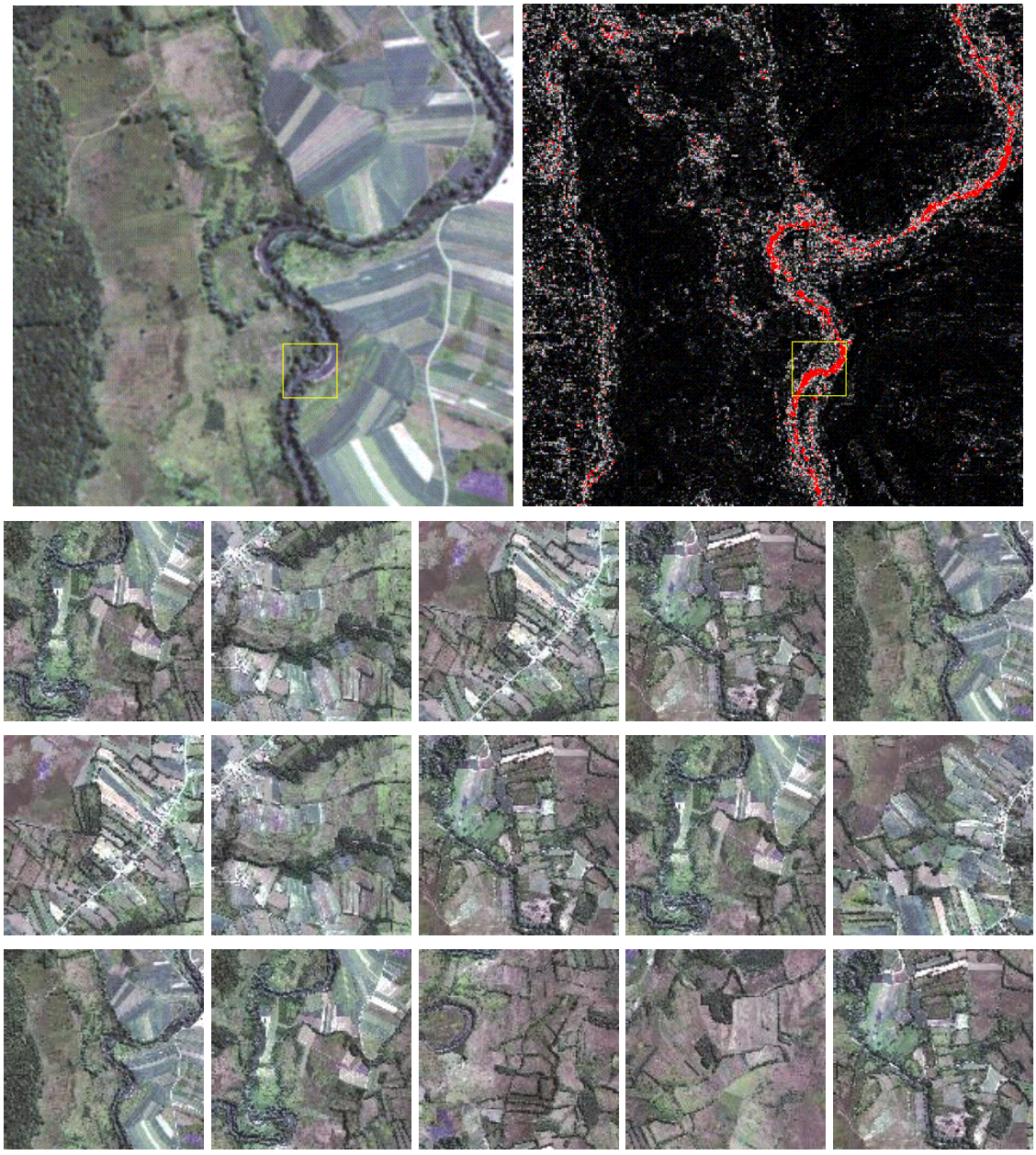


Figure 5.13: Interactive learning and probabilistic search of semantic label 'water' from hyperspectral and polarimetric data using 4 different signal models. Images from left to right: quick look and posterior map (1st row), top-ranked retrieved images according to coverage (2nd row), posterior probability (3rd row) and separability (4th row).

k-means clustering algorithm we presented an efficient mining tool to reduce the large amounts of primitive features. Particularly the tree structure of the algorithm and the potential for parallelization may lead to a further speed-up.

- In a final step, the user's interests in form of semantic image content interpretation were linked to the content-index by using simple Bayesian networks. By fusing information of the different feature models, the system generates a supervised classification of the whole archive and retrieves the most relevant images.

- The applied concept of unsupervised image content description and supervised cover-type training is very advantageous since it enables the separation between the time-consuming content-index generation and the fast interactive learning. Even if the algorithms for feature extraction and clustering make great demands on the computational performance, they do not impact the system since the images are processed at the time of data ingestion and processes can be distributed accordingly. On the other side, the computational complexity of the Bayesian learning algorithm — updating the hyper-parameters and their normalization — at the client's site is rather low.

- In order to open new applications for the mining system and to make it an operational tool for discovering and understanding highly complex data, an extended verification and evaluation procedure is necessary.

6

Semantic Grouping and Category Learning

With the learning paradigm from the preceding chapter we can link user-specific semantic cover-type labels L_ν to signal classes ω_i of the unsupervised content-index by probabilities $p(\omega_i|L_\nu)$ that are derived from the user's feedback. However, the success of this scheme for data fusion and interactive learning depends on the fact that a semantic cover-type is fully described by the content-index. In order to be independent of the applied datasets, we therefore introduce another level, level 5, of information representation: semantic grouping of individual cover-type labels (Fig. 4.1). The elements at this level are composed of different semantic cover-type labels of level 4 and are denoted as Λ_τ hereafter.

Before we deal with the modeling of semantic classes Λ_τ in Sec. 6.2, we will first point out the prerequisites of the new method for image content modeling in Sec. 6.1. Then, in Sec. 6.3, we expand the scheme of supervised semantic grouping at the indexing of heterogenous collections of images and in Sec. 6.4 we deal with the learning of ontological categories.

6.1 Prerequisites and Motivation

While the fusion of image information computed from data of a specific sensor is not crucial, the combination of image attributes extracted from co-registered image data is only valid under certain conditions. In the case of a low image co-registration quality and very different resolutions between the image datasets, the approach to fusing information using simple Bayesian networks produces poor results. Furthermore, the approach from the previous section does not allow the application to non-registered images. The limitation of accessing image data across sensors and data collections can be circumvented with the new level of image content abstraction as depicted in (Fig. 6.1).

Another limitation is that only the training of specific semantic cover-types such as 'river' or 'forest', is supported, but not complex structures or objects of higher-level semantic concepts, e.g. 'a city on a river'. The fusion of such contradictory

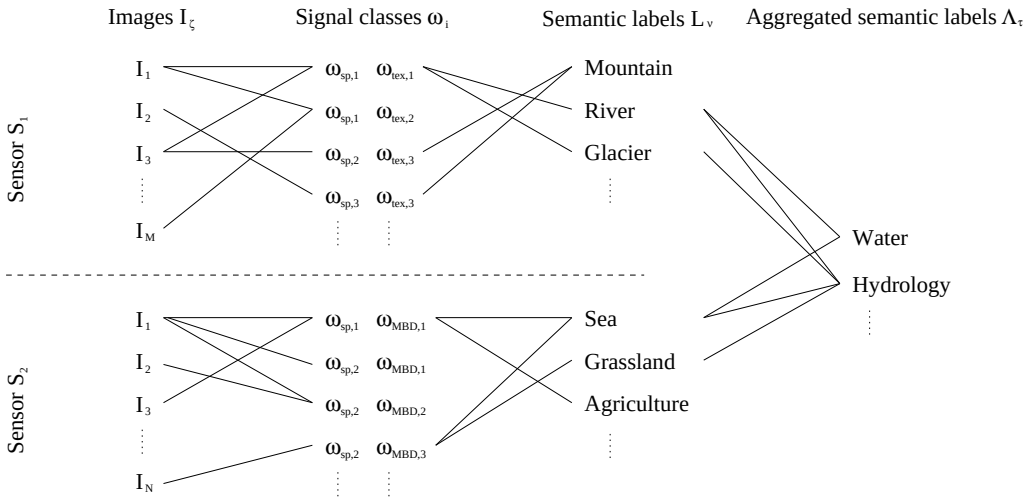


Figure 6.1: Semantic grouping of cover-type labels defined on two different datasets. Note, cover-type labels L_ν at each dataset are linked via a specific content-index ω_i to images I_ζ while the aggregated semantics are valid across the whole archive.

image data in the Bayesian classifier results in poor and non-acceptable classifications. To demonstrate this difficulty, we show the semantic labeling of cover-types 'river', 'sea' and the training of both in (Fig. 6.2). While the individual semantic labels can be well-separated from the other content, the joined training of 'sea' and 'river' reduces the classification accuracy and also marks non-relevant structures. As we can see in the example, the learning of aggregated semantics at signal class level causes problems, but the learning of individual labels and a fusion afterwards yields better results. Thus, we define a new level of semantic abstraction, where we connect user-specific semantic cover-type labels to complex and higher-level semantic concepts.

Another reason for aggregating individually trained cover-types at semantic level is time. In order to obtain a well-defined semantic label, the user needs various iterations based on several images (iterative incremental learning). With the new method the time for querying the archive can be considerably reduced as the user only has to weight already existing labels in the semantic inventory.

6.2 Modeling Semantic Classes

In this section, we carry out a scheme for the aggregation of cover-type labels to complex and higher-level semantics.

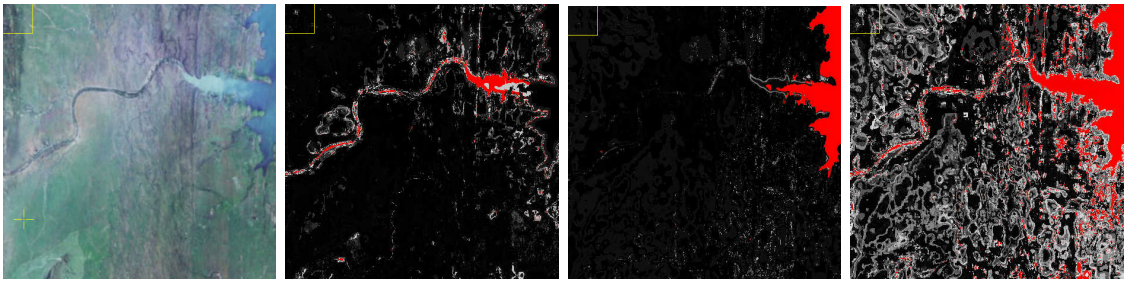


Figure 6.2: Interactive learning of different semantic labels. Images from left to right: original quick look, trained cover-type 'river', label 'sea' and the common classification of 'sea' and 'river'. While the separate classification of 'sea' and 'river' yields good results, the training of the two labels together is difficult and results in many irrelevant structures.

Inference from data (level 0) to grouped semantics (level 5)

Based on Eq. 5.18, where we associated semantic labels L_ν to the image data, we can now link higher-level semantic concepts Λ_τ to the image data D by

$$p(\Lambda_\tau|D) = \sum_\nu p(\Lambda_\tau|L_\nu)\, p(L_\nu|D) \ . \tag{6.1}$$

In accordance with Sec. 5.3, we can apply Bayes' formula to express the posterior probability of Eq. 6.1 as

$$p(\Lambda_\tau|D) = p(\Lambda_\tau) \sum_\nu \frac{p(L_\nu|\Lambda_\tau)\, p(L_\nu|D)}{p(L_\nu)} \tag{6.2}$$

with $p(\Lambda_\tau)$ as the prior probability of aggregated label Λ_τ and the prior of semantic cover-types $p(L_\nu) = \sum_\tau p(L_\nu|\Lambda_\tau)p(\Lambda_\tau)$.

Before we can link aggregated semantics to individual cover-type labels by probabilities $p(L_\nu|\Lambda_\tau)$, we make the connection between cover-types L_ν and image data D as follows:

$$p(L_\nu|D) = \sum_i p(L_\nu|\omega_i)\, p(\omega_i|D) \ . \tag{6.3}$$

This calculation is a computationally intensive task since it demands the summation over the joint class space ω_i. Especially for a large number of signal classes and cover-type labels this computation is not tractable and therefore requires an approximation. We do this by computing the maximum a-posteriori classification of the image data and obtain

$$p(L_\nu|D) \approx p(L_\nu|\omega_i) \ , \tag{6.4}$$

where the training samples of given cover-type labels L_ν serve as input training data for the Bayesian classification algorithm. To find out to which cover-type a data

Input: semantic cover-types given by $p(\omega_i|L_\nu)$,
set of N data points with signal classes $\omega_i\,, i = 1, \ldots, r$
Output: set of N classified data points
Method: 1) compute the discrimination functions
$g_\nu(\omega_i) = p(L_\nu|\omega_i)$ for all c individual labels L_ν
3) **for** $k = 1, \ldots, N$
4) assign each point to semantic label L_ν if
5) $g_\nu(\omega_i) > g_\mu(\omega_i)$ **and** $g_\nu(\omega_i) > \frac{1}{2} \quad \forall\, \nu \neq \mu$
6) **end**

Figure 6.3: Pseudo-code of the supervised Bayesian classifier.

sample belongs, we have to define discriminant functions $g_\nu(\omega_i)\,, \nu = 1, \ldots, c$, one for each semantic label L_ν. In the case of Bayesian classification, where $p(L_\nu|\omega_i)$ serve as posterior probabilities, we can define the following simple discrimination functions (Duda et al. 2001)

$$g_\nu(\omega_i) = p(L_\nu|\omega_i) \;\;, \tag{6.5}$$

where $p(L_\nu|\omega_i)$ are computed for each cover-type label using the user's positive and negative training samples. Then, the classifier assigns image data with a particular signal class ω_i to a cover-type label L_ν if

$$g_\nu(\omega_i) > g_\mu(\omega_i) \quad \forall\, \nu \neq \mu \;\;. \tag{6.6}$$

This procedure results in supervised classification maps with as many semantic classes as labels. In (Fig. 6.3), we show the pseudo-code for the supervised Bayesian classification algorithm and in (Fig. 6.4) the obtained classification map for a single quick look. In order to avoid poor classification, we introduce another criterion for label membership as $p(L_\nu|\omega_i) > 0.5$.

Posterior maps of aggregated semantic cover-types

We can visualize the posterior probabilities $p(\Lambda_\tau|D)$ in the same way as for $p(L_\nu|D)$ and call it "semantic posterior map". Based on this visualization, the user gets an intuitive feedback about the relevance of his/her semantic grouping. The stochastic link $p(L_\nu|\Lambda_\tau)$ is derived from the user's training samples using a vector of hyper-parameters $\boldsymbol{\alpha}$ as outlined in the following.

Interactive learning of the probabilistic link $p(L_\nu|\Lambda_\tau)$

As previously stated, before we can aggregate defined semantic labels, we have to learn the probabilities $p(L_\nu|\Lambda_\tau)$ from the user's feedback samples. Therefore we

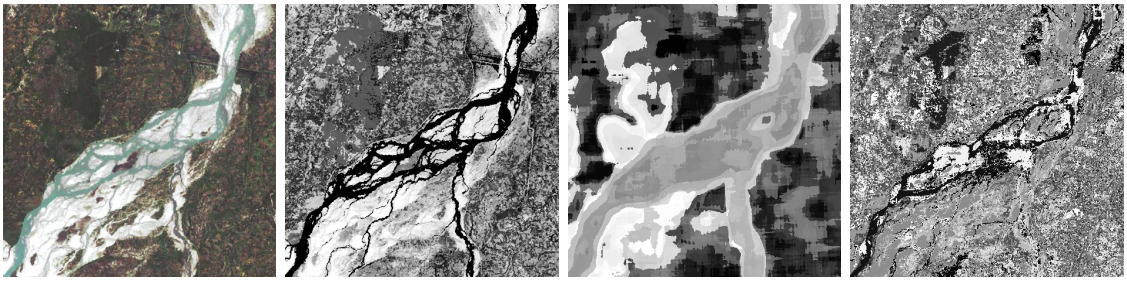


Figure 6.4: Generation of semantic cover-type maps. Images from left to right. original quick look image (level 0), unsupervised classification maps (level 3) of spectral feature model, unsupervised classification map of textural feature model and the supervised classification map based on a couple of user-specific semantic labels (level 4) ingested in the database. Note that semantic labels are trained with different combinations of signal feature models, e.g. spectral, spectral and texture or texture at different scales. In order to avoid poor supervised classification maps, image regions (pixels) that are only weakly touched by semantic labels remain unclassified.

use a modified version of the Bayesian learning algorithm as applied to derive the stochastic link $p(\omega_i|L_\nu)$ between signal classes and cover-type label. We start again with the assumption of having a set of c probabilities $p(L_\nu|\Lambda_\tau)$ derived from the user's feed-back T given by $\{N_1, \ldots, N_c\}$, where N_ν indicates the weight factor assigned to label L_ν in T. In order to provide an efficient and robust approach for semantic aggregation, we assign a certain state to each N_ν and define the following probabilistic model

$$p(L_\nu|\Lambda_\tau, \boldsymbol{\psi}) = \psi_\nu \tag{6.7}$$

with the parameter vector $\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_c\}$. If an aggregated semantic Λ_τ is newly defined, we obtain for the prior

$$p(\boldsymbol{\psi}) = \Gamma(c) = (c-1)! \tag{6.8}$$

a constant distribution with c as the number of semantic cover-types and the gamma function $\Gamma(\cdot)$. After the user's first feedback, the posterior probability of $\boldsymbol{\psi}$ is given as

$$\begin{aligned} p(\boldsymbol{\psi}|T) &= \frac{\Gamma(c+N)}{\prod_\nu \Gamma(1+N_\nu)} \prod_\nu \psi_\nu^{N_\nu} \\ &= \mathrm{Dir}(\boldsymbol{\psi}|1+N_1, \ldots, 1+N_c) \\ &= \mathrm{Dir}(\boldsymbol{\psi}|\boldsymbol{\alpha}) \end{aligned} \tag{6.9}$$

with Dirichlet function $\mathrm{Dir}(\cdot)$ and the hyper-parameters

$$\alpha_\nu = 1 + N_\nu \ . \tag{6.10}$$

If the user performs another learning iteration, the probabilities are updated according to

$$p(\boldsymbol{\psi}|T,T') = \mathrm{Dir}(\boldsymbol{\psi}|\alpha_1 + N'_1, \ldots, \alpha_c + N'_c) \ , \tag{6.11}$$

where T' denotes the new set of training samples. The new (updated) hyper-parameter can be expressed as

$$\alpha'_\nu = 1 + N'_\nu \ . \tag{6.12}$$

Starting with the initial state of the hyper-parameter

$$\boldsymbol{\alpha}_0 = \{1, \ldots, 1\} \ , \tag{6.13}$$

we can describe the link between cover-types and aggregated semantic labels by

$$p(L_\nu|\Lambda_\tau) = \frac{\alpha_\nu}{\sum_\nu \alpha_\nu} \ , \tag{6.14}$$

where the α_ν are updated as given in Eq. 6.12. With this definition, the computation of probabilities $p(L_\nu|\Lambda_\tau)$ is rather simple and allows their on-line application. From the perspective of data transfer, semantic aggregation is less time-consuming than interactive learning of user-specific semantic labels. With the new method only one supervised classification map has to be transferred whereas 'standard' learning requires an unsupervised classification map for each signal model.

In (Fig. 6.5), we illustrate the graphical interface applied. The user defines an aggregated label by associating weights for and against labels L_ν. After each assigned weight the semantic posterior map is updated and gives the user an intuitive feedback about the quality of the performed training iteration. In (Fig. 6.6), we show an example sequence of several iterations performed by a user to group 'river' and 'riverbed' labels to higher semantic concepts. After just a few interactions, one obtains good results that can be further used to search the archive for relevant images.

Probabilistic search

Searching for relevant images in the archive is not only confined to user-specific semantic cover-types, but can be extended to aggregated semantics, too. In principle, we can apply all three similarity measurements — posterior probability, coverage and separability — in practical experiments, however, we found out that only posterior probability delivers satisfying search results. We obtain the posterior probability of Λ_τ given a particular image I_ζ as

$$p(\Lambda_\tau|I_\zeta) = \sum_\nu p(\Lambda_\tau|L_\nu)\, p(L_\nu|I_\zeta) \ , \tag{6.15}$$

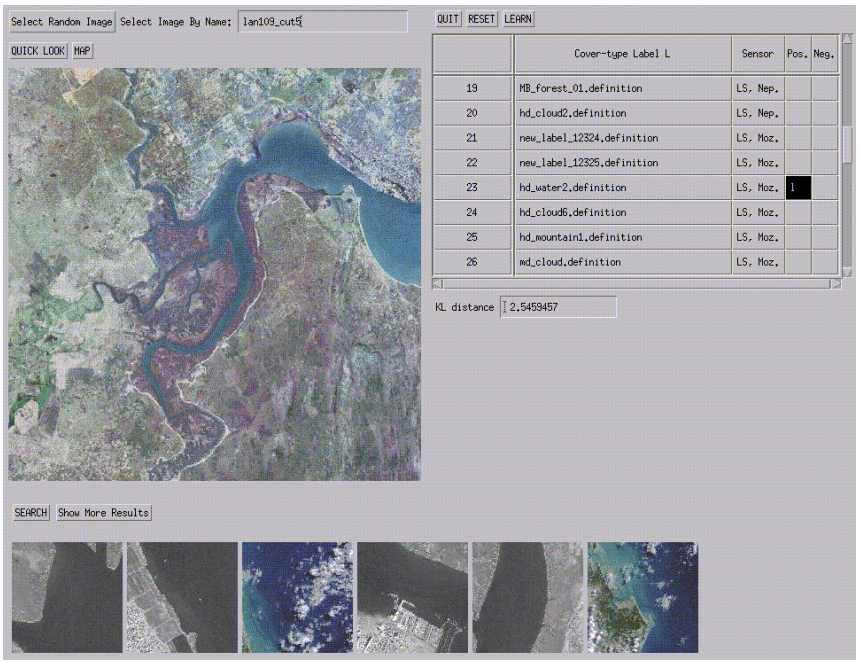


Figure 6.5: Experimental off-line graphical interface for semantic grouping of user-specific cover-type labels. The example image is a Landsat TM scene covering Mozambique, where the user aggregates the semantic classes 'river' and 'riverbed'. Grouping is performed by giving weights for and against semantic labels (table, right). The user can switch between quick look and semantic posterior map representation to get a feedback about the relevance of the performed training step. The posterior map is computed according to Eq. 6.2. In (Fig. 6.7), we illustrate top-ranked Landsat TM and Ikonos images together with their posterior maps.

where $p(L_\nu|I_\zeta)$ denotes the frequency of cover-type label L_ν in a certain image. Note, the complexity of searching the archive at semantic level, Eq. 6.15, is much less than searching at signal class level, Eq. 5.28, since the summation only has to be performed for few cover-types in comparison to the joint space of signal classes. Thus, semantic grouping is an efficient method for querying large image archives.

Remaining problems

We demonstrated an advanced concept of applying user-specific cover-type labels and grouping them into higher-level semantic concepts. The method aimed at overcoming some of the limitations that impair interactive learning of semantic labels based on signal classes. Examples show that this approach delivers promising results. However, we should not forget to mention the prerequisites and shortcomings of the method outlined. First, the performance of semantic grouping depends on the

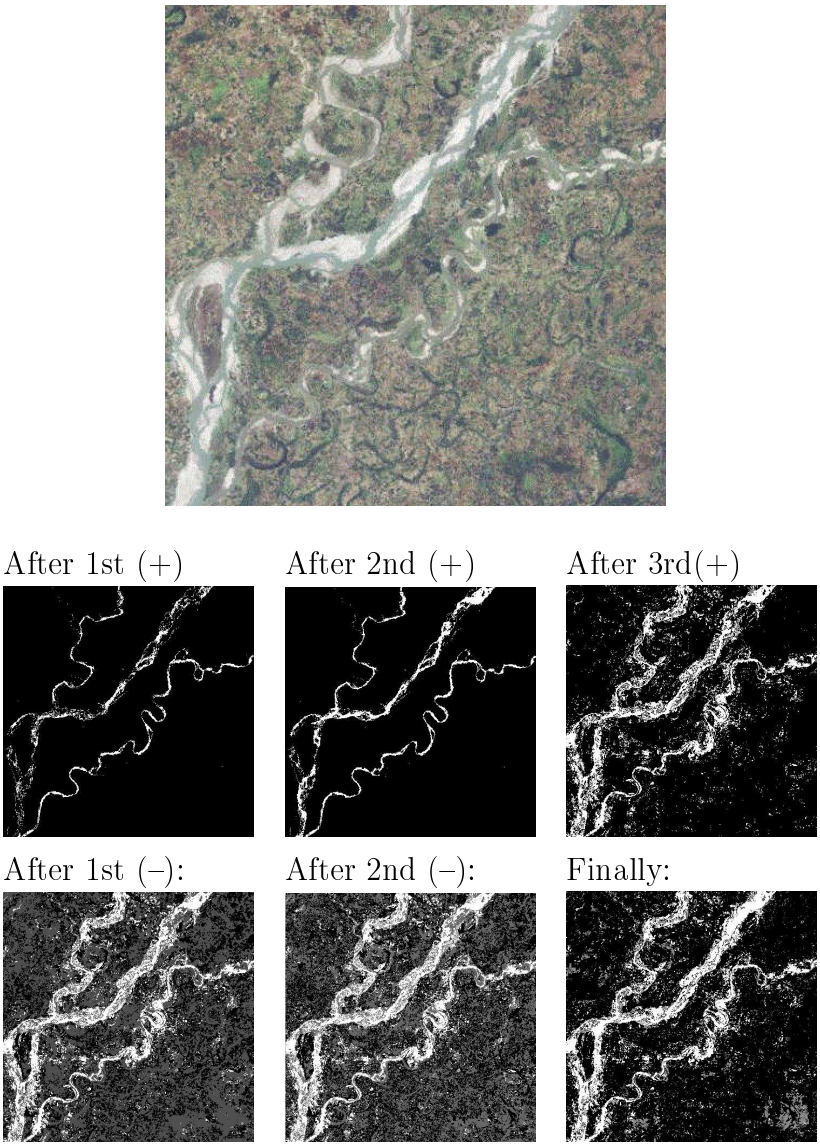


Figure 6.6: User example sequence for interactive grouping of semantic labels. The upper image shows the original Landsat TM quick look and the next two rows depict the development of the semantic posterior map after the user's feedback. After three positive iterations, the user followed with two negative ones to exclude irrelevant structures. After finishing the aggregation, the user can search the entire archive for relevant images containing the defined content. Note that the semantic labels used for aggregation are defined by various users and different combinations of signal models.

quality of the given semantic labels. If the labels are well-defined for the datasets, semantic grouping yields good results. Problems occur if the labels are of poor quality or if they are associated with the wrong semantic meaning, e.g. a user trained a cover-type 'coastline' and stored it as 'water'. To overcome these problems and to make the method more robust for large-scale practical applications, the database index must be modeled as a multidimensional dynamic process. A function that prevents poor labels from contribution must be implemented, too.

A minor limitation is that supervised classification maps have to be computed from existing cover-type labels for all images in the archive. However, the time requirements are not a limiting factor and computing semantic cover-type maps can be repeated regularly.

6.3 Semantic Grouping of Heterogenous Image Collections

A limiting factor of interactive learning of user-specific semantic cover-types is that labels are only valid for a certain collection of images which is defined via the global content-index. Only images in this collection can be linked to the semantic label and queried, but not images from other collections. Sometimes, however, it is helpful to query the overall archive for images that are covered by certain labels, regardless of which image collection they belong to, e.g. to group the semantic labels 'river' and 'lake' from Landsat TM with 'water tank' from Ikonos. The demonstrated method of semantic grouping of individual cover-type labels provides an efficient way to query the entire archive for relevant data.

Assuming we have two different sets of cover-type labels

$$L^{S_1} = \{L_1^{S_1}, L_2^{S_1}, \ldots, L_m^{S_1}\} \tag{6.16}$$

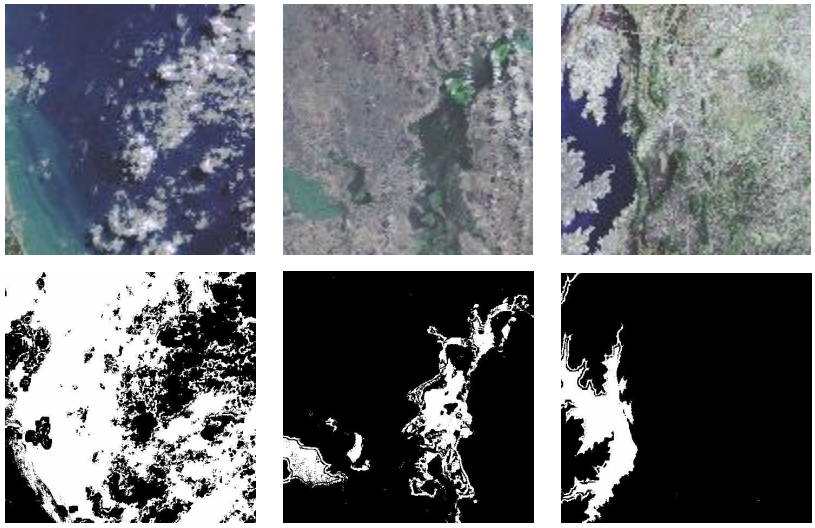
and

$$L^{S_2} = \{L_1^{S_2}, L_2^{S_2}, \ldots, L_n^{S_2}\} \ , \tag{6.17}$$

where the m labels from set L^{S_1} refer to sensor S_1 and the n labels from L^{S_2} to S_2 as depicted in (Fig. 6.1), we apply the described concept of semantic grouping to index images from both collections. We follow the same formalism as in the previous section and express the user's feedback T by $\{N_1, \ldots, N_{m+n}\}$ with N_ν being the occurrence of label N_ν in T. Again, we model the probabilities $p(L_\nu|\Lambda_\tau, \boldsymbol{\psi})$ in terms of a parameter vector $\boldsymbol{\psi}$ as outlined in Eq. 6.7. We model a new aggregated semantic label with the prior distribution

$$p(\boldsymbol{\psi}) = \Gamma(m+n) = (m+n-1)! \tag{6.18}$$

Landsat TM, Mozambique, resolution: 25m



Ikonos, Mozambique, resolution: 1m

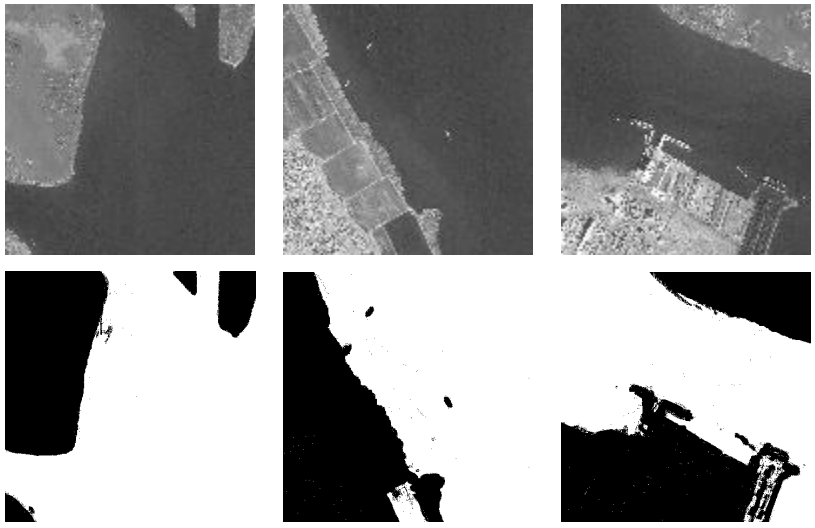


Figure 6.7: Semantic grouping of heterogenous collections of images using semantic labels 'water' of Landsat TM and 'river' of Ikonos. The first and second row show the most covered Landsat TM images of the Mozambique dataset together with the semantic posterior maps. Row 3 and 4 depict the most relevant Ikonos quick looks covering 'river', again with posterior maps.

with Gamma function $\Gamma(\cdot)$ and $m+n$ semantic labels. After obtaining a training set T with its instances N_ν, we can express the posterior probability as

$$\begin{aligned} p(\boldsymbol{\psi}|T) &= \frac{\Gamma(m+n+N)}{\prod_\nu \Gamma(1+N_\nu)} \prod_\nu \psi_\nu^{N_\nu} \\ &= \mathrm{Dir}(\boldsymbol{\psi}|1+N_1,\ldots,1+N_{m+n}) \\ &= \mathrm{Dir}(\boldsymbol{\psi}|\boldsymbol{\alpha}) \ , \end{aligned} \tag{6.19}$$

again with Dirichlet function $\mathrm{Dir}(\cdot)$, total number of training samples $N = \sum_\nu N_\nu$ and hyper-parameters

$$\alpha_\nu = 1 + N_\nu \ . \tag{6.20}$$

Finally, we can express the probabilities $p(L_\nu|\Lambda_\tau, T)$ just by normalizing the hyper-parameters as

$$p(L_\nu|\Lambda_\tau, T) = \frac{\alpha_\nu}{\sum_\nu \alpha_\nu} \ . \tag{6.21}$$

We see, indexing heterogenous collections of images is nothing but an extension of the traditional model for new datasets. In (Fig. 6.7), we illustrate the indexing using Landsat TM and Ikonos data from Mozambique. Each dataset was separately clustered and indexed and cover-type labels were aggregated at semantic level.

6.4 Learning Ontological Categories

The user group of an information retrieval system usually consists of people with various degrees of expertise and the structure of the group depends on the specificity of the people's interests. Apart from this structure, the group can taxonomically be partitioned into areas of interest. Within each area of interest, the members can be further grouped according to the level of expertise or degree of knowledge in that specific area. However, there is no clear definition for the separation into areas of interest since interests usually overlap to a certain extent. Consequently, the concepts of an expert in his area of interest are more associated to users whose interests are similar to the expert's ones.

Before we outline how users of the image information mining system utilize and share domain categories to learn and represent semantic categories, we will first point out the prerequisites of understanding certain contexts and context sharing from other domains.

6.4.1 Prerequisites

The user group of the mining system consists of people from very heterogenous working fields, such as agriculture, urbanism, sensor design and image analysis, for

instance. They all have different background knowledge and intentions. What they have in common, however, is the interest in earth observation domains, products and sensor data. When a user enters the system, he first has to register and select a certain sensor and combination of feature models for further data exploratory purposes. Sensors may include both optical, e.g. panchromatic, multi- or hyperspectral and radar with different polarizations. The user's decision for a certain signal model also involves resolution aspects since the representation of various image structures is scale-dependent. After choosing a sensor (or several sensors for multi-mission image exploration) and corresponding signal models, the user can select a certain image, train semantic labels of his interest and search the entire archive for relevant images. The classes of scenes or objects selected by the user depend on the domain ontology [1] he belongs to. Since each user belongs to a certain ontology domain, misunderstandings in the semantic space can occur. For instance, a climatologic expert could have a different association of water in comparison to an oceanography expert or an urban architect may have a different vision of city than an ecologist. In order to have an effective communication between different users without running the risk of having misinterpretations, the users have to share a common ontology. With this ontology-based approach, it is possible to achieve a more effective search and data exploration.

Similar ontological approaches are applied in e-commerce as amazon.com and grouplens. Amazon.com aims at analyzing the customer's behaviour and comparing it with other customers in order to recommend the most evident products. Grouplens (Resnick et al. 1994) helps readers of netnews to find suitable articles by making choices based on the recommendations of other users.

6.4.2 Learning and Representation of Semantic Categories

In the image information mining system we implemented techniques from statistical analysis and machine learning to provide users with innovative tools to "explore and explain" the image data in the archive: a user can select several feature models that are most suitable for him, train a specific cover-type of his interest by giving positive and negative samples and insert the cover-type semantic in the DBMS in a free-form text manner. With standard computer vision like methods the resulting associated semantic content is difficult to be obtained.

Thus, the idea is to build intelligent visual interfaces which explain image data and high-level user concepts in order to achieve a compromise for a common understanding. This results in the difficulty of knowledge sharing, involving the related problems of ontology, i.e. the specification of a conceptualisation. In this context,

[1] The term 'ontology' has a long history in philosophy and meets with increasing interest in the field of artificial intelligence. According to T. Gruber, ontology is defined as a specification of a conceptualization (Gruber 1993). Gruber further specifies ontology as a description of the existing concepts and relationships for or against a community agent.

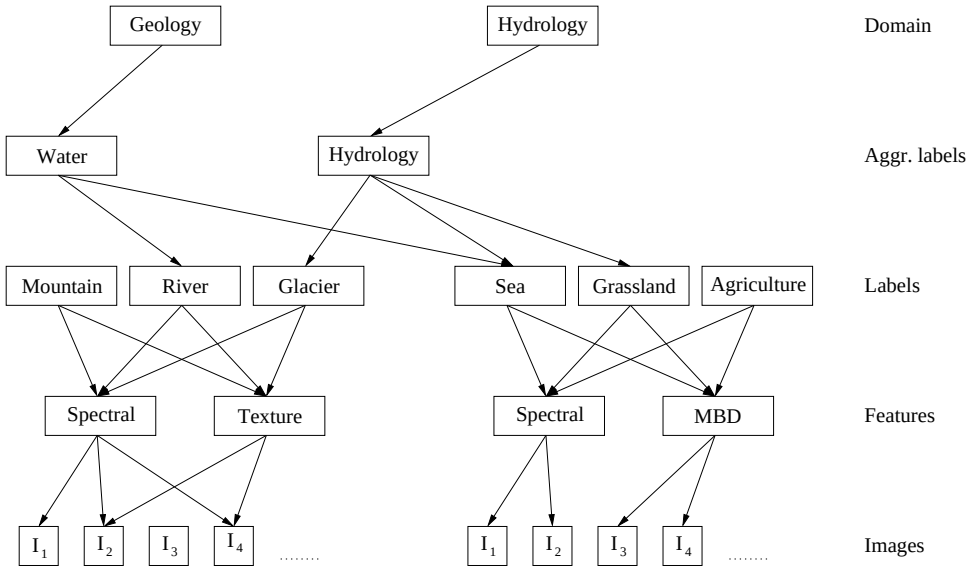


Figure 6.8: Domains (and sub-domains) are linked to individual and aggregated semantic labels and further to image features and image data. The two domains share the semantic label 'sea', however, they consider it from different views (domains). As a result, conceptual and terminological confusion at the semantic level may occur.

we understand an ontology as a description of the categories and associations that could be established between the components of the set of hierarchical representation of information and for a given user conjecture, as presented in (Fig. 6.8). Semantic conflicts can arise between information communities, when methodology or ontology is not shared. Resolving the semantic antagonisms is equivalent to harmonizing different scientific models, and involves the establishment of equivalencies between feature types and between feature instances and equating model types and instances. Therefore, the notion of context must be enhanced.

6.5 Conclusions

In this chapter, we have discussed the following items:

- We outlined the limitations of the basic hierarchical scheme of image content abstraction and the motivation to add a new level. With the same stochastic model with which we linked user-specific cover-types to the unsupervised content-index, we can define higher-level semantic concepts by grouping existing labels. Whereas the signal-oriented supervised semantic labeling of image content only provides the query for images that are assigned to the content-

index, the new level of information representation allows to query archives across datasets and sensors.

- Users of the mining system have different domain-specific background knowledge and therefore can taxonomically be distributed into several ontology domains. At the level of defining semantic categories, problems may arise if users do not share a certain methodology or ontology. Avoiding these semantic conflicts requires the incorporation of semiotic aspects.

Part III

System Evaluation

7

System Evaluation Methods

In Chap. 5, we presented basic concepts and functions implemented in the information mining system and in Chap. 6 we described an enhanced framework for content-based image retrieval at semantic level. Why did we give such a detailed system description? From the author's point of view, it is essential for the further understanding and assessment of this work to get a survey of implemented modules and mining functions. We included not only individual tools in the presentation but also their interactions and the information flow during system operation. Altogether, we aimed at demonstrating the system complexity and the imperative for an overall evaluation.

For the assessment of the effectiveness of a content-based image retrieval system, there are several criteria and measurements. Since there is almost no standardization in this domain, evaluation measures can be arbitrarily selected. A current method is to analyze the queried image set in terms of relevant and non-relevant images. Then, the retrieval performance is reflected by the number of relevant images in the search results (precision) and the number of returned relevant images from all relevant images in the archive (recall). However, precision/recall only indicates the retrieval capability of a system but is not sufficient to present the overall performance.

In addition to the problem of the selection of suitable evaluation measurements, we are further faced with rather complex remote sensing data with a much higher diversity of structures and objects than in multimedia applications. A mining system dealing with this kind of images requires specific functions to exploit the information content and needs specially-designed and application-dependent tools for its evaluation. To perform an overall verification of our system, we developed an algorithmic protocol that will be presented in this chapter. The protocol consists of methods for the objective technological evaluation of each system module, takes into account the subjective evaluation component as well, and measures the information flow from the archive to the user.

We begin this chapter with an analysis of the information content of extracted spectral and textural image parameters in Sec. 7.1. Sec. 7.2 deals with the representation of significant image structures in the individual feature spaces by clusters. To identify the information content of clusters, we determine their discrimination

performance and the ability to represent details. After analyzing the off-line data processing chain, the extraction of primitive image features and their unsupervised grouping, we focus on the interactive learning module in Sec. 7.3. The quality of interactive learning — the supervised classification of the entire image archive — is represented by the stochastic link between subjective semantic cover-types and objective signal classes, the classification accuracy/selectivity and the separation of semantic labels against each other. Based on the stochastic link between semantic label and content-index, the whole archive can be searched in a probabilistic way for relevant images. In Sec. 7.4, we measure the quality of the probabilistic search function by applying standard measurements like target/misclassified images and precision/recall. The system operation complexity is analyzed in Sec. 7.6 and human-machine interactions in Sec. 7.7. After having evaluated the technical quality of individual system modules and the communication with the user, we measure the information flow from the archive to the user by applying different levels of information abstraction as outlined in Sec. 7.8. We conclude this chapter in Sec. 7.9 with a summary of other evaluation criteria that have not been explicitly pointed out but are worth mentioning.

7.1 Information Content of Primitive Image Features

A comprehensive evaluation requires the analysis of all system modules and their connections. According to the described scheme of hierarchical Bayesian image content modeling we begin the system verification with an analysis of the information content of image data and extracted primitive features (Daschiel and Datcu 2002b). To describe the content of remote sensing images in the archive, we use spectral and textural attributes. The textural parameters are calculated at different scales since the appearance of several structures is scale-dependent. In the following verification, spectral and textural attributes are considered.

7.1.1 Spectral Features

If we are interested in the information content of spectral image parameters, we have to measure the degradation of the data D by noise. We will only analyze optical images which are considered to be distorted by zero-mean white Gaussian noise with unknown variance σ_η^2. Multiplicative noise will not be taken into account. The estimated noise variance indicates the reliability in the given data.

To determine σ_η^2, we apply several of the methods for noise variance estimation proposed by Olsen (Olsen 1993b). From the described algorithms we use the 'Average' and the 'Median' ones because they delivered the best results in Olsen's experiments (Olsen 1993a). In both methods, a filtered image is subtracted from

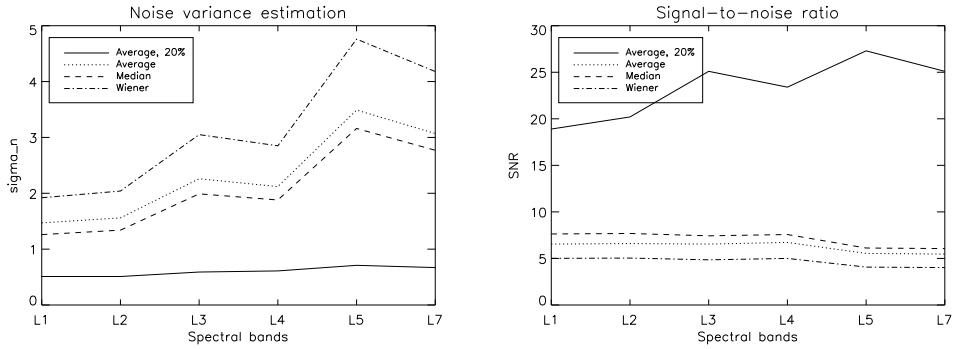


Figure 7.1: Standard deviation σ_η and SNR for Landsat TM spectral bands in the visible spectrum. Differences between 'Average' and 'Median' are due to a restriction to homogeneous regions in the 'Average' method; 'Average, 20%' means that 20% of the image data contribute to σ_η and SNR. The other filters include all data samples.

the given non-filtered one to derive a measure of noise at each image element. We apply a 3×3 filter kernel for both methods as recommended by Olsen. Whereas in the 'Median' method all data points are considered, they can be specified in the 'Average' method. With a selection of homogeneous regions using the gradient, edges and even small-scale image structures can be excluded from contributing to the noise measure. Olsen demonstrated that the 'Average' method achieved the best estimates for various datasets. Since all samples contribute in the 'Median' approach, the obtained variance is overestimated in comparison to the other method. In addition to Olsen's methods, we apply an adaptive Wiener filter for denoising that is sensitive to local image variances. If the estimated noise variance is large for data samples, the smoothing is low and vice versa. Again, the image noise variance is determined by subtracting the filtered image from the original one. In (Fig. 7.1), we depict the yielded noise standard deviation σ_η and the corresponding signal-to-noise ratio (SNR) for the spectral bands of a Landsat TM image.

7.1.2 Textural Features

The content of images in the archive is not only described by spectral attributes, but also by estimated parameters based on Gibbs-Markov random field texture models (Datcu et al. 1998). In general, such models are given as parametric data models via the likelihood $p(D|\boldsymbol{\theta}, M)$ as outlined in Chap. 5. Thereby, different structures in the data are characterized by different values of the elements of $\boldsymbol{\theta}$. We realize the extraction of spatial information by calculating the maximum a posteriori estimate of the model parameter vector. To ensure a fast and robust estimation of the parameters, we apply the conditional least-squares (CLS) estimator (Lele and

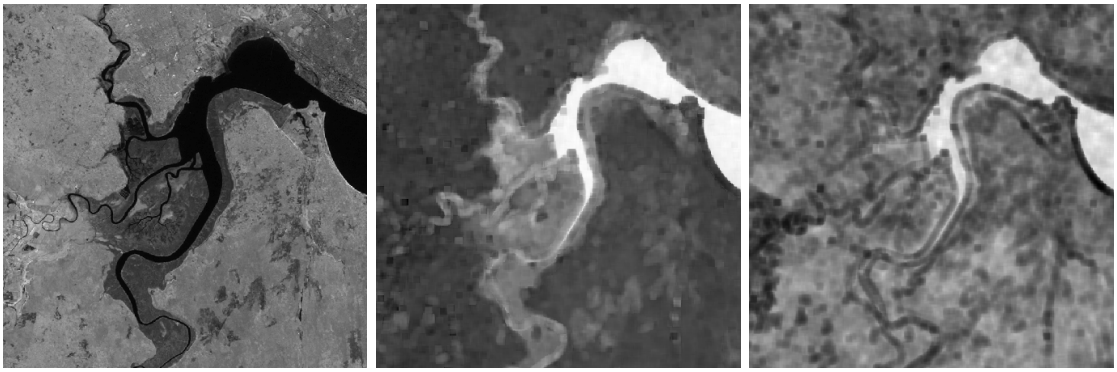


Figure 7.2: Assessing the estimation performance of Gibbs random field texture model parameters by using the Cramér-Rao bound. Images from left to right. Landsat TM with its $4th$ band from where texture parameters $\hat{\boldsymbol{\theta}}$ are estimated, the norm $|\hat{\boldsymbol{\theta}}|$ of the estimated parameters and the Cramér-Rao bound $\sigma^2_{|\hat{\boldsymbol{\theta}}|}$. For visualization, we used a logarithmic scaling for norm and CRB. For smooth regions with almost no texture, e.g. water, the CRB indicates large values, whereas with regard to areas with significant structures and patterns the CRB is small.

Ord 1986) for the auto-binomial model from the Gibbs family. Then, the parameter estimation is reduced to robust linear regression (Schröder et al. 1998).

After having estimated the model parameter vector $\boldsymbol{\theta}$, we focus on the quality of the estimation process. A common way of determining the accuracy of an estimated parameter is to calculate its variance and a possibly occurring bias. Although there exist several bounds to limit the variance of an estimated parameter, we confine ourselves to the Cramér-Rao bound (van Trees 1968). For a scalar parameter, the CRB delivers the lower bound on the mean-square error in the estimate. Since the extracted spatial information of the image is implicitly contained in vector $\boldsymbol{\theta}$, a bound on the variance of each element of $\boldsymbol{\theta}$ must be placed to get a measurement for the accuracy of the estimated parameter vector. If we consider an unbiased estimate of $\boldsymbol{\theta}$,

$$\sigma_{\hat{\boldsymbol{\theta}}}{}^2 \geq \boldsymbol{I}(\boldsymbol{\theta})^{-1} \tag{7.1}$$

is the bounded variance of the estimated parameter vector $\boldsymbol{\theta}$. The square matrix $\boldsymbol{I}(\boldsymbol{\theta})$ denotes the Fisher information matrix which is given by

$$[\boldsymbol{I}(\boldsymbol{\theta})]_{ij} = -E\left\{\frac{\partial^2 \log p(D|\boldsymbol{\theta}, M)}{\partial\theta_i\,\partial\theta_j}\right\} \quad . \tag{7.2}$$

$E\{\cdot\}$ denotes the expectation and θ_i is the ith component of $\boldsymbol{\theta}$. Eq. 7.1 and 7.2 show that the CRB only depends on the likelihood $p(D|\boldsymbol{\theta}, M)$. Since the total probability cannot be determined for Gibbs random fields, the theoretical lower limit usually

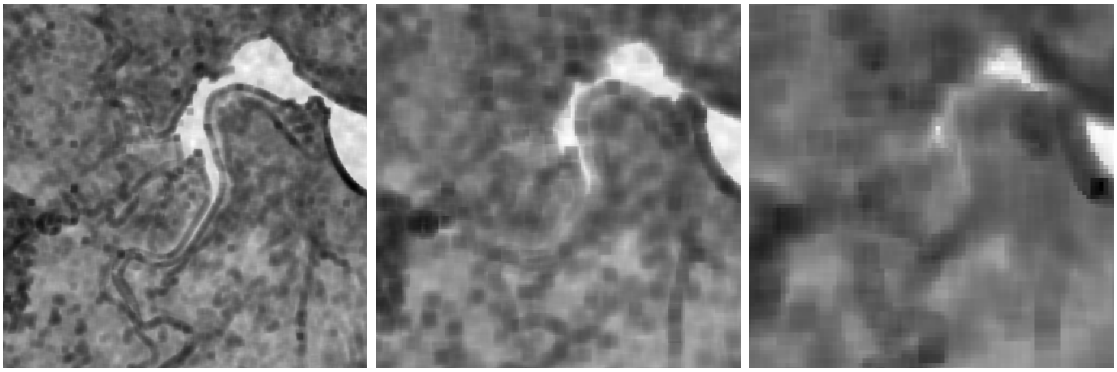


Figure 7.3: Estimation performance of Gibbs-Markov random fields model parameters at different scales by using the Cramér-Rao bound. Images from left to right. Cramér-Rao bound of norm $|\hat{\boldsymbol{\theta}}|$ at 30m, 60m and 120m. Images are logarithmically scaled for a higher visualization quality.

cannot be computed. Only for the auto-binomial model in combination with the conditional least-squares estimator, the Cramér-Rao bound can be calculated. This bound results in the variance of the estimated parameter vector given the data D. To obtain a scalar measurement for the CRB (Papoulis 1984), we use the norm $|\hat{\boldsymbol{\theta}}|$ as the strength of the texture instead of the parameter vector $\hat{\boldsymbol{\theta}}$ as exemplified in (Fig. 7.2).

7.1.3 Features at Multiple Scales

In order to obtain a quasi-complete description of the scene, we perform a multiscale approach as outlined in Sec. 5.1. Therefore, we scale the image in a dyadic way and apply a Gibbs estimation kernel of a limited size to the data at each scale. For the scale-dependent analysis of the accuracy of the estimation process, we compute the Cramér-Rao bound for the obtained parameters at each scale. The Cramér-Rao bound of the norm $|\hat{\boldsymbol{\theta}}|$ reflects the capability of a certain scale to describe homogeneous texture with respect to the applied model order. To exemplify the CRB of estimated parameters at different scales, we analyze the image from (Fig. 7.2) and show the obtained results in (Fig. 7.3). In addition to the 2D visualization of the scale-dependent Cramér-Rao bound in (Fig. 7.3), we also computed the normalized histogram (pdf) for norm and CRB as depicted in (Fig. 7.4).

7.2 Unsupervised Clustering

The first steps in the I^{2}M data ingestion chain are the extraction of primitive image features and their reduction ensured by an unsupervised classification. Primitive

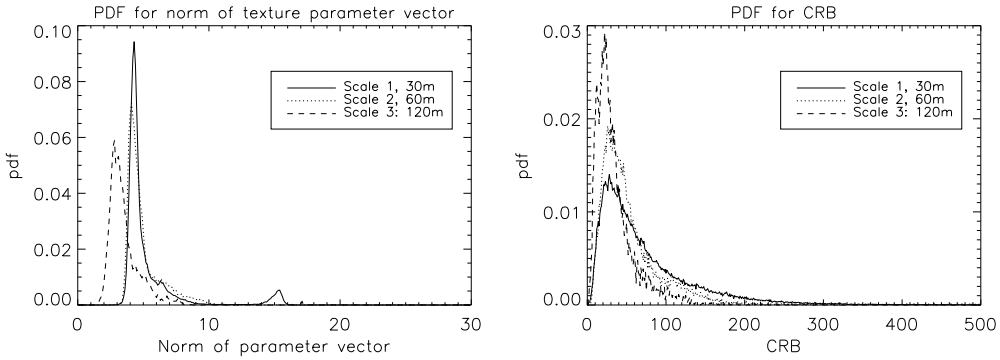


Figure 7.4: Normalized histogram of norm (left) and CRB (right). The image content at scale of 30m shows the most significant texture reflected by the highest values for the norm $\hat{\boldsymbol{\theta}}$. However, the difference between first and second scale (30m and 60m) is rather small. As far as accuracy of the estimated model parameters is concerned, texture at scale of 120m depicts the best results.

visual parameters must be compressed and reduced since the feature extraction produces large volumes of data that cannot be managed in practice. Clustering, which is similar to vector quantization (McLean 1993), reduces the accuracy of the system, but justifies its practical use due to a large data reduction.

Each image element is located in a multidimensional-dimensional space at a certain position specified by the values of the contributing, pre-extracted primitive features. In this space, the 'pixels' tend to group themselves into specific regions. Then, the clustering process substitutes the 'clouds' of primitive features by parametric models $p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_j \in \omega_i)$ of their groups and therefore makes a more compact description of feature space points possible. In our system, primitive features are compressed by the use of a dyadic k-means algorithm.

In order to determine the information content of signal classes (clusters) ω_i, we have to identify the results of unsupervised clustering in a multidimensional space. An often applied criterion for cluster analysis is the separability of clusters with methods like scatter matrices, for instance. In order to circumvent the 'curse of dimensionality' for cluster analysis, we project multidimensional feature clusters in a 2D image space. Thus, cluster analysis is reduced to studying unsupervised classification maps.

7.2.1 Cluster Analysis

Popular techniques for cluster discrimination are scatter matrices, divergence, Bayes' probability of error and non-parametric feature space density estimation that we will now proceed to describe.

Scatter matrices

With scatter matrices, we can analyze the compactness of clusters as well as the isolation between them. Therefore, scatter information is an important measurement of separability. Assuming that a d-dimensional feature space R^d consists of n data points $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ and all points are assigned to one of r clusters $\omega_1, \omega_2, \ldots, \omega_r$, a scatter matrix $\boldsymbol{S}_i$ for the ith cluster ω_i is given by

$$\boldsymbol{S}_i = \sum_{\boldsymbol{\theta}_j \in \omega_i} (\boldsymbol{\theta}_j - \boldsymbol{\mu}_i)(\boldsymbol{\theta}_j - \boldsymbol{\mu}_i)^t \ , \tag{7.3}$$

where $\boldsymbol{\mu}_i$ denotes the mean vector for the ith cluster. If we further define the within-cluster scatter matrix

$$\boldsymbol{S}_w = \sum_{i=1}^{r} \boldsymbol{S}_i \ , \tag{7.4}$$

the between-scatter matrix

$$\boldsymbol{S}_b = \sum_{i=1}^{r} n_i (\boldsymbol{\mu} - \boldsymbol{\mu}_i)(\boldsymbol{\mu} - \boldsymbol{\mu}_i)^t \tag{7.5}$$

and the total scatter matrix

$$\boldsymbol{S}_T = \boldsymbol{S}_W + \boldsymbol{S}_B \tag{7.6}$$

with n_i number of samples in cluster ω_i and $\boldsymbol{\mu}$ the total mean vector, we can derive scalar measurements from scatter matrices as (Duda et al. 2001)

$$\xi_1 = \frac{|\boldsymbol{S}_W|}{|\boldsymbol{S}_T|} \tag{7.7}$$

and (Aksoy and Haralick 2001)

$$\xi_2 = \log |\boldsymbol{S}_W^{-1}(\boldsymbol{S}_W + \boldsymbol{S}_B)| \tag{7.8}$$

with $|\cdot|$ denoting the determinant of a matrix. These quantities deliver the quality of the applied clustering method as well as the information content of clusters.

Divergence

Another measurement for the discrimination effectiveness of clusters ω_i is divergence (Kailath 1967). To introduce divergence for cluster analysis, we start with the problem of how to classify a single point $\boldsymbol{\theta}_j$ in a multidimensional feature space to one of two groups ω_k or ω_l. Then, the divergence between clusters (classes) ω_k and ω_l is given as

$$D_{kl} = \sum_{\boldsymbol{\theta}_j} [p(\boldsymbol{\theta}_j|\omega_k) - p(\boldsymbol{\theta}_j|\omega_l)] \log \frac{p(\boldsymbol{\theta}_j|\omega_k)}{p(\boldsymbol{\theta}_j|\omega_l)} \tag{7.9}$$

Feature	Scale	Dim.	Normal. method	ξ_1 $(\cdot 10^{-5})$	ξ_2	$\langle D\rangle_{div}$	$\langle D\rangle_{div'}$	$\langle D\rangle_{Bhat}$
Spectral	30m	6	Gaussian	4.84	9.94	1596	1.913	0.951
Spectral	30m	6	Linear	0.88	11.64	942	1.932	0.961
GRF	30m	4	Gaussian	21.3	8.45	531	1.910	0.947
GRF	60m	4	Gaussian	13.80	8.89	119	1.917	0.953
GRF	120m	4	Gaussian	13.72	8.89	162	1.930	0.961
EMBD	60m	1	Gaussian	2.23	6.10	13858	1.967	1.192
EMBD	120m	1	Gaussian	2.45	6.01	14988	1.967	1.209
GMRF	60m	1	Gaussian	2.11	6.16	83931	1.960	1.448
GMRF	120m	1	Gaussian	0.94	6.97	24481	1.956	1.063

Table 7.1: Clustering performance for the Mozambique multi-mission datasets and features (Tab. 5.1).

and finally the average discrimination between the feature clusters

$$\langle D\rangle = \sum_{k=1}^{r}\sum_{l=1}^{r} p(\omega_k)\, p(\omega_l)\, D_{kl} \tag{7.10}$$

with $\langle\cdot\rangle$ indicating the average, r the total number of clusters and $p(\omega_i)$ the prior probability of cluster ω_i.

In general, the divergence $\langle D\rangle$ of Eq. 7.10 cannot be easily calculated, however, in the case of clusters with a Gaussian distribution $p(\boldsymbol{\theta}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, the divergence D_{kl} between two clusters ω_k and ω_l can be expressed as (Therrien 1989)

$$\begin{aligned} D_{kl} \;=\; & \tfrac{1}{2} trace\left\{\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_l + \boldsymbol{\Sigma}_l^{-1}\boldsymbol{\Sigma}_k - 2\boldsymbol{I}\right\} + \\ & \tfrac{1}{2}\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\right)^t \left(\boldsymbol{\Sigma}_k^{-1} + \boldsymbol{\Sigma}_l^{-1}\right)\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\right) \quad . \end{aligned} \tag{7.11}$$

The computation of divergence in Eq. 7.11 caused difficulties since small variations in the mean vector difference resulted in large divergence changes. Thus, the transformed divergence was proposed (Swain and Davis 1978)

$$D'_{kl} = 2\left(1 - \mathrm{e}^{-D_{kl}/8}\right) \quad . \tag{7.12}$$

Through the exponential nature, the transformed divergence D'_{kl} has a more saturating behaviour for growing class discrimination.

Bhattacharyya distance

In the same way as we defined the divergence between a pair of probability distributions $p(\boldsymbol{\theta}_j|\omega_k)$ and $p(\boldsymbol{\theta}_j|\omega_l)$, we obtain the Jeffries-Matusita distance as (Duda et al. 2001)

$$J_{kl} = \sum_{\boldsymbol{\theta}_j}\left(\sqrt{p(\boldsymbol{\theta}_j|\omega_k)} - \sqrt{p(\boldsymbol{\theta}_j|\omega_l)}\right)^2 \quad . \tag{7.13}$$

In the case of Gaussian distributed clusters $p(\boldsymbol{\theta}_j|\omega_i)$, Eq. 7.13 can be expressed as

$$J_{kl} = 2\left(1 - \mathrm{e}^{-B}\right) \tag{7.14}$$

with the Bhattacharyya distance

$$B = \frac{1}{8}\left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\right)^t \left(\frac{\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_l}{2}\right)^{-1} \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\right) + \frac{1}{2}\log\left(\frac{\left|\left(\boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_l\right)/2\right|}{\sqrt{\left|\boldsymbol{\Sigma}_k\right|\left|\boldsymbol{\Sigma}_l\right|}}\right) \ . \tag{7.15}$$

In the same way as the average divergence of pairwise clusters is defined, we can write

$$\langle D \rangle = \sum_{k=1}^{r} \sum_{l=1}^{r} p(\omega_k)\, p(\omega_l)\, J_{kl} \tag{7.16}$$

for the average Bhattacharyya distance (Tab.7.1).

Bayes' probability of error

Another measure to express the quality of the obtained clusters to represent the image content is given by the so-called probability of error (Duda et al. 2001). The basic idea behind this quantity is to estimate the probability of each single point $\boldsymbol{\theta}_j$ in the feature space and to analyze if it is really assigned to a certain cluster. After the grouping process, all points belong to one of the r clusters $\omega_1, \omega_2, \ldots, \omega_r$ and we obtain the conditional densities $p(\boldsymbol{\theta}_j|\omega_i)$ for each point given a certain class ω_i. Using Bayes' formula

$$p(\omega_i|\boldsymbol{\theta}_j) = \frac{p(\boldsymbol{\theta}_j|\omega_i)p(\omega_i)}{p(\boldsymbol{\theta}_j)} \ , \tag{7.17}$$

we can infer the posterior probability $p(\omega_i|\boldsymbol{\theta})$ for a class ω_i given a data point $\boldsymbol{\theta}_j$ by combining the prior probabilities $p(\omega_i)$ with the likelihood $p(\boldsymbol{\theta}_j|\omega_i)$ of ω_i with respect to $\boldsymbol{\theta}_j$. The term $p(\boldsymbol{\theta}_j)$ is called the evidence and serves as a normalization constant for the posterior probability. After calculating the posterior probability for each point, we can determine Bayes' probability of error as (Therrien 1989)

$$p(error|\boldsymbol{\theta}_j) = 1 - \max\{p(\omega_i|\boldsymbol{\theta}_j)\} \ . \tag{7.18}$$

This way, we derive the influence of all clusters on each point, and consequently, the relationship of a point to its true cluster. In (Fig. 7.5), we illustrate Bayes' probability of error for a classified image using spectral image parameters.

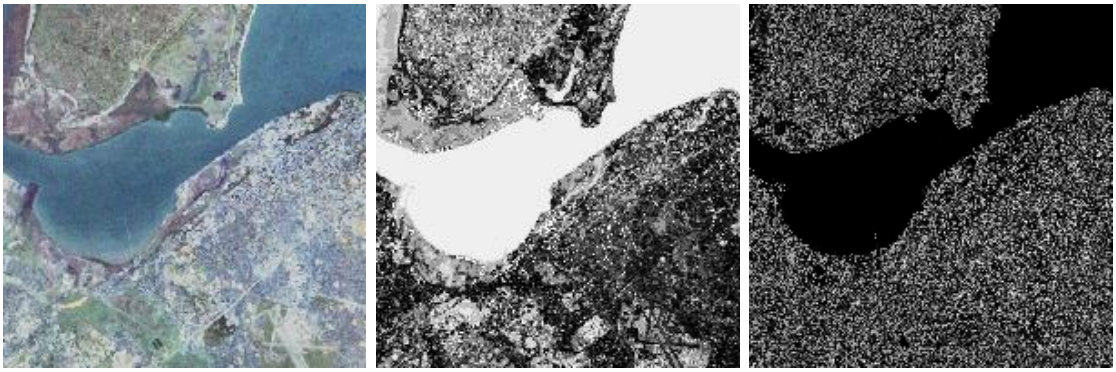


Figure 7.5: Visualization of Bayes' probability of error for the unsupervised classification of spectral features. Images from left to right: quick look of a Landsat TM image, its unsupervised classification and the computed probability of error $p(error|\boldsymbol{\theta}_j)$. The latter is encoded from white (high error) to black (low error). Image points that belong to strong and well-separated clusters, e.g. sea, indicate a low probability of error.

Non-parametric density estimation

How much information do clusters contain? Are they located in the feature space by accident and are their forms correlated with the distribution of the dataset? To answer these questions, we have to analyze the points' distribution in the multi-dimensional feature space. This analysis is equivalent to non-parametric density estimation. From the two main approaches for this task, the Parzen-window density estimation (Parzen 1962) and the k_n-nearest-neighbour estimation (Patrick and Fischer 1970), we apply the first one due to its lower computational complexity. Using one of Parzen's window functions $\varphi(\cdot)$, the density estimate can be expressed as

$$p_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{\boldsymbol{\theta}-\boldsymbol{\theta}_j}{h_n}\right) \quad , \tag{7.19}$$

where n denotes the number of samples and V_n a multidimensional hypercube with edge length h_n centered at $\boldsymbol{\theta}$. Achieving a good density estimate means finding a good extension of the hypercube for the applied window function. In the most simple case, we can use a unit hypercube centered at the origin. This corresponds to a gridding of the feature space in regular cells. In (Fig. 7.6), we show the result of the estimated density for textural and spectral feature spaces.

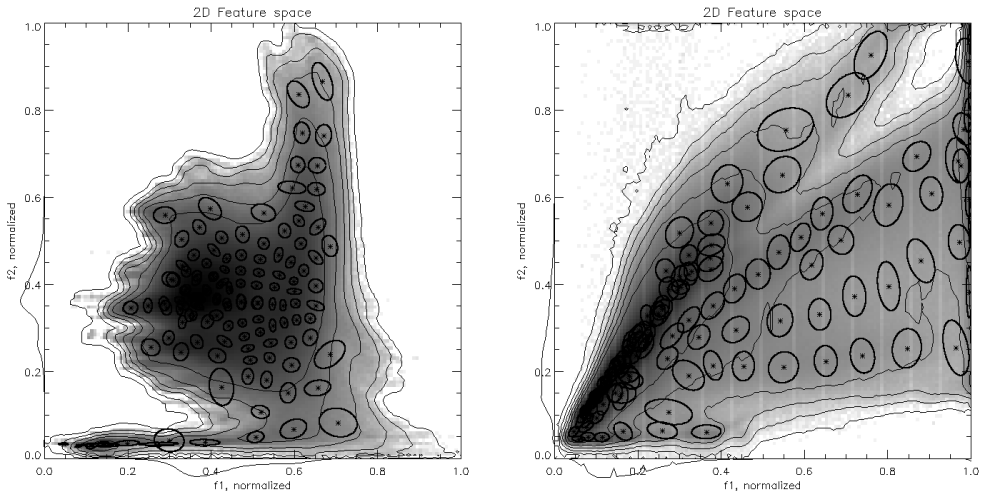


Figure 7.6: Projection of a 4-dimensional texture feature space (left) and a 6-dimensional spectral feature space (right). Each feature space consists of about 100 Mio. data points. The plots show the estimated feature space density, marginal distributions, cluster centers and their shapes. For clearness reasons, feature space density and contour lines are logarithmically scaled.

7.2.2 Accuracy of Unsupervised Content-index

We pointed out methods to measure the performance of unsupervised clusters for different feature models. As opposed to analyzing the clusters in a multidimensional space, we project the feature clusters in the 2D image space. Thus, discovering the clustered groups is reduced to an analysis of the classification maps. We follow the approach to determine classification accuracy with confusion or error matrices, which compare the outcome of a supervised or unsupervised classification with 'ground-truth' information. As it is rather difficult to have access to such data, we compare the produced unsupervised classification maps with reference classification maps. In the next section, we outline error matrices in more detail. Then, we use them to determine the accuracy of interactively trained cover-type labels.

In order to obtain reference classification maps computed for spectral features, we assumed to have noise-free image data. Noise was removed by filtering the data as applied in Sec. 7.1. To analyze the influence of different levels of noise distortion, we added successive zero-mean white Gaussian noise to the noiseless data. Then, the unsupervised classification result of the noisy data was compared with the reference classification. We calculated the overall and the average accuracy to obtain the reliability in yielded classes. Furthermore, the κ-coefficient as a combination of the overall and average classification accuracy was computed. Finally, we compared the obtained results with Olsen's noise variance estimates to derive an approximated measurement for the loss of information due to existing noise. The results of this experiment are illustrated in (Fig. 7.7). Comparing the obtained classification ac-

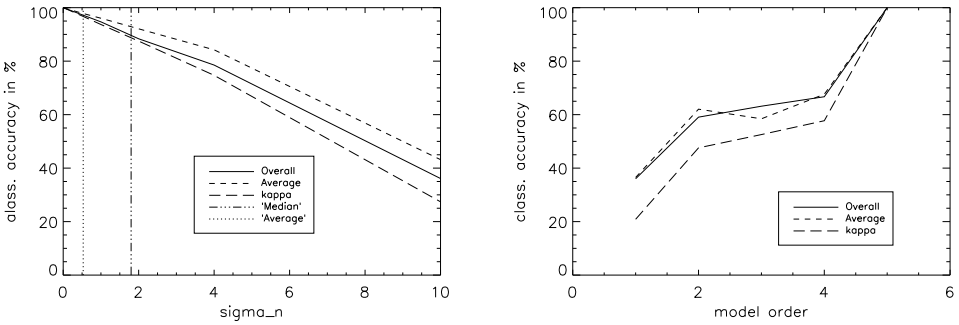


Figure 7.7: Accuracy of classification results generated by spectral (left) and textural (right) features using measurements from confusion matrices. Spectral signal classes are analyzed according to the degree of noise in the Landsat spectral bands and signal classes derived from texture model parameters according to the applied model complexity (order).

curacies with the estimated noise variance σ_η^2, we find a loss of information due to noise of about 5% to 10%.

We repeated the experiment and classified textural parameters. To test the influence of estimated parameters on the unsupervised clustering, we compared the classifications of textural features based on different orders (complexity) of the autobinomial texture model. The outcome for model order 5 is considered as the reference classification and results based on lower model orders are compared with the reference data. In (Fig. 7.7), we depict the experimental results.

7.3 Interactive Learning

In our evaluation procedure, the verification of the interactive training is focused on the quality assessment of user-defined cover-type labels L_ν. Each semantic label is the outcome of probabilistic calculations based on signal classes ω_i. Therefore, the accuracy of a label indicates the objective quality of the system to detect the cover-type in the whole database.

Since each semantic label L_ν can be seen as a supervised classification of the entire image archive in label L_ν and 'non'- label $\neg L_\nu$, we make use of accuracy assessment strategies to analyze (1) the quality of a stochastic link between a subjective cover-type label L_ν and objective signal classes ω_i, (2) the classification accuracy of a single label L_ν and (3) the separation between all labels $\{L_\nu\}$ [1].

[1] If we speak of 'all' labels $\{L_\nu\}$ in this context, we mean all defined labels for a special combination of signal classes, e.g. spectral and texture at a certain scale.

7.3.1 Quality of Stochastic Link

The question of the quality of the stochastic link between subjective semantic labels and objective signal classes can be directly answered by using information-theoretic quantities. In (KULLBACK 1997), the divergence between two complete sets of probabilities $\mathcal{L}_\nu = \{p(\omega_1|L_\nu), \ldots, p(\omega_r|L_\nu)\}$ and $\neg\mathcal{L}_\nu = \{p(\omega_1|\neg L_\nu), \ldots, p(\omega_r|\neg L_\nu)\}$ is defined as

$$D(\mathcal{L}_\nu, \neg\mathcal{L}_\nu) = \sum_{i=1}^{r} [p(\omega_i|L_\nu) - p(\omega_i|\neg L_\nu)] \log \frac{p(\omega_i|L_\nu)}{p(\omega_i|\neg L_\nu)} \ , \tag{7.20}$$

which can be seen as the distance between the two probability distributions $\mathcal{L}_\nu$ and $\neg\mathcal{L}_\nu$. The divergence $D(\mathcal{L}_\nu, \neg\mathcal{L}_\nu)$ can be calculated either for a combination of signal models or separately for each model. Additionally it supports the system operator during the interactive training and continuously gives him relevance feedback about the performed training (Fig. 5.7). The usefulness of the divergence is shown in the selection of strong signal models, the removal of low ones and in finding similarities between semantic cover-types defined by different users.

7.3.2 Classification Accuracy and Selectivity

As mentioned in the beginning of this section, each single label can be considered as a partition of the dataset in labels L_ν and $\neg L_\nu$. To determine the classification accuracy and quality of a user-defined cover-type, we use a standard method in remote sensing: error or confusion matrix (RICHARDS and JIA 1999). The basic principle of such a matrix is to compare an obtained classification with ground-truth data or a reference classification.

Since ground-truth information is usually not available or can only be gained to a limited extent, we will not use it. Another problem in a ground-truth based classification assessment is the fact that ground-truth classification maps are normally generated by different datasets or are acquired at different times. Different data acquisitions, in particular, make it difficult to derive objective measurements from the confusion matrix, e.g. a semantic label 'mountain' defined in a Landsat image can change completely if mountains are snow-covered, for example. Finally, these reference datasets are normally available for a fixed number of classes and are quite inflexible, especially for small classes.

To avoid a long-term generation of ground-truth maps, we compute a maximum-likelihood (ML) reference classification for a selected test dataset. The ML classification can be considered as a classification of superior accuracy since it is based on large training regions in comparison to the trained cover-type labels L_ν that are defined with just a few samples. One of the advantages of this reference classification is that both maps are based on the same dataset. Besides, the data volume can be partitioned into as many classes as semantic labels and the computation can be

Figure 7.8: Comparison of a reference classification with a defined semantic label. Images from left to right: Landsat TM image from the Nepal dataset, its ML reference classification of 'Dense-mixed forest' with classes B_ν (white) and $\neg B_\nu$ (black), and the supervised Bayesian classification with cover-type label L_ν (white) and $\neg L_\nu$ (black). Only small differences are obvious between B_ν and L_ν, although the latter is defined with only a few training samples.

repeated due to a short processing time. Weakly defined classes can be evaded by using a threshold in the ML classification algorithm.

For semantic labels we determine the classification accuracy by using the reference ML classification as previously described. We compare a label L_ν with the corresponding class B_ν in the reference classification and $\neg L_\nu$ with $\neg B_\nu$, that is, a combination of all other classes except B_ν (Fig. 7.8). Thus, for each semantic cover-type we obtain a 2×2 confusion matrix with elements $x_{\nu\mu}$. From $x_{\nu\mu}$, we calculate the overall proportion of area correctly classified

$$Q_o = \frac{\sum_\nu x_{\nu\nu}}{N} \tag{7.21}$$

with the total number of N observations and the average

$$C_\nu = \frac{U_\nu + P_\nu}{2} \tag{7.22}$$

of the user's accuracy U_ν and the producer's accuracy P_ν (Congalton 1991). In (Tab. 7.2), we summarize the calculated classification accuracy measurements for several labels. In this experiment, we only used the information from spectral signals for label training to achieve an objective comparison between the spectral-based reference classification and semantic labels.

We should not forget to mention the limitations and drawbacks of the proposed method. First, in the supervised reference classification only the main classes can be well separated. It is almost impossible to reach a classification of small structures

Label name	Overall Acc. P_o	Avg. Acc. C_ν
river	97.8	54.5
riverbank	98.4	65.8
clouds	94.1	50.4
sisau forest	93.4	53.5
dense-mixed forest	99.0	76.4
soil	59.8	52.5
grassland	82.5	39.4
Average:	89.3	56.1

Table 7.2: Classification accuracy of spectral-based cover-type labels in %. All semantic labels L_ν were defined in the system evaluation test week, this time, to analyze the classification performance of the system.

with only spectral information. For an entire evaluation of the classification potential of the I^{2}M system, further experiments have to be done. Another drawback is that the results of (Tab. 7.2) do not only represent the classification quality of the labels, but also the inaccuracies of the performed ML classification. In order to avoid the generation of a reference classification, we present a method that is completely independent of reference data.

In order to avoid the disadvantages of ground-truth based classification accuracy techniques, we perform an alternative classification evaluation method. Instead of using other reference classification maps, we stay on the user's training samples. Each time a user gives a new positive or negative training example to the system, the probabilities $p(\omega_i|L_\nu)$ of the stochastic link are updated, and, consequently, so are the posterior probabilities $p(L_\nu|\omega_i)$ (see Sec. 5.4). If the posterior probability of a training sample is beyond a certain threshold, this sample is classified with label L_ν, or otherwise, with label $\neg L_\nu$. Next, we analyze all positive and negative training samples and determine the ones associated with cover-type L_ν and $\neg L_\nu$. Finally, a measurement for the average classification accuracy is given as the frequency of correctly classified positive and negative samples to all training samples (Fig. 7.9). High classification accuracy means that most of the positive training samples belong to label L_ν and most of the negative examples to label $\neg L_\nu$.

7.3.3 Separation Between Semantic Labels

An entire evaluation of user-defined semantic labels does not only include the accuracy assessment of each individual cover-type, but also the analysis of how far labels can be separated against each other. If the defined semantic labels L_ν can be represented by the unsupervised content index ω_i, we compute the stochastic

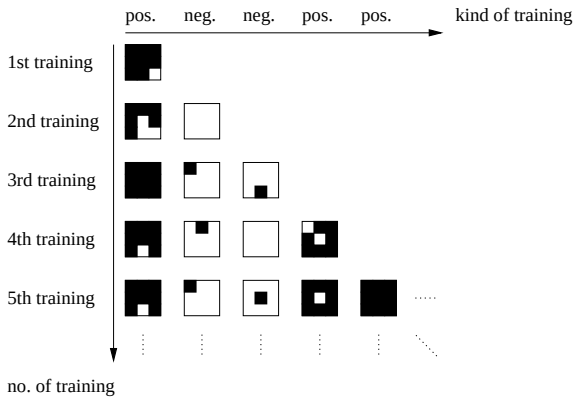


Figure 7.9: Schematic representation of a classification accuracy assessment based on training samples. Each positive or negative training is given by a set of 3×3 points. Black pixels are associated with label L_ν and white ones with label $\neg L_\nu$. For each new click in the system, the number of training samples associated with L_ν and $\neg L_\nu$ changes. In this example, after the 5th training iteration we obtain a classification accuracy of about 91%.

confusion matrix $M_{\nu\mu}$ with elements

$$p(L_\nu|L_\mu) = \sum_i p(L_\nu|\omega_i)p(\omega_i|L_\mu) \ . \tag{7.23}$$

For a perfect representation of labels by the signal content ω_i, the matrix $M_{\nu\mu}$ is diagonal, mathematically expressed: $p(L_\nu|L_\mu) = \delta_{\nu\mu}$. This means that a trained label L_ν would generate a label L_ν with probability 1 and 0 for all other labels. We compute the stochastic confusion matrix for a set of labels as demonstrated in (Tab. 7.3).

We want to mention that an entire evaluation procedure of the interactive training should also contain subjective issues. From a subjective point of view, a semantic label is the result of a series of positive and negative examples given by the user during the interactive training. Therefore, the quality of a cover-type depends on actions performed by the user and his capability to 'learn' the label, e.g. when a system operator defines a label 'water' and gives positive examples on 'clouds' or when a user is satisfied with a badly trained label. 'Good' training samples produce a well-defined label. A more detailed (subjective) evaluation involving the complexity of system operations and man-machine interactions is presented in Sec. 7.6 and 7.7.

L_μ	$p(L_\nu\|L_\mu)$ in %						
	L_1	L_2	L_3	L_4	L_5	L_6	L_7
L_1 = 'River'	**75**	4	5	4	4	4	4
L_2 = 'Riverbank'	5	**63**	7	6	5	7	7
L_3 = 'Cloud'	4	5	**72**	5	4	5	5
L_4 = 'Sisau forest'	2	2	3	**85**	3	2	3
L_5 = 'Dense-mixed forest'	3	5	6	6	**70**	5	5
L_6 = 'Soil'	3	4	5	4	4	**75**	5
L_7 = 'Grassland'	3	3	4	3	3	4	**80**

Table 7.3: Stochastic confusion matrix with elements $p(L_\nu|L_\mu)$ computed for the labels presented in (Tab. 7.2). Considering that the set of labels $\{L_\nu\}$ is trained only with spectral signals, they show a quite high separability.

7.4 Probabilistic Search

In addition to the quality of the stochastic link and the interactive training, also the "quality of user examples" and the "user's judgment of a completely defined label" have to be taken into account. Whether retrieved images are relevant for the system operator's training is difficult to answer because of a high diversity in the content of remote sensing images. Therefore, an evaluation of the retrieved results is quite a complex task.

In this dissertation, we verify the quality of the query results by collecting target and mis-detected images and measuring precision/recall. Moreover, we deal with the probability to forget an image and to over-retrieve images. We analyze the potential of the mining system to 'explore' the archive, that means, to provide user access to many stored images and not to limit the results to a small number of images.

Target and misclassified images

To obtain a measurement for the quality of the probabilistic search, we first analyze the queried images according to 'relevant' or 'irrelevant'. Relevant images are targets. We assume that each retrieved image is either a target or a misclassified one, independent of how strong they are covered by the semantic label. The relation of target/misclassified images in the query results can be verified both by visual inspection and by using ground-truth information.

In the first method the operator controls the queried images by visual inspection after finishing the cover-type training. To ensure a fast evaluation, the operator has only a look at the displayed top-ranked images of all relevant images. A queried image is a target if it contains the trained label from the operator's point of view. In (Tab. 7.4), we show the results of target and misclassified images for several semantic labels. Target and misdetected images in the query results can not only

label name	'Vis. inspection', %		'ground-truth', %		P_o	P_f
	targ.	misclass.	targ.	misclass.		
river	88	12	87	13	0.13	0.16
riverbank	100	0	75	25	0.25	0.64
clouds	61	39	64	36	0.36	0
sisau forest	94	6	80	20	0.20	0
dense-mixed forest	100	0	52	48	0.48	0
soil	100	0	97	3	0.30	0
grassland	100	0	98	2	0.20	0
Average	92	8	79	21	0.21	0.12

Table 7.4: Evaluation of the probabilistic search system function using semantic cover-types from (Tab. 7.2). Results are obtained by the user's visual inspection, ground-truth information and the probability to over-retrieve P_o and to forget an image P_f. Differences in the results are due to the different nature of evaluation methods.

be determined by visual inspection, but also by ground-truth information. An advantage of this method is that it can be performed in post-processing and it is not limited to the top-ranked images in the gallery. For our test dataset, we calculated an index for each image from the results of ML classification (Sec. 7.3). We derived target and misclassified images by comparing the query results with the generated index as presented (Tab. 7.4).

Precision and recall

In this section, we extend the retrieval performance evaluation from the last section by considering not only target and misdetected images in the query results, but also the potential of the probabilistic search for the whole image archive. In content-based image retrieval, 'recall' and 'precision' measurements are most often used to visualize how many relevant (target) and irrelevant (misdetected) images are in the highest ranked images(Korfhage 1997). Precision is the fraction of the retrieved images that are relevant to the query and recall is the fraction of the total number of relevant images (contained in the archive) that are retrieved. If we denote T the set of returned images and R the set of images relevant to the query (Fig. 7.10),

$$precision = \frac{|R \cap T|}{|T|} \tag{7.24}$$

and

$$recall = \frac{|R \cap T|}{|R|} \tag{7.25}$$

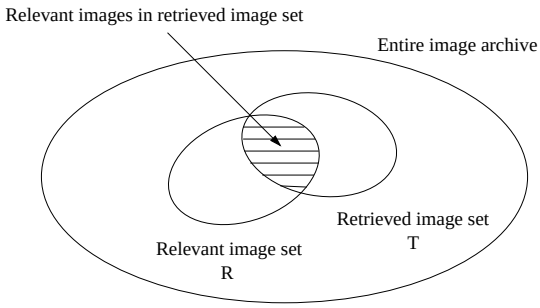


Figure 7.10: Recall and precision for an example probabilistic search and its answer set.

with $|\cdot|$ as the cardinality of a set. Usually, results of *recall* and *precision* are presented in form of precision-recall graphs. These graphs indicate retrieval effectiveness quite well since they include information about images that were not retrieved in the top-ranked images either. For instance, a low precision at high recall shows that the system has problems in capturing the diversity of the content of the images for a semantic label. In (Fig. 7.11), we present the PR graph for a selected label of (Tab. 7.2).

False alarms and non-detected images

Measurements that are closely related to both precision and recall are the probability to over-retrieve images P_o and the probability to forget an image P_f in the archive. Whereas *precision* and *recall* deliver the retrieval performance using the queried results from the database, P_o and P_f give a measurement of lost and confused

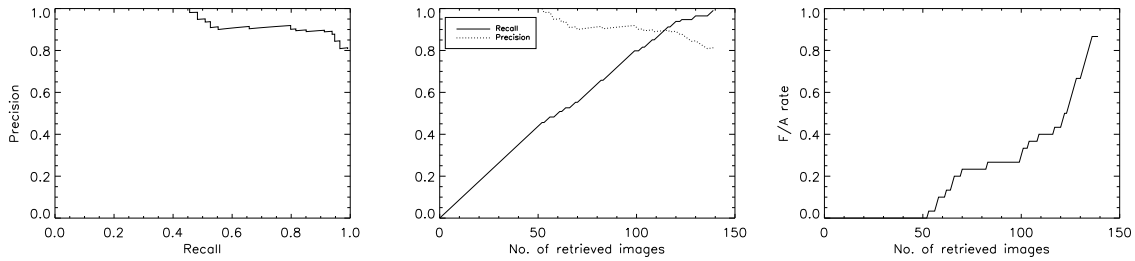


Figure 7.11: Precision-recall graph (left), precision and recall (middle) and false alarm (F/A) rate (right) indicating the quality of probabilistic search for the semantic label 'forest' computed for the Nepal dataset. For the label 'forest', the mining system returned in the 50 top-ranked images only relevant ones.

images in the database. Using R and T from the last section, we obtain

$$P_o = \frac{|T \setminus R|}{|T|} \tag{7.26}$$

and

$$P_f = \frac{|R \setminus T|}{|R|} \ . \tag{7.27}$$

In contrast to the last section, we do not analyze P_o vs. P_f graphs. Instead, we compute the probabilities for a number of semantic labels as outlined in (Tab.7.4).

Exploratory image archives

In the last sections, we determined the retrieval performance of our I^{2}M system by verifying queried images for individual labels L_ν. What we have not evaluated yet is the amount of information, that means images from the archive that were used by system operators for training and searching. Our aim is to measure the potential of the system in order to explore large remote sensing image archives, e.g. to gain a large number of ingested images or to have connection to only a small amount of data. Since the connection to images is realized via defined semantic cover-types L_ν, discovering the archive is equivalent to the analysis of all defined labels and their relations to images. In detail, for each cover-type L_ν we consider the most similar and top-ranked images from the query and consequently obtain the overall amount of retrieved images (Fig. 7.12).

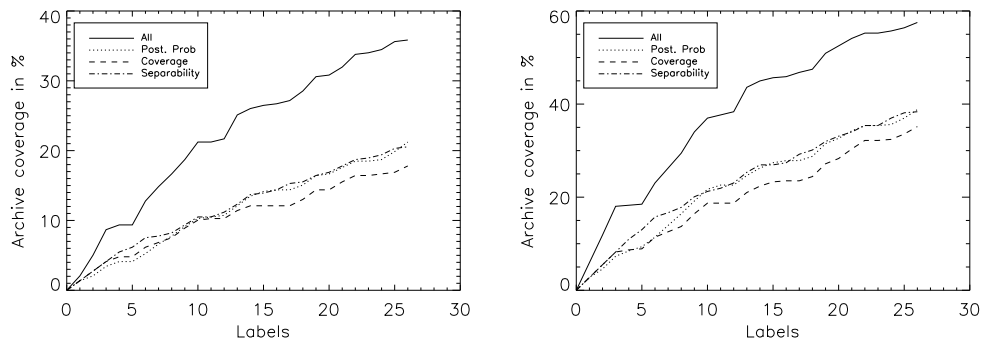


Figure 7.12: Exploration image archive for 26 defined semantic labels. The plots show the archive coverage using the first 6 (left) and 12 (right) top-ranked images in the search results. With all 26 labels, about 35% respectively, 60% of the entire archive is assigned to the semantic content.

7.5 Semantic Preservation and Generalization of Subspace Clusters

In Sec. 7.3 we outlined methods to analyze the classification accuracy of the interactive learning system module and in Sec. 7.4 we identified the system performance to query for images in the archive with structures similar to a defined cover-type. All approaches have in common the evaluation of individual system modules, but did not explicitly reflect the dependence on system performance and database size. According to the influence of the archive size on the semantic content, questions that arise are if the semantic image content is preserved for increasing volumes of data, how the number of primitive feature clusters influences the quality of semantic labels, and if clusters — generated for a subset of the archive — can represent the image content of the entire database. Answering these questions is rather important for system evaluation and enhancement since they involve the capability of the image information mining system to explore large volumes of data and how much information details are lost.

Semantic content vs. database size

In a first experiment, we analyze the interdependence of semantic image content and database size. Therefore, the whole database is partitioned into a sequence of subsets with increasing size. Each individual dataset is composed of the next smaller one and a fixed number of new images. Then, an unsupervised content-index is generated for the different feature spaces and volumes of data using the same clustering parameters. Various semantic cover-type labels are trained based on images from the smallest (initial) dataset. In order to apply the 'same' semantic labels on images of the different datasets, we fix the cover-type tracing parameters: location and type of training samples. With this information, we compute the stochastic link between content-index of a certain dataset and semantic labels. The quality/accuracy of cover-types to represent the semantic content for each dataset is measured by the sum of Kullback-Leibler divergence, Eq. 7.20, of the different feature models. Additionally, we analyzed the semantic image content by retrieval accuracy as the frequency of relevant images in the retrieval set. (Fig. 7.13).

Semantic content vs. number of clusters

As stated in Chap. 5, for learning a cover-type label from data using the naive Bayesian classifier, it is essential that structures of a label are fully described by at least one signal class model. In general, the more complicated the structures or patterns of a cover-type label, the higher the number of required clusters. However, the number of clusters influences the complexity of the Bayesian classifier and therefore has to be carefully selected. If we group the spectral and the textural feature

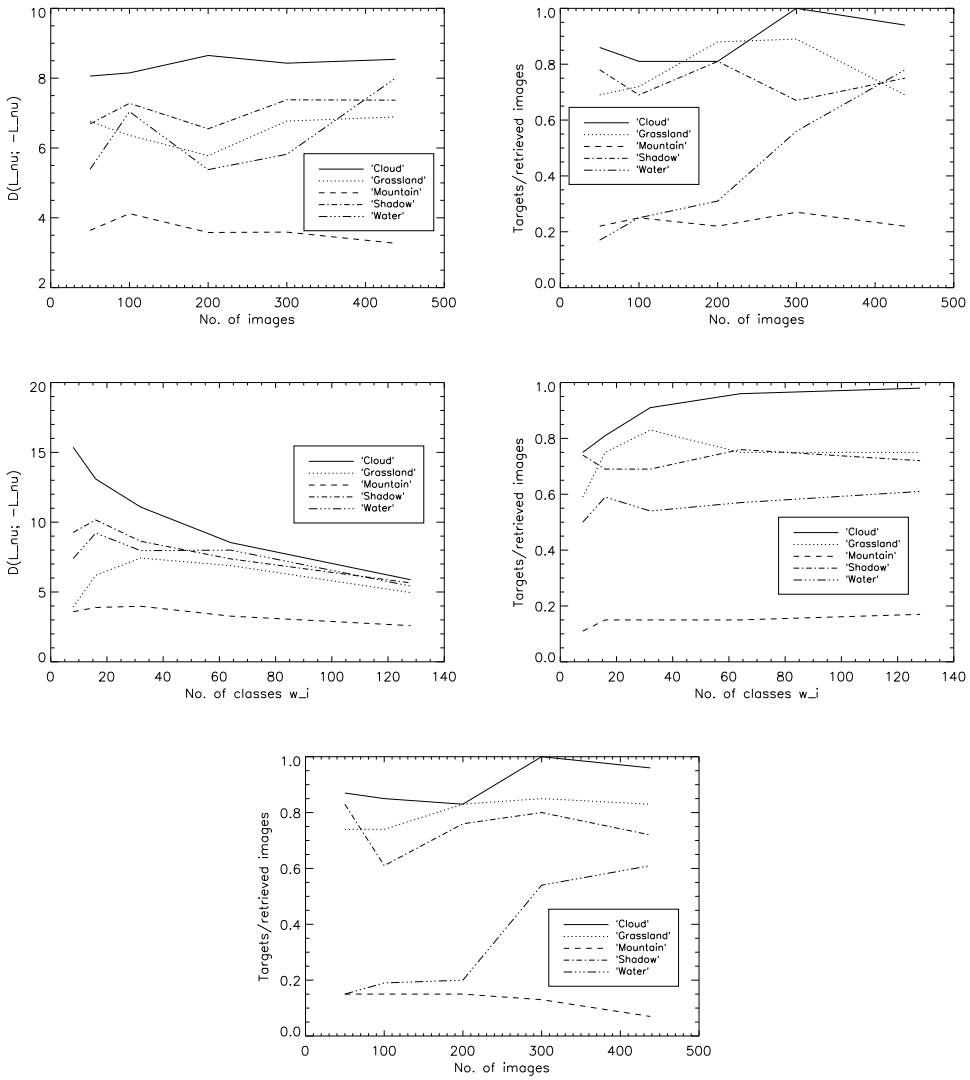


Figure 7.13: Evaluating the preservation of semantic image content and the generalization of subspace clusters. 1st. row: sum of Kullback-Leibler divergence $D(\cdot)$ (left) and retrieval accuracy (right) determined for several cover-type labels and volumes of data (right). 2nd. row: sum of Kullback-Leibler divergence $D(\cdot)$ (left) and retrieval accuracy (right) determined for several cover-type labels and number of feature clusters (classes) ω_i (right). 3rd. row: performance identification of subspace clusters generalization.

space into 50 clusters, we obtain $2 \cdot (50 + 50) = 200$ variables for pairwise disjunct labels L_ν and $\neg L_\nu$ in comparison to $2 \cdot (150 + 150) = 600$ variables for 150 spectral and textural clusters, for instance. Thus, both from the algorithmic complexity and diversity of structures of the image content point of view, verifying the influence of different numbers of clusters on the semantic image content is important. For this experiment we use one of the generated datasets as previously described. This time, we applied different numbers of clusters and computed Kullback-Leibler divergence and retrieval accuracy for several semantic cover-types as outlined in (Fig. 7.13).

Generalization of subspace clusters

The last experiment deals with the evaluation of the consistency of subspace clusters to represent the semantic content for the whole image archive. Since semantic labels L_ν are linked to the content-index (and finally images in the database) by the stochastic link $p(\omega_i|L_\nu)$, it is of interest to explore how far a certain vocabulary of signal classes ω_i — generated from a small subset of data — is sufficient to retrieve relevant images from the entire dataset. This verification can also be considered a first test in the direction of incremental clustering: producing a number of significant groups and classifying all samples according to the generated clusters.

We started this experiment by selecting 10 images from the whole archive (438 images) by random. Then, we clustered these data in an unsupervised way and derived a characteristic vocabulary of signal classes. According to signal classes ω_i, we assign each sample to a particular cluster using the minimum distance criterion. In order to test if the generated content-index is appropriate to query relevant images from the entire archive, we verify the retrieval results in terms of relevant (target) and irrelevant (misdetected) images. In (Fig. 7.13), we outline the results of this short case study.

7.6 System Operation

To analyze and evaluate the system operation and all user actions in the man-machine communication dialogue in I^{2}M, we trace all user actions during system interactions. In the graphical user interface as presented in (Fig. 5.7), we provide selected actions the user can apply for training labels, querying the database and analysing the images (Tab. 7.5). From human-machine interactions, we determine several measurements (Jermyn et al. 2002) (Gee and Cipolla 1999) to analyze and evaluate both the complexity of the system and the user's capability to use the possibilities the system offers.

action A^i	action name	explanation
A^1	click_center	user clicks with central mouse button, no action
A^2	click_right	user gives negative training sample
A^3	click_left	user gives positive training sample
A^4	image	user chooses new image for training
A^5	image_type	user changes image type between Landsat TM and ERS1
A^6	learn	user clicks on learn button
A^7	lut	user checks the lut checkbox
A^8	model	user chooses a certain model
A^9	order	user selects an image from a certain search criteria
A^{10}	pan	user clicks on pan radio button
A^{11}	reset	user clicks on reset button (restart learning)
A^{12}	search	user queries the database
A^{13}	tab_change	user changes the current tab
A^{14}	undo	user cancels the last training sample
A^{15}	zoom	user inserts a new zoom factor

Table 7.5: List of different types of user actions A^i. Each user action belongs to one of the listed types of actions.

Complexity $\mathcal{C}_1$: individual user actions

Assume we have a series of user-performed actions ${A_j}^i$ with consecutive action number j in a session and action type i, e.g. 'positive training' where each type of user action A^i is treated as a certain event E_i in the event space $\{E_i\} = \{E_1, E_2, \ldots, E_{N_A}\}$. For the events, the corresponding discrete probabilities p_i are given by

$$p_i = \frac{\sum_{j=1}^{N} {A_j}^i}{\sum_{j=1}^{N}\sum_{i=1}^{N_A} {A_j}^i} \tag{7.28}$$

as the frequencies of the occurrence of actions ${A_j}^i$ of type i in relation to the total number of performed actions. In Eq. 7.28, N indicates the total number of actions in a session and $N_A = |A^i|$ the number of actions of type i.

In (Fig. 7.14), we point out the communication principle between user and system. From the obtained probabilities p_i we can calculate the amount of information or complexity $\mathcal{C}_1$ contained in the man-machine communication by using Shannon's entropy (Cover and Thomas 1991)

$$H(p_1, p_2, \ldots, p_{N_A}) = -\sum_{i=1}^{N_A} p_i \log p_i \ . \tag{7.29}$$

High entropy or complexity can be interpreted as the user's capability to apply many of the functions the system provides. By contrast, low entropy or complexity

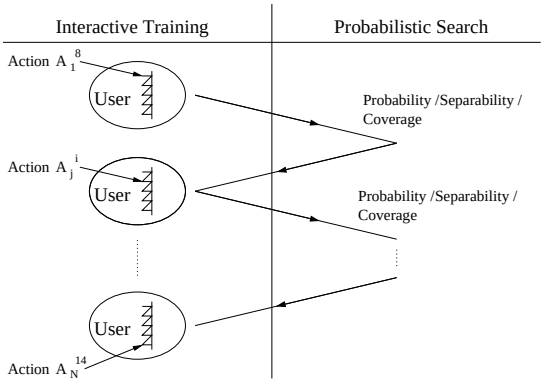


Figure 7.14: A user translates his interests to the system by a series of N interactions ${A_j}^i$. During the session the system learns from the user's actions and specifies his interests. From the user's actions, we obtain a measurement of the complexity of the man-machine dialogue.

indicates that the interacting person does not use the whole potential of the mining system.

Complexity $\mathcal{C}_2$: timing of actions

Now, we extend the complexity $\mathcal{C}_1$ from the previous section to the time domain. In this section, not only performed individual actions will be taken into account but also the time it takes to make an action.

The time ${t_j}^i$ for performing a certain action acts as a kind of weighting factor for ${A_j}^i$, e.g. if a user takes some time to determine a high quality training, the training action is higher weighted than a fast and imprecise training. From ${t_j}^i$ we derive the probability p_i for an event E_i as

$$p_i = \frac{\sum_{j=1}^{N} {t_j}^i}{T} \quad , \tag{7.30}$$

with $T = \sum_{j=1}^{N} \sum_{i=1}^{N_A} {t_j}^i$ as the total time for the complete session. Once again, we use Shannon's entropy to determine the complexity $\mathcal{C}_2$ of the man-machine dialogue. In order to consider only user actions for $\mathcal{C}_2$, we exclude the time for data transfer via Internet from our calculations. [2]

Complexity $\mathcal{C}_3$: probabilistic search

Having measured the operator-system complexity based on a set of individual actions or the corresponding times, we will now analyze the probabilistic search. Our aim

[2] During our experiments we observed great time variations for the data transfer from the server to the clients. The differences are due to overloaded networks, different browser, server problems and low level machines at the client site, for instance.

is to obtain a measurement of how far a client really uses information given in the probabilistic search results. High complexity means that the user takes all kinds of provided methods into account to specify and improve his search results: posterior probability, separability and coverage. By contrast, low complexity shows that the user is only partially using the information contained in the probabilistic search.

We define three independent classes ${A_j}^{i,m}$ with running action number j, action type i and class membership $m \in \{prob, sep, cov\}$ for posterior probability, separability and coverage. Each class is assumed to consist of the same action types A^i and, consequently, the same event space $\{E_i\}$. After separately calculating the probabilities ${p_i}^m$ for each of the m classes, we obtain the complexity of the system according to the probabilistic search as

$$\mathcal{C}_3 = -N_A \sum_{m=1}^{N_m} \sum_{i=1}^{N_A} {p_i}^m \log {p_i}^m \tag{7.31}$$

with N_A and N_m as the total number of events and classes.

Complexity $\mathcal{C}_4$: classes of actions

In this section, we describe how we define an overall complexity measurement based on classes of actions. Therefore, we partition action types A^i into three main classes $m \in \{training, search, image\}$, namely training, probabilistic search and analyzing images. Additionally, we divide the probabilistic search into posterior probability, separability and coverage. From the users' actions we compute the probabilities ${p_i}^m$ and consequently the complexity $\mathcal{C}_4$ as given in (Eq. 7.31).

Complexity $\mathcal{C}_5$: image type

As the final system operation (complexity) measurement, we analyze the user behaviour for applying different types of images, e.g. Landsat TM or ERS1. We aim at measuring the user's preference for a certain image type as well as his/her capability to incorporate different sources of information during the interactive learning process. In a similar way as outlined for $\mathcal{C}_3$, we define two identical event spaces, one for each image type. Again, low complexity indicates that the user is mainly focused on a certain category of image whereas high complexity reflects that the user is able to exploit the possibilities the mining system offers in terms of image type.

Assuming we have co-registered Landsat TM and ERS1 image datasets, we can define the classes ${A_j}^{i,m}$ with $m \in \{LS, ERS\}$, consecutive action number j and action type i. With calculated probabilities p_i^m for each image type class, we acquire the system operation complexity $\mathcal{C}_5$ according to Eq. 7.31. Of course, this measurement makes only sense if the user can choose between two or more categories of image types.

7.7 Human-machine Interaction

The development of intelligent human-machine interfaces for information mining applications is a difficult task since no well established guidelines and models of the functions that such systems should have are available (Guida and Tassa 1994). Despite this difficulty, we designed and implemented a graphical, intuitive and powerful visual interface that aims at controlling an underlying mining system, directly interacting with the users and enabling them to retrieve relevant images without the support of a human intermediary.

In this section, we survey both the functionalities of the interface and the information representation and communication aspects in the human–machine interactions. Finally, a tracing module enables us to follow user–performed actions to fully analyze the potential of the interface.

7.7.1 Functions of the Graphical User Interface

In the image information mining system, users express their interests with the help of a graphical user interface. The learning human-machine interface consists of an applet that is mainly developed to provide user interactions. Several servlets give all necessary data to the applet, apply the defined semantic label to all images in the archive and present the results to the system operator in form of a gallery of image thumbnails. The visual interface consists of four panels as outlined in (Fig. 5.7). In the image panel (top left of the GUI), an image selected by the user for training a semantic cover–type of his interest is displayed. The system operator can switch between different sensor data visualizations by clicking on one of the mutually exclusive buttons in the upper part of the panel. To perform a very precise training sample, the user can magnify an image portion at a location of his interest in the zooming panel (lower left of the GUI). Additionally, the visual interface assists the user in choosing the magnification ratio to optimally adapt to the user's needs. The posterior map (top right of the GUI), continuously gives the system operator a feedback about the current state of the trained semantic cover–type. Posterior probabilities $p(L_\nu|D)$, Eq. 5.18, are visualized and encoded from black (probability 0) to white (probability 1). If for individual image pixels the probability $p(L|D)$ is above a certain threshold, they are depicted in red color. After each positive or negative training sample (mouse click), the user is informed about the quality of the training. The divergence bars — one for each selected feature model according to Eq. 7.20 — in the histogram panel (lower right of the GUI) are updated after each interaction and indicate the quality of the link between the subjective semantic label and objective signal classes. The higher these bars, the stronger is the link. An overall quality of the link is given by the sum of the divergence bars.

The user can either click with the left mouse button to give an example point of the desired semantic cover–type or with the right mouse button to give an example

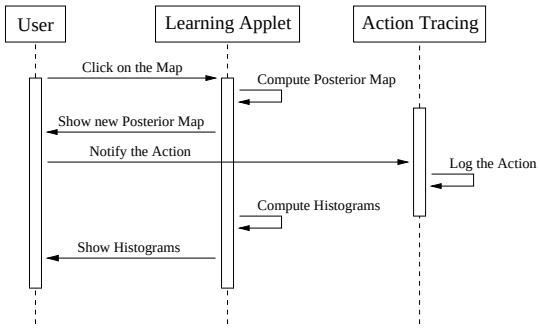


Figure 7.15: HMI state–diagram I: training a semantic label. A positive or negative training sample brings a newly computed posterior map in the learning applet and shows it to the user. Each 'learning' action — where, the kind of training and the panel on which the action was performed — is stored in a user–log. Not only the posterior map but also the divergence bars (histograms) are updated.

of what is not the desired cover–type. When the user moves the mouse pointer over one of the three panels, a linked cursor is shown on the other two panels in order to precisely indicate the chosen region. The graphical interface is used to track the user interactions, too. Each click on the panels, on the buttons or on another control is written in a user–log. Having presented the basic properties of the I^2M interface, we consider the implemented communication and information representation methods between user and system in the next section.

7.7.2 Human-machine Interactions

To enter the learning graphical interface, the user first has to register and to select up to four image feature models. The information flow is from the user to the system: the user knows about the relevance of certain feature models for learning a cover-type and the system makes use of this information and incorporates it in all further calculations and visualizations. Then, a gallery of randomly chosen images is presented to the user from where he has to select one. Again, the information flow is from the human to the mining system. The image servlet gives the image data to the learning applet and the label definition servlet gives the classification maps — according to the selected feature models — to the learning applet, too. When the data transfer is finished, the interface (Fig. 5.7) shows the information to the user and is ready for interactions.

The user can either define a new label or update an existing one by progressively giving positive and negative training samples using the left and right mouse button. The examples can be placed in the image, the zoomed image or the posterior map. After each click, the posterior map is updated with different colors indicating the probability of each point. Additionally, the divergence bars in the histogram panel

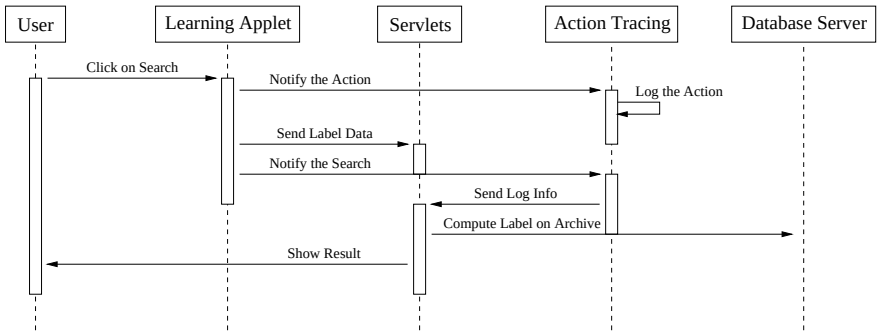


Figure 7.16: HMI state–diagram II: probabilistic search. After clicking on the search button, label information is sent from the learning applet to the servlet and probabilistic calculations are performed. The top-ranked images are shown to the system operator. He can select an image for further training or ingest the label in the inventory.

change accordingly to the user–performed actions. Both posterior map and divergence bars give the system operator a feedback about the quality of the performed training and the state of the semantic label. The interactions between user and system during the learning process are depicted in (Fig. 7.15).

Being satisfied with the trained cover–type, the user can click on the search button and the data containing the user interactions log is passed to the user management servlets. The divergence bars of the models are sent back to the label definition servlet that performs the computation of coverage, Eq. 5.29, posterior probability, Eq. 5.28, and separability, Eq. 5.31, on all the images in the archive. The results are presented to the user as a gallery of top-ranked images. With one of these images, the user can continue to learn the cover-type label or store the label in the database (Fig. 7.16). We call the continuous refinement of a semantic cover-type by interactive training and probabilistic retrieval 'iterative incremental learning'.

A tool worth mentioning in the image information mining system is user tracing. All user actions are traced and buffered from the learning applet. Each time the user presses the search button, the information is sent back to the user tracing servlet. This servlet inserts the performed actions and the user information in the database.

7.7.3 Communication and Information Representation Aspects in the HMI Dialogue

The characteristic of I^2M is that it fully incorporates the user in the information retrieval process: the system operator trains a semantic cover-type of his interest, the system queries the image archive for relevant data, delivers and displays the most relevant images to the user, who selects one image from a certain category of returned images, interprets the image, trains again, etc. This dialogue is repeated until the

Communication methods	Information representation
dialogue	
user $\longrightarrow$ system	
user $\longleftarrow$ system	
training samples	
positive	sign
negative	sign
visualization	
multiscale	signals
multi-sensor	signals
knowledge	
selected signal models	symbols
prob. search	signals
divergence bars	symbols

Table 7.6: Communication methods and information representation in I^2M. The mining system is based on a man-machine dialogue: the user learns the system and the basic properties of image data in the archive whereas the system learns the user's conjecture. The user trains a semantic cover-type of his interest with a series of positive and negative samples. For training, the user can choose multiscale and multi-sensor image data as visualizations of 2D signals. In the information mining system, the acquired knowledge is presented at the user's site as well as at the site of the system: the user selects two or more signal models and the system delivers the search results ordered according to posterior probability, coverage and separability. The communication between system and user is completed by divergence bars, one for each model.

user is satisfied with the trained label and the retrieved images. To help the user during the interactions and to find out whether a label is well-defined, several communication and information representation methods are applied as summarized in (Tab. 7.6). Before starting the interactive learning, the user has to express his prior knowledge about suitable image feature models that capture the desired cover-type characteristics. The semantic cover-type 'water' can be well described by spectral features whereas texture captures relevant structures for 'mountain', for instance. The system ingests this information in the DBMS and all further calculations are based on it. To perform precise training samples, the user is supported by different visualizations in the image panel. Especially the zooming panel has shown to be a quite useful tool for both training and image interpretation. Each training iteration — positive or negative sample — is related to a certain structure or object (sign) with location in the image space. All visualizations incorporate multiscale and multi-sensor two-dimensional signals. Moreover, the user is supported by divergence bars (symbols) in the histogram panel. They indicate how strong the selected

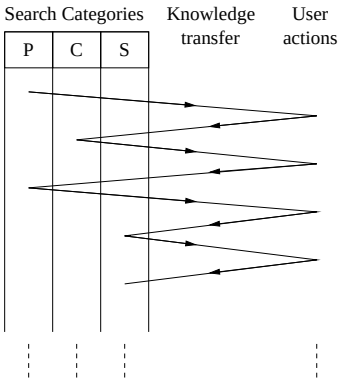


Figure 7.17: Iterative incremental learning of a cover-type label and knowledge transfer between mining system and user. The user provides the system with training samples, searches the entire archive for relevant images and continues learning with retrieved images. In this HMI dialogue, the user obtains information about the data and the system learns the user's conjecture.

models are for the desired cover-type and enable the user to remove weak feature models or substitute them by other (stronger) ones.

Based on all this information, the system queries the entire archive for images containing the trained semantic cover-type. The query results (returned top-ranked images) can be regarded as the knowledge of the system in terms of delivering the most relevant images according to user training. For further training, the user has to select one image from all retrieved images. The operator implicitly decides for a particular probabilistic search measurement and passes the knowledge of retrieval effectiveness and relevance back to the system (Fig. 7.17). On the other hand, the system incorporates this transmitted knowledge and presents the selected image to the user for further learning.

7.7.4 Analysis of Human-machine Interactions

Up to now we have presented the graphical human-machine interface, the interactions between user and system and the communication and information representation aspects in the HMI dialogue. The experiments reported in this section focus on the evaluation of the performance of the HMI.

First, we classify and identify the user's target structure by tracing the man-machine interactions. Then, the convergence of the learning process is analyzed using information-theoretic measurements and finally we apply the training feedback to predict which cover-types the user might be interested in. For the evaluation task, we trace the training actions of a professional image analyst that we obtained during a one-week system evaluation.

Iteration	Training (+/−)	User's comment
1	+	Starting sample, 'road' in flat terrain
Search archive, selected image from coverage retrieval set		
2	+	'road' in another image
3	−	negative training sample for smooth terrain
4	−	negative training sample, exclude flat terrain
5	−	negative training sample close to a 'road'
6	+	include 'road'
7	−	exclude rough terrain
Search archive, selected image from separability retrieval set		
8	+	include highway 'road' on a new image
9	−	specify 'road' by clicking next to 'road'
Search archive, selected image from coverage retrieval set		
10	+	include strong 'road'
11	−	exclude 'city'
12	+	include 'road' crossing
Search archive, selected image from separability retrieval set		
13	+	include 'road' on another image
14	+	include 'road'
15	−	exclude 'road'

Table 7.7: Information about the training samples shown in (Fig. 7.18). The user's aim was to include many similar structures in his positive training whereas the negative training shows a high range of contrary patterns.

Target structure classification and identification

In order to follow the user's training iterations, we use the information stored in the user-log. With this information, the following details about target structures from the traced human-computer interactions are extracted: the location of positive and negative training samples and the reasons why the examples are placed at a certain location. We superimpose the training samples on the images with indication of training iteration and the kind of training (positive or negative). We show the man-machine interactions for training using five Landsat images in (Fig. 7.18). The user both performed 15 samples on these five images to learn the semantic cover-type 'road' and commented on each iteration. With the positive examples, the operator tried to include many label-relevant (linear) structures. Negative ones are supposed to cover a high diversity of irrelevant objects and structures reflected by different feature models (Tab. 7.7).

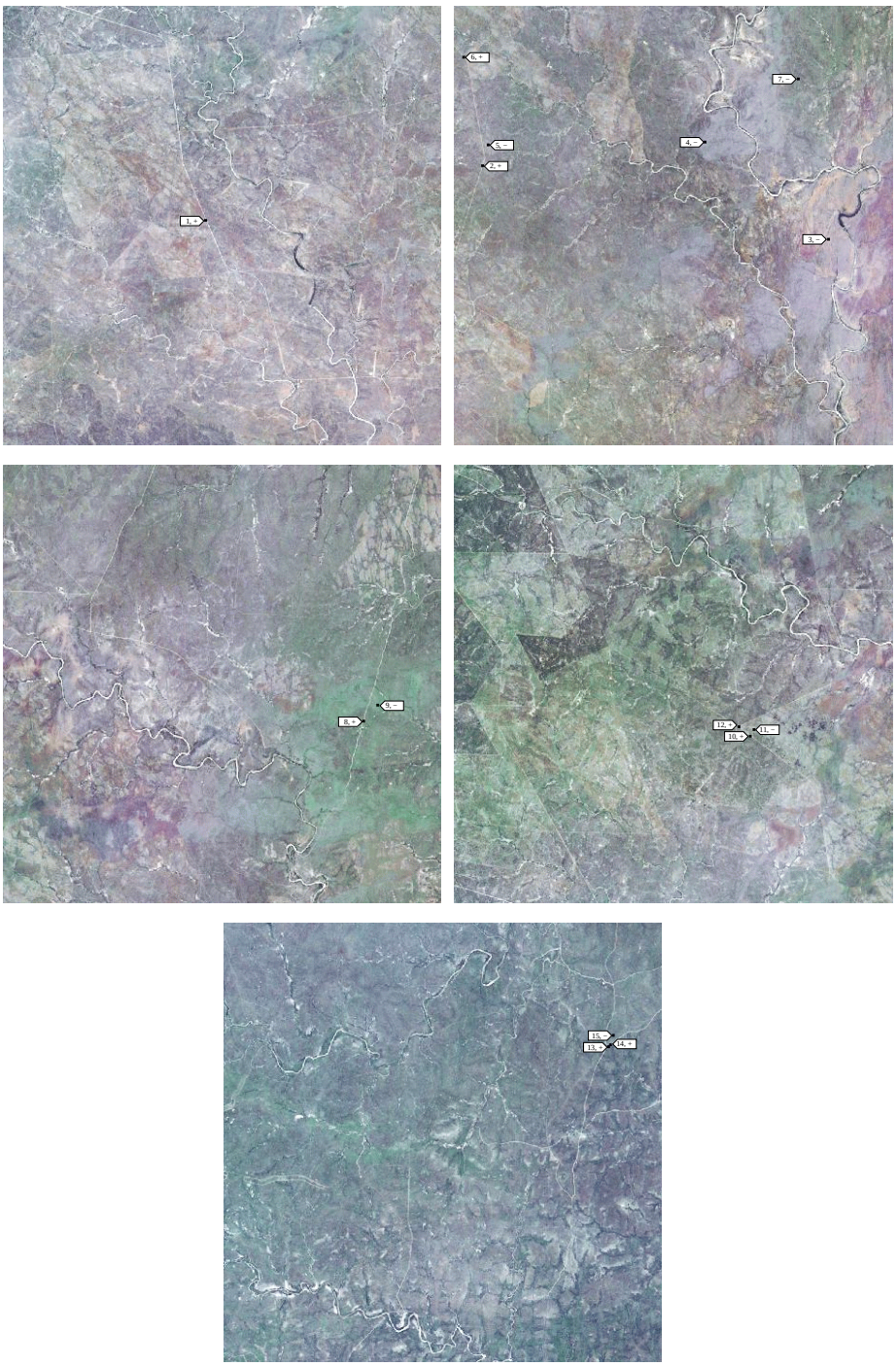


Figure 7.18: Definition of a semantic label 'road' based on spectral and textural feature models (Landsat TM) with 15 training iterations. Altogether, the user performed the training samples (8 positive and 7 negative) with 5 different images. The reasons for individual training samples are summarized in (Tab. 7.7).

Convergence of learning process

How efficient is the I^{2}M system in learning convergence? To answer this question, we analyze the human-computer interactions and measure the learning progress using Kullback-Leibler divergence. As previously mentioned, each positive and negative training iteration (mouse click) implicitly causes the update of the probabilistic link between content-index and semantic cover-type label. The amount of information in each learning iteration is reflected by the increase or decrease of the divergence bars. The questions that arise are (1) how the divergence for the different feature models behaves and (2) whether there is convergence.

In Fig. 7.19, we analyze the Kullback-Leibler divergence for the traced interactions as displayed in (Fig. 7.18). After the first two positive training samples, the divergence for both feature models almost equally increases. The third and all following interactions show that the defined semantic label is well represented by texture, the spectral model is of minor importance. Although the last three positive training iterations lead to a linear increase of the divergence for texture, the sum of both models just like the average quality of the training shows convergence. This example shows the real behaviour of the interaction between human and system: with only a few training samples an operator can define a cover-type label and query the database for relevant images. Of course, the number of iterations necessary to train a label depends on the selected feature models and the complexity of the label.

Quantities that influence the convergence of the learning process are the average number of training samples per image needed to classify the cover-type and the time for loading the interactive learning applet from the server. In Sec. 9.4, we depict these measures and show other evaluation results — acquired by a one-week system verification — indicating the performance of the graphical user interface, e.g. the

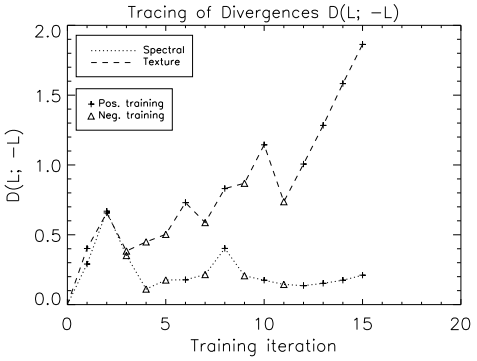


Figure 7.19: Kullback-Leibler divergence for interactive incremental learning across several images as depicted in the example in (Fig. 7.18). The two graphs, each one for a certain feature model, represent the divergence bars in the graphical interface after each interaction.

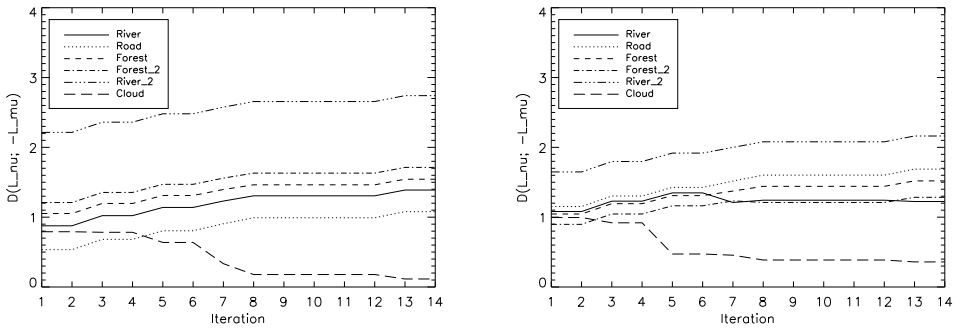


Figure 7.20: Results of matching a trained cover-type label 'cloud' using spectral (left) and textural (right) feature model.

overall time for defining semantic labels and the time for searching the database.

Matching user-specific semantic labels

The example in (Fig. 7.19) demonstrates that a user is able to define a particular semantic cover-type with just a few training iterations. Now, we want to predict the user's intentions. By monitoring the interactions with the database system it might be possible to somehow predict which semantic label or category of labels the user is interested in. This kind of forecast is called 'matching of user interests' in research literature and tries to identify the user's target by analyzing his actions (Koh and Mui 2001). In a similar way as we computed the Kullback-Leibler divergence $D(\mathcal{L}_\nu, \neg\mathcal{L}_\nu)$ between the two probability distributions $\mathcal{L}_\nu$ (positive training) and $\neg\mathcal{L}_\nu$ (negative training) in Eq. 7.20, we can extend this formula and determine the 'similarity' between a certain label and other labels in the inventory. Denoting L_ν the label a user is training, we can assess the similarity to any other label L_μ as

$$D(\mathcal{L}_\nu, \mathcal{L}_\mu) = \sum_i p(\omega_i|L_\nu) \log \frac{p(\omega_i|L_\nu)}{p(\omega_i|L_\mu)} \quad . \tag{7.32}$$

Labels that are 'close' [3] to L_ν are characterized by a very low divergence $D(\cdot)$ and labels dissimilar to L_ν show a high divergence. In (Fig. 7.20), we depict the similarity of a particular label to others during the interactive learning process. After just a few feedback samples, one semantic label close to 'cloud' is visible for both the spectral and texture feature model. Of course, the performance of this method depends on the applied signal models for interactive learning and the capability of the user to learn the system.

[3] 'Close' in the semantic sense.

7.8 System Information Flow

In this chapter, we have focused on the evaluation of individual system modules and the interaction between operator and system so far. What we have not analyzed yet is the communication between system levels and the flow of information during system operation. Thus, we explore the mining system from the communication channel view by determining information transmission between different system levels using information-theoretic quantities. However, before we measure the information between the basic levels of image content abstraction — image, class (cluster) and semantic — we will point out the characteristics of the mining system from the communication channel view. In addition to the information-theoretic treatment, we show how user-specific cover-types are linked to primitive clusters in the spaces of different feature models.

7.8.1 Communication Channel View

From the communication channel point of view, the image data at the lowest level in the hierarchical abstraction of image content is regarded as a message transmitted by an imperfect communication channel as two-dimensional signals to the user. The difficulty in understanding the received image information in form of symbols and semantics in a certain semiotic context, and inferring about the original image causes the problem of unsupervised image content modeling. Through the hierarchical image content characterization, the image retrieval system can be viewed as a composed communication channel. The imperfect nature of the system in combination with the well-known statement of information theory, which says that data processing cannot increase information, entails that each level in the hierarchical scheme is associated with a certain loss of information. The purpose of this section is to measure the information (association) between different system levels using information-theoretic quantities.

7.8.2 Information-theoretic Measures Between System Levels

In Sec. 3.3, we summarized basic quantities used in information theory and statistics to measure information. Now, we apply them to determine the information flow in the mining system. Therefore, we define three basic levels of different semantic abstraction as depicted in (Fig. 4.1): image space $\boldsymbol{I}$, content-index space or class space $\boldsymbol{\omega}$ and semantic label space $\boldsymbol{L}$.

Image space — class space

First, we evaluate the correlation between image space $\boldsymbol{I}$ and class space $\boldsymbol{\omega}$. For this verification, we use the multi-mission datasets consisting of 438 geocoded and co-

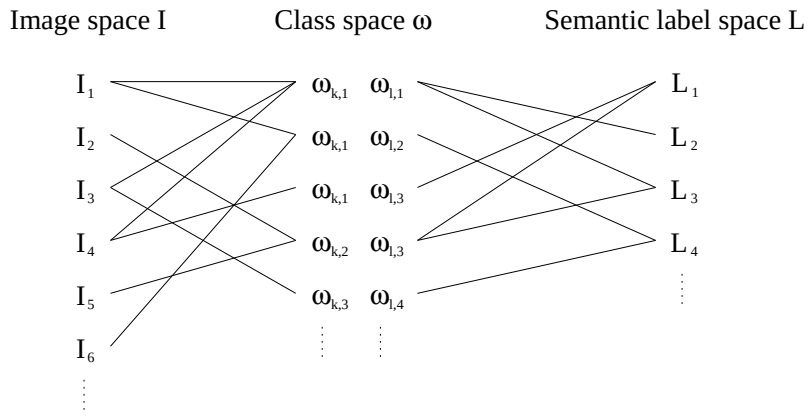


Figure 7.21: Association between the elements of different levels of the hierarchical image content characterization: image space $\boldsymbol{I}$, class space $\boldsymbol{\omega}$ and semantic label space $\boldsymbol{L}$. Note, images I_ζ are linked to cover-types L_ν via the joint space of signal classes $\omega_{k,i}$ and $\omega_{l,i}$ of model k and l, e.g. spectral and texture at a certain scale.

registered Landsat TM and ERS1 images (see Sec. 9.1 for more details). From optical Landsat TM images we use spectral and textural properties at different scales. From ERS1 radar data we use enhanced model-based despeckled intensity information and spatial characteristics based on the Gauss-Markov random field texture model. From the generated content-index we can compute the mutual information between image space $\boldsymbol{I}$ and class space $\boldsymbol{\omega}$ as

$$I(\boldsymbol{I};\boldsymbol{\omega}) = \sum_{\zeta,i} p(\omega_i|I_\zeta)\, p(I_\zeta)\, \log \frac{p(\omega_i|I_\zeta)}{p(\omega_i)} \quad , \tag{7.33}$$

where $p(\omega_i|I_\zeta)$ indicates the posterior probabilities of signal classes ω_i given a certain image I_ζ from the archive. Prior probabilities for signal classes and images are given by $p(\omega_i)$ and $p(I_\zeta)$, respectively.

In (Tab. 7.8) we summarize the calculations between image and class space. The measures indicate the information transmitted from image data through feature extraction and unsupervised content-index generation (clustering) to the class space. Note that for radar data the computed mutual information $I(\boldsymbol{I};\boldsymbol{\omega})$ is much lower than for Landsat TM. For Landsat TM images, mutual information is minimal for texture at lowest scale.

Image database complexity

The association between image space and class space can further be used to measure the complexity of images in the archive. Since the query performance of content-based image retrieval systems depends on the complexity of the data, analyzing the image database that is used for testing is rather important for evaluation. Similar

	Sensor	Signal models	Scale	I($\boldsymbol{I}$; $\boldsymbol{\omega}$)
A	Landsat TM	spectral	30m	1,41
B	Landsat TM	GRF	30m	0,92
C	Landsat TM	GRF	60m	1,23
D	Landsat TM	GRF	120m	1.39
E	ERS1	MBD	60m	0,53
F	ERS1	MBD	120m	0,59
G	ERS1	GMRF	60m	0,43
H	ERS1	GMRF	120m	0,56

Table 7.8: Mutual information $I(\boldsymbol{I};\boldsymbol{\omega})$ between image space and class (content-index) space. The class space $\boldsymbol{\omega}$ was separately generated for 438 co-registered Landsat TM and ERS1 images based on different signal models.

to the method of (Rao et al. 2002) that applies image database statistics and information theory to determine the complexity of image databases, we measure the information between image space and class based on Kullback-Leibler divergence.

In comparison to mutual information, Kullback-Leibler divergence can be applied to determine the complexity of a single image in the entire archive. Thus, we define the prior probability p_i as the probability $p(\omega_i)$ of a particular class ω_i in the global (across image) class space and the posterior probability q_i as the probability $p(\omega_i|I_\zeta)$ of a class ω_i given a particular image I_ζ. For these two quantities, Kullback-Leibler divergence is given according to Eq. 3.20 as

$$D(p_i;q_i) = \sum_i p(\omega_i|I_\zeta)\ \log\ \frac{p(\omega_i|I_\zeta)}{p(\omega_i)} \tag{7.34}$$

and can be interpreted as the complexity of a certain image in relation to the entire archive. A high complexity is associated with a low divergence and vice versa. Of course, the image complexity expressed by Eq. 7.34 highly depends on the ability of the applied signal models to describe the image content and to capture characteristic image structures. We show the image complexity of five images as depicted in (Fig. 7.22). These images belong to the Landsat TM dataset that is composed of 438 images. For each of the depicted images we calculated the relative entropy $D(\cdot)$ for the applied signal models and compared the results (Fig. 7.23). Although there are differences between the signal models, a correlation between signal models is visible. An interesting fact is that the GRF texture model at a scale of 30m (original resolution) delivers the smallest entropies. It is the model that captures most of the image structures in the archive.

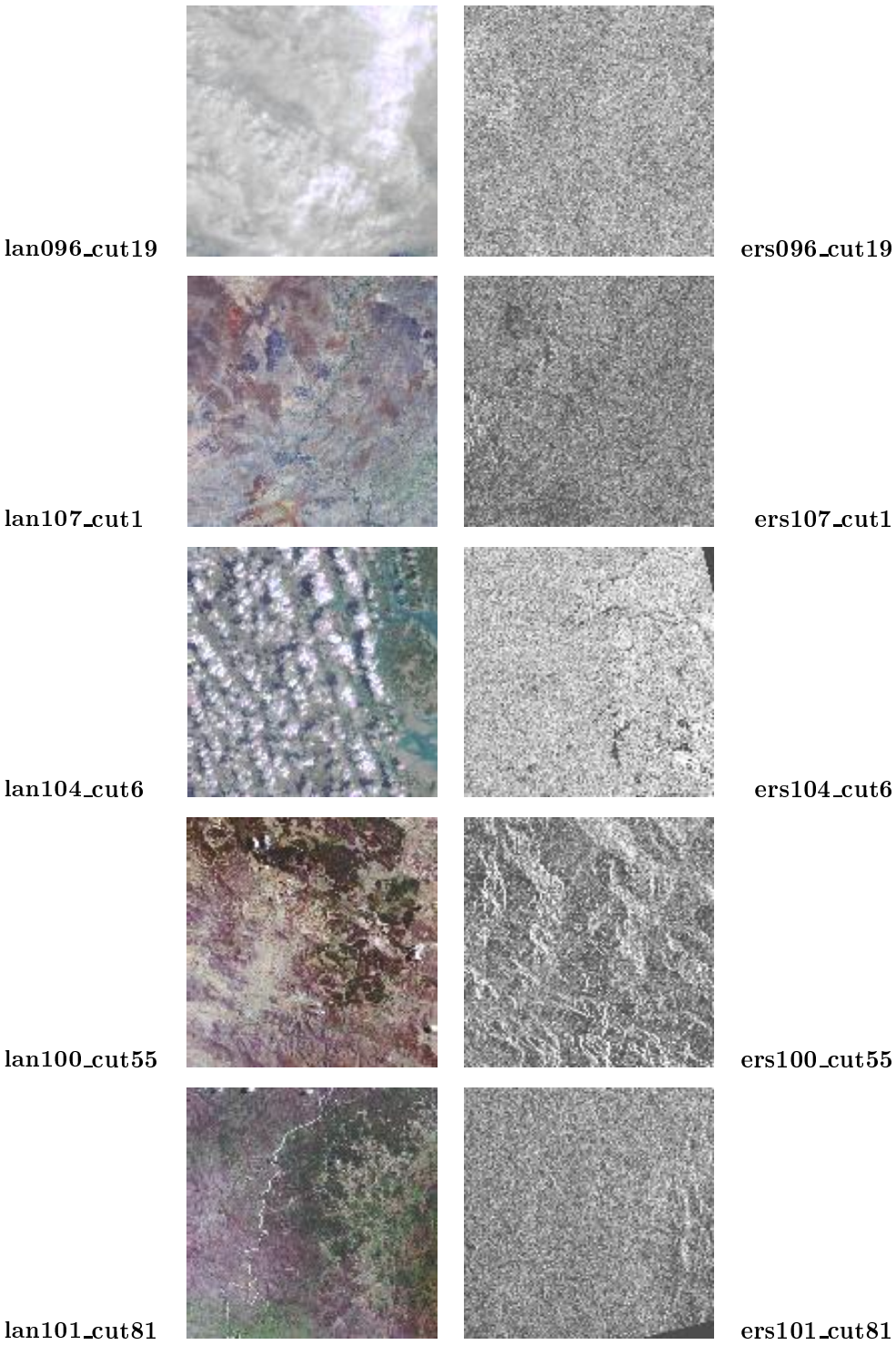


Figure 7.22: Example multi-mission dataset containing five co-registered Landsat TM (left) and ERS1 (right) images. For each image we computed the complexity in the archive using different signal models as outlined in (Fig. 7.23).

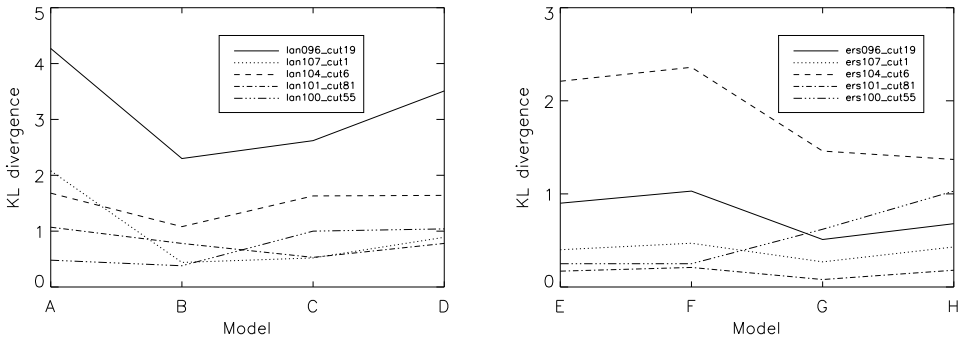


Figure 7.23: Information-theoretic complexity of Landsat TM (left) and ERS1 (right) images as depicted in (Fig. 7.22). To describe the content of both datasets, the models from (Tab. 7.8) are applied. The complexity measures show that 'lan096_cut19' is a very simple image in the Landsat TM archive whereas 'lan107_cut1' and 'lan101_cut81' are quite complex. For ERS1 data, 'ers104_cut6' shows to be rather simple and 'ers101_cut81' is the image with the highest complexity.

Class space — space of semantic labels

In the same way as we calculated the mutual information between image space and class space, we can compute the mutual information between the next basic levels in our hierarchical representation: class space $\boldsymbol{\omega}$ and semantic label space $\boldsymbol{L}$. As mentioned in Sec. 4.1, the first three levels in the hierarchy are determined in a complete unsupervised and application-free way. Consequently, the information between image and class space can be seen as a complete objective measure. Subjective user-related concepts neither have an influence on $I(\boldsymbol{I};\boldsymbol{\omega})$, Eq. 7.33, nor on $D(\cdot)$, Eq. 7.34.

Since a user-defined semantic cover-type label L_ν is the result of several human-machine interactions, the information between $\boldsymbol{\omega}$ and $\boldsymbol{L}$ can be seen as subjective and objective. More precisely, the stochastic link $p(\omega_i|L_\nu)$ derived from the user's feedback connects objective signal classes ω_i to the user-specific interpretation of the image content in the form of semantic cover-types L_ν. Therefore, the set of probabilities $p(\omega_i|L_\nu)$ is the central element of our analysis. With $p(\omega_i|L_\nu)$ as the posterior probabilities and the priors $p(\omega_i)$ and $p(L_\nu)$, we can compute the mutual information between signal class space and semantic label space as

$$I(\boldsymbol{\omega};\boldsymbol{L}) = \sum_{i,\nu} p(\omega_i|L_\nu)\, p(L_\nu)\, \log \frac{p(\omega_i|L_\nu)}{p(\omega_i)} \quad . \tag{7.35}$$

Note that $I(\boldsymbol{\omega};\boldsymbol{L})$ is separately computed for each signal model that the user selected to learn a cover-type of his interest. We can interprete this measure as the quality of semantic cover-types L_ν to capture the entire diversity of structures and patterns

represented by the content-index. In (Fig. 7.24) we show the behaviour of mutual information for a sequence of semantic labels defined by various users. If semantic labels are inserted that differ from the existing ones in terms of association to signal classes ω_i of the different feature models, mutual information increases. By contrast, mutual information decreases if cover-type labels are similar to the existing ones in the DBMS.

The mutual information between class space and semantic space can further be applied to analyze how far the whole diversity of image structures in the archive is captured by semantic cover-types. It can result in a new function to dynamically control the semantic image content, that is, to filter out overlapping cover-types or to support users in the training process.

The information-theoretic quantity $I(\boldsymbol{\omega}; \boldsymbol{L})$ represents the connection between a set of semantic labels L_ν and the characteristic vocabulary of signal classes ω_i in just one number. Neither does it yield information about the relevance of certain clusters for the definition of semantic labels nor about the location of training samples in the global feature spaces. Before we show how feature space clusters are 'filled' by training samples of inserted semantic labels in the last part of this section, we will first point out how the mutual information between image and semantic label space is derived.

Image space — space of semantic labels

After computing the mutual information between image–class space and class–semantic space, we can directly make the connection between cover-type labels L_ν and images I_ζ. Therefore, we start with Bayes' formula

$$p(L_\nu|\omega_i) = \frac{p(\omega_i|L_\nu)\, p(L_\nu)}{\sum_\nu p(\omega_i|L_\nu)\, p(L_\nu)} \tag{7.36}$$

to obtain the posterior probabilities $p(L_\nu|\omega_i)$ from the likelihoods $p(\omega_i|L_\nu)$ and the priors $p(L_\nu)$. Note, in comparison to Eq. 5.19, where we calculated the probabilistic link between ω_i and L_ν for a pair of disjunct cover-types L_ν and $\neg L_\nu$, the integration in the denominator of Eq. 7.36 has to be performed over the whole semantic label space. Having defined the posterior probabilities $p(L_\nu|\omega_i)$, we can infer the probability of a semantic label L_ν given a certain image I_ζ as

$$p(L_\nu|I_\zeta) = \sum_i p(L_\nu|\omega_i)\, p(\omega_i|I_\zeta) \ . \tag{7.37}$$

These probabilities indicate how strong certain cover-type labels L_ν are linked to images in the database.

In a similar way as we calculated the mutual information in the previous sections,

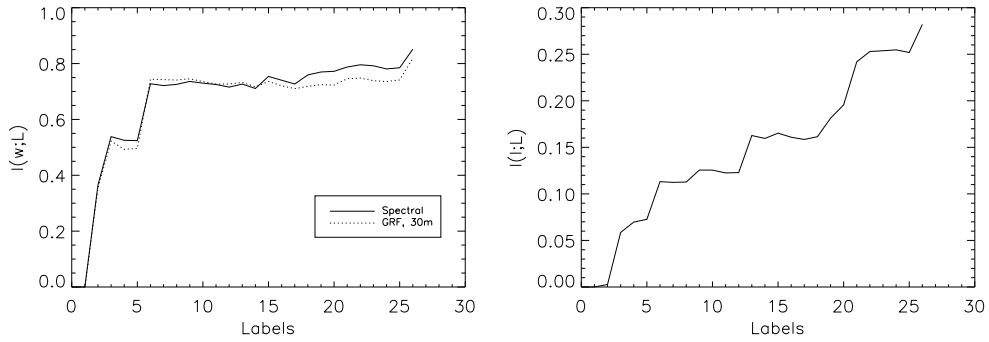


Figure 7.24: Mutual information $I(\boldsymbol{\omega};\boldsymbol{L})$ (left) and $I(\boldsymbol{I};\boldsymbol{L})$ (right) for a sequence of defined semantic cover-type labels. Note, the mutual information $I(\boldsymbol{\omega};\boldsymbol{L})$ was separately computed for each signal model whereas $I(\boldsymbol{I};\boldsymbol{L})$ is given via the joint space of signal classes (Fig. 7.21). The increase of mutual information depends on the diversity of defined labels.

we obtain this information by

$$I(\boldsymbol{I};\boldsymbol{L}) = \sum_{\zeta,\nu} p(L_\nu|I_\zeta)\, p(I_\zeta)\, \log \frac{p(L_\nu|I_\zeta)}{p(L_\nu)} \tag{7.38}$$

with posterior probabilities $p(L_\nu|I_\zeta)$ from Eq. 7.37 and priors $p(I_\zeta)$ and $p(L_\nu)$. In (Fig. 7.24) we display the computational results based on the same semantic cover-types that are used to analyze the association between class and semantic space. Whereas $I(\boldsymbol{\omega};\boldsymbol{L})$ indicates how much information the clusters contain about semantic label, $I(\boldsymbol{I};\boldsymbol{L})$ directly shows the semantic coverage of the image archive. Consequently, we can — at least qualitatively — infer the amount of images in the archive that are connected to cover-types. And if a new semantic label is inserted in the DBMS, we can assess its novelty in relation to the existing ones. Notice the similarity between (Fig. 7.12) and (Fig. 7.24). While the first figure shows the coverage of the database according to retrieved top-ranked images, the second shows the information-theoretic association between image and semantic label space.

7.8.3 Cluster Occupation by Semantics

As previously outlined, representing the association between signal classes ω_i and semantic cover-type labels L_ν by mutual information is rather abstract and a conclusion about the actual coverage of the primitive feature spaces by the semantic image content cannot be drawn. In order to verify the information flow from users to image data, we have to analyze how user-specific semantic labels are connected to primitive clusters in the multi-dimensional spaces of the different feature models. In this context, the questions arise (1) if there are few significant clusters to represent all semantic labels, (2) where the occupied clusters are located and (3)

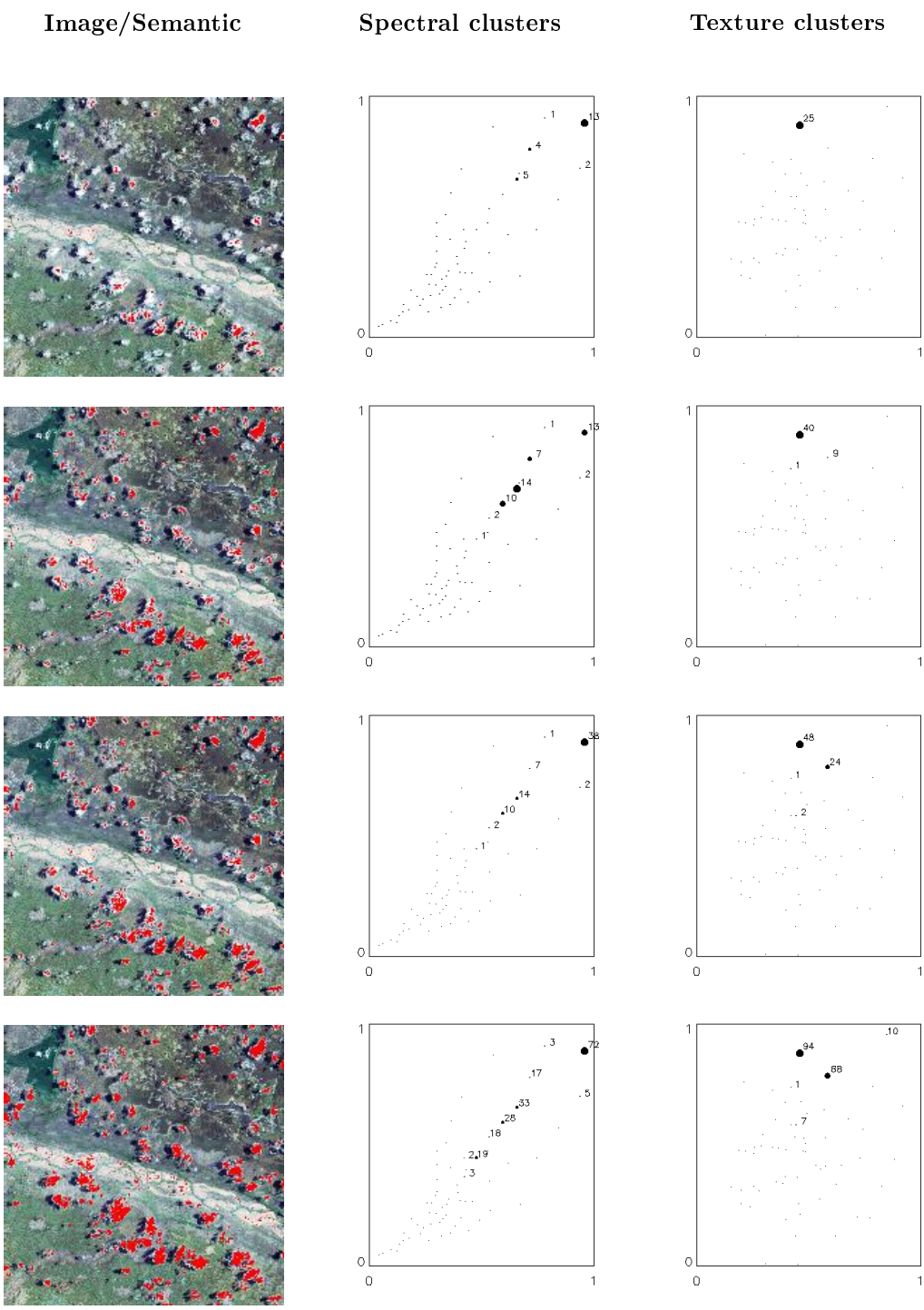


Figure 7.25: Information flow from semantic (user concept) to spectral and textural feature clusters (unsupervised image content). Row 1 to 3: 1st., 2nd. and 3rd. positive training iteration. Row 4: final (inserted) semantic cover-type label. Image content that belongs to the semantic cover-type is displayed in red.

how the different model feature spaces vary. Before we present the association of all semantic labels — inserted in the DBMS after an evaluation test week — to feature space clusters in Sec. 9.5, we describe the methodology by means of an example (Fig. 7.25). A user trains a cover-type label 'cloud' based on spectral and textural image parameters. After each positive or negative training iteration (left/right mouse click), we computed the 'filling' of the clusters in the different feature spaces by performed training samples. While the number of occupied clusters by training samples increases from iteration to iteration in the spectral feature space, there are two predominant clusters in the textural feature space. Consequently, the image structure complexity for cover-type 'cloud' can be well described by a few compact clusters in the textural feature space while it is distributed to several clusters in the spectral feature space.

7.9 Further Evaluation Issues

Evaluating an image information mining system is a challenging and complex task since both functions and their interrelations have to be taken into account to indicate the overall system performance. Particularly the interaction between system and user via an interface requires tools to analyze both the objective technical system quality and the user-related subjective concepts. With the algorithmic protocol presented in this chapter, we meet the requirements for the detailed validation of an interactive image retrieval system. However, there are other important evaluation aspects that have not explicitly been outlined yet but are worthy of equal consideration.

From the man-machine interface design point of view, the adaptivity to the users' abilities, preferences and predilections demonstrates the flexibility of a system. Important for an efficient image content access are personal differences (Egan 1988) in background knowledge (sensor characterization, physics of scattering, etc.) and application domain, for instance. In order to suit these requirements, the I^2M system makes the distinction between registered novice and expert users. Whereas users from the first category can only select certain pre-defined combinations of feature models, expert users can choose models from the whole range of ingested data. Not only the adaptivity to personality differences, but also the suitability for various applications reflects the flexibility of a system. Since the content-based retrieval of remotely sensed imagery deals with very heterogenous datasets in terms of type and resolution, a system should be adapted to the characteristics of the image without major modifications in system design and implementation. The mining system fulfilled these requirements and was successfully applied for retrieving optical and radar images, remote sensing change detection and medical help diagnosis.

Another parameter that should be tested in an image retrieval system is user guidance. The difficulty in evaluating a system in terms of user guidance is its overall

relation to operators. User guidance should give inexperienced users necessary help so they do not get frustrated. But at the same time expert users should not be limited in communicating with the system. What makes a system robust is the use of self-explaining communication objects like buttons with a certain text field. I^{2}M uses these tools and only in the interactive learning GUI the user can ask for help (Fig. 5.7).

What is not very much validated in image retrieval is the long-term stability or the error resilience of a system. Rather often only the outcomes of single user sessions are analyzed. In order to evaluate the long-term stability of the I^{2}M system, we therefore counted the errors during a one-week extensive system verification test as outlined in Chap. 9.

7.10 Conclusions

In this chapter, we have discussed the following items:

- A detailed evaluation methodology adapted to the architecture of the mining system has been carried out. We described methods to validate the objective technological quality of individual system components, included subjective human factors and verified the transmission of information from the user's site to the archive during system operation.

- We showed how the information content of primitive spectral and textural image parameters can be determined. Based on this information, we can analyze which image structures and objects (at a certain scale) can be well described by the applied signal models and where problems may arise.

- In order to reduce the large amounts of extracted features, a global unsupervised clustering for the different feature spaces is performed. We demonstrated several measurements like divergence or Bayes' probability of error to analyze the quality of the obtained clusters and, additionally, how inaccuracies of the features impair the unsupervised content-index.

- Next to primitive features and clusters we verified the performance of the interactive learning and the probabilistic search system modules. We outlined how the quality of a training sample given by the user can be measured in an information-theoretic way, how the classification accuracy of semantic cover-types across the entire archive can be determined, how far the system retrieves images according to the user's conjecture and how far the system preserves the semantic image content for different volumes of data. Whereas the information content of features and clusters was determined in a completely objective way, analyzing interactive learning and probabilistic search includes both objective and subjective factors.

- The third category of evaluation methods focuses on purely subjective concepts. We described methods to quantize the complexity of interactions between user and system, the capability of a user to 'learn' the system with a limited number of training iterations and the matching of semantic labels.
- We applied information-theoretic quantities to determine the association between the elements at the different levels of image content abstraction. Particularly the application of Kullback-Leibler divergence to analyze the complexity of images in the entire archive demonstrated its relevance.

8

Evaluation Procedure

In the previous chapter we pointed out various methods to assess the objective technical performance of the mining system and included subjective evaluation concepts by analyzing the interactions between users and system. With these methods we have a number of powerful tools to analyze both individual system components and their interrelations during system operation. What we have not identified yet is the degree of how far users find the image information mining system helpful and whether they are satisfied with its functions. However, such a performance test is not an easy task since many real and expertized users from different application fields have to be included in the experiments to obtain reliable and statistically correct results. Furthermore, tests with humans are usually hard to perform, subjects must be carefully selected and experiments well-designed in order not to influence the outcomes and shift them in the desired direction. Large-scale tests with many participants also put strong requirements on the organization since a high number of tests have to be carried out in a comparatively short time, results have to be recorded and the retrieval system must be kept stable during the experiments.

Because of the difficulties mentioned, an overall system evaluation with objective and subjective issues is a rather difficult task. In order to verify the overall effectiveness of I^2M under real world conditions, a particular organization, certain tools and an efficient procedure are needed. They are outlined in the following sections.

8.1 Organization

The evaluation of a complex image retrieval system like I^2M requires a special organization and development of appropriate tools as shown in (Fig. 8.1). For the analysis and quantification of an objective performance of the system, a tool is implemented to both trace the users' interactions and to statistically analyze the results of the traced parameters. The objective evaluation is based on measures of

- classification error to assess the quality of the interactive training (see Sec. 7.3),
- information transfer to evaluate the quality of learning semantic labels (see Sec. 7.3) and

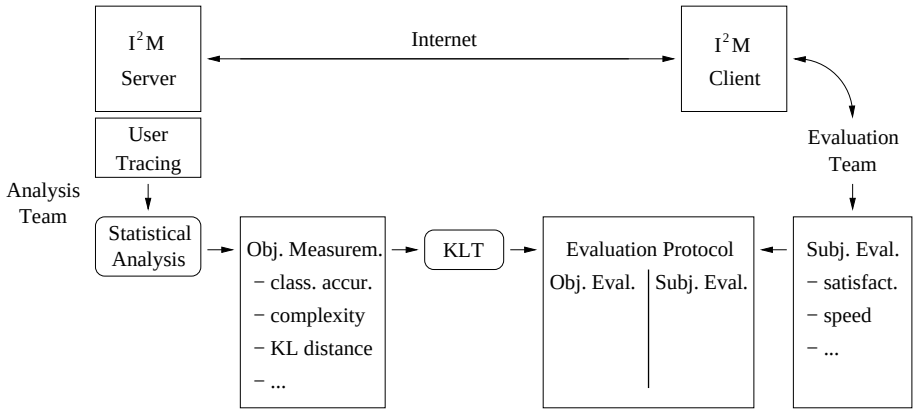


Figure 8.1: I²M system architecture for the evaluation purpose.

- complexity for the man-machine communication dialogue (see Sec. 7.6).

During the operation of the mining system, the report of objective evaluation measures is generated by off-line processing based on the user tracing information as shown in (Tab. 8.1).

At the subjective level, the user (evaluator) was asked to qualitatively rank the degree of satisfaction after the operation of the system. The evaluation was formalized as a questionnaire, where labels could be marked as 'very good', 'good', 'acceptable' and 'unsatisfying'. The overall system evaluation was realized as a one-week performance test with various participants: image analysts from the European Union Satellite Center (EU-SC), scientists from Nansen Environmental and Remote Sensing Center (NERSC) and technical stuff from the European Space Agency (ESA) at ESRIN. The evaluation results were analyzed by merging the two reports — objective (technical quality) and subjective (user satisfaction). The overall aim of this system verification was to test how well objective and subjective measurements meet, an attributed semantic label of 'very good' and the objective measurements for this cover-type should agree, for instance. Examples are shown in (Tab. 8.2).

8.2 Experimental Results

After defining the basic conditions of the evaluation procedure in the last section, we are interested in the relevance of the measured objective quantities to reflect the user's satisfaction. The approach we follow is similar to the work of Healey and Picard (Healey and Picard 1997). Whereas both authors analyzed physiological signals like skin conductivity, blood volume pressure, respiration and an electromyogram on the masseter muscle, we deal with measurements directly derived from human-computer interactions. Healey and Picard described a method to collect training data, extract features from the recorded signals and determine

time	action type	parameters
10:05:20	MODEL	1013 1032, 2, Nepal_spectral
10:05:20	MODEL	1013, 1038, 2, Nepal_texture_band4
10:05:20	IMAGE	1013, 5, TM_991106r_4_14, New label
10:05:25	TAB_CHANGE	1013, Learn
10:05:33	APPLET_LOADED	1013
10:05:44	PAN	1013
10:05:58	LEARN	1013, [x=127,y=179]
10:06:01	CLICK_LEFT	1013, Zoom Panel, [x=252,y=207]
10:06:11	PAN	1013
10:06:22	LEARN	1013, [x=262,y=12]
10:06:25	CLICK_LEFT	1013, Zoom Panel, [x=407,y=68]
10:06:30	PAN	1013
10:06:54	LEARN	1013, [x=0,y=265]
10:06:58	SEARCH	1013
10:07:39	IMAGE	1013, 6, TM_991106r_14_7, mb_river_01
10:07:39	ORDER	1013, PERCENTAGE_DESC, HIGHEST
10:07:50	TAB_CHANGE	1013, Learn
...	...	...

Table 8.1: Example DBMS entry of user tracing information. After selecting a certain combination of feature models and an image from the initial gallery, the user started to analyze the image data and to give (positive) training samples. Note that each performed action is assigned to action types as summarized in (Tab. 7.5). The time information is used to compute the complexity measurement $\mathcal{C}_2$ and the quantity $\mathcal{C}_1$ is derived from individual actions.

the relevance of their measurements using the Fisher linear discriminant and the leave-one-out test method. The quantities (features) we use to evaluate the quality of user-defined semantic cover-types are given in (Tab. 8.2). Additional features are the sum of the Kullback-Leibler divergence for both models and the maximal divergence of the two models. The first one reflects the overall quality of the stochastic link between image content and semantic cover-type and the second one the ability of at least one model to capture the label. With these two additional measurements, we obtain a set of nine different quantities for each semantic label. However, some of the measures are correlated or may have only limited discrimination power to distinguish weak and strong cover-types. Finally, analyzing the relevance of the observed data results in a Karhunen-Loéve transform expressed as (Duda et al. 2001)

$$\boldsymbol{y} = \boldsymbol{G}\,\boldsymbol{x} \quad . \tag{8.1}$$

In this equation, $\boldsymbol{x}$ denotes an observed d-dimensional vector for a trained semantic label, $\boldsymbol{G}$ a linear transformation matrix and $\boldsymbol{y}$ the representation of the observations in the transformed space (see App. A for details about computing the KLT). In our case, the yielded eigenvalues λ_i are of major interest since they indicate the

label name	model 1	model 2	$D(\mathcal{L}_\nu, \neg\mathcal{L}_\nu)$		Class. acc	User tracing				User Satisfaction			
			model 1	model 2		$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_5$	VG	G	A	U
ht_water_1	spectr	GRF, 30m	1.82	2.17	77.6	2.97	2.77	26.7	35.7		x		
ht_road_2	spectr	GRF, 30m	0.81	1.46	98.3	1.93	2.51	17.8	30.7		x		
ht_sand_1	spectr	GRF, 30m	2.46	1.24	87.4	2.9	2.78	49	46.5		x		
ht_sand_2	spectr	GRF, 30m	2.3	1.55	59.2	2.99	2.8	43.8	44.9		x		
ht_riverbed_2	spectr	GRF, 30m	2.18	1.27	94.7	2.67	2.8	44.6	44	x			
PGM_riverbed_1	spectr	GRF, 30m	2.15	1.68	71.3	2.66	1.73	26.6	34.9			x	
hd_water_2	spectr	GRF 30m	4.46	4.25	87.5	2.83	2.55	38.1	50.9	x			
hd_cloud_6	spectr	GRF, 30m	4.45	3.79	98	2.66	2.47	25.5	45.1	x			
md_cloud_1	spectr	GRF, 30m	0.83	0.8	80	2.73	0.69	23.3	22.4			x	
md_sea_1	spectr	GRF, 30m	1.99	1.88	100	2.73	1.32	11	46.2			x	
hd_riverbank_1	spectr	spectr	4.31	4.31	98	2.5	2.26	39.03	46.94	x			
ap_water_3	spectr	EMBD, 60m	1.1	0.63	95	2.45	2.09	15.02	24.94			x	
ac_cloudshadow_1	EMBD, 60m	tex, 30m	0.32	2.31	64	2.53	0.8	16.5	12.96			x	
ap_river_3	EMBD, 60m	EMBD, 60m	0.71	0.71	33	2.7	2.48	20.57	25.93				x

Table 8.2: Evaluation protocol with semantic labels and associated objective and subjective measurements. For each defined label we computed the KL-distance $D(\mathcal{L}_\nu, \neg\mathcal{L}_\nu)$ separately for each signal model, the classification accuracy based on training samples and several complexity measurements $\mathcal{C}_i$ from the user tracing parameters. Additionally, the user validated each semantic label with one of the quantized satisfaction categories very good (VG), good (G), acceptable (A) and unsatisfying (U).

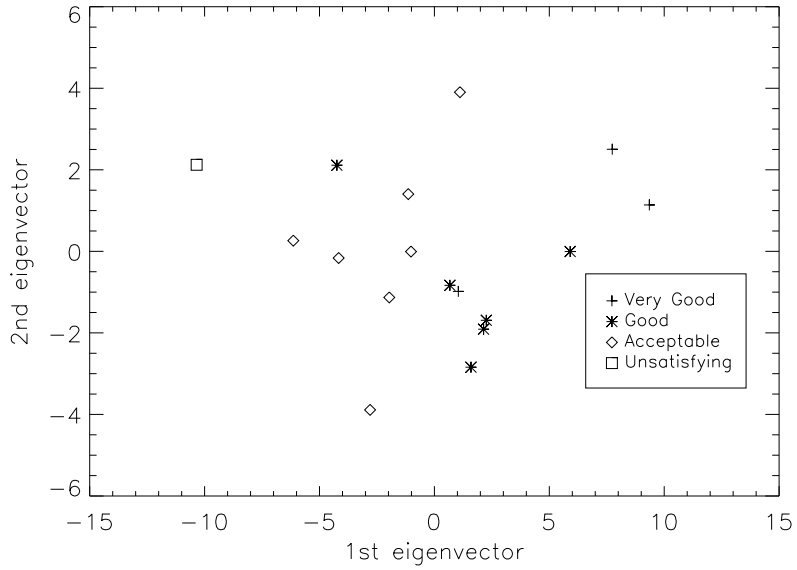


Figure 8.2: Defined and classified labels that were trained with spectral and textural image parameters. The labels can be separated in terms of user satisfaction using only the first two components of the Karhunen-Loéve transform. These two components contain about 80% of the total variance of the original measurements.

correlation of the full dataset and the amount of information in the first k principal components. Usually, there are only few large eigenvalues with the consequence that k is the inherent dimensionality of the subspace governing the 'measurements' (signal) while the remaining $d - k$ dimensions generally include noise. Thus, the Karhunen-Loéve transform can be interpreted as a projection of the data into a k-dimensional subspace. As depicted in (Fig. 8.2), we computed the KLT for a set of 17 semantic cover-types and plotted the first 2 principal components. Each label was marked as 'very good', 'good', 'acceptable' or 'unsatisfying' according to the user's degree of satisfaction. In this two-dimensional space, semantic cover-types group themselves by means of training quality (user satisfaction). Consequently, our measurements can be used to guide the user during the learning process to indicate relevance feedback.

During the evaluation period 31 labels for Landsat TM data, 7 labels for for ERS1 data and 10 labels multi-mission (ERS-1 and Landsat TM) were defined. The analysis of the objective and subjective criteria according to the described procedure resulted in the label validation of 10% as 'very good', 60% as 'good', 20% as 'acceptable' and 10% as 'unsatisfying'.

Remaining problems

The outlined evaluation procedure delivered good results for almost all defined semantic cover-type labels in the DBMS. However, there are situations in which discrepancies between the different measurements occurred, e.g. a user gives a high number of arbitrary training samples or he tries to discriminate image structures that are not reflected by primitive features. Although the system operation complexity increases, there is no learning progress and the stochastic link between semantic labels and content-index is weak. A solution to such a 'random' training could be both the separate analysis of Kullback-Leibler divergence and system operation measurements to exclude weak-defined labels from the semantic inventory. Altogether, the applied evaluation procedure performed well for the majority of labels and demonstrated its usefulness due to a low computational complexity.

8.3 Conclusions

In this chapter, we have discussed the following items:

- We demonstrated the organization of a system evaluation architecture and the tools designed therefore. The link of the user-tracing information to the tools for statistical analysis and the computation of a set of evaluation measurements based on human-computer interactions for each inserted semantic label are the core of the overall system evaluation. Since the computational complexity of the proposed measures is rather low, they can be computed on-the-fly.

- In order to obtain the relevance of the evaluation tools we performed an overall system effectiveness test and included several persons from different application fields. After computing for a number of semantic cover-types objective evaluation measures and comparing them with the subjective degree of user satisfaction, we performed a Karhunen-Loéve transform to analyze the outcomes. The obtained results reflect the performance of the implemented evaluation functions.

9

Executive Summary of Evaluation Results

In the preceding chapters we described a methodology to determine the objective technical quality of the image information mining system, we identified subjective user-related concepts and compared both components in the evaluation protocol. Unlike most commonly used techniques for assessing the performance of a content-based image retrieval system, our approach aims at analyzing individual system modules and their interrelations. Since the system operator is an inherent part of the retrieval loop, we included psychophysical aspects in the evaluation, too. With the proposed system evaluation architecture we arrived at a set of objective and subjective measurements for each defined semantic cover-type. Based on the obtained measurements, we aimed at analyzing their relevance to represent the overall system performance and finding out how strong they are correlated.

In this chapter, we outline the results of an extensive one-week system test. The preparation and organization of this evaluation was not an easy task as it had to be based on the following requirements: the careful selection and processing of appropriate datasets, the association of tasks with different degrees of difficulty to the evaluators, and the fast reporting of the outcomes to give users a feedback about performed actions. On the other hand, the clear partitioning of off-line data processing and on-line system operation enabled us to evaluate the off-line data ingestion chain and image archive complexity a priori. Thus, we could focus on the interactions between users and system during the evaluation week.

In Sec. 9.1, we give an overview of the ingested datasets in the mining system, describe their characteristics and their relevance for the evaluation procedure, and the type of implemented mining function. Then, in Sec. 9.2, we point out the time requirements for off-line data ingestion — feature extraction, clustering and index generation. After having outlined the requirements for off-line data ingestion, we focus on the image archive complexity in Sec. 9.3 and on how this quantity is influenced due to a certain subsampling factor. Whereas Sec. 9.2 and 9.3 describe the objective performance of the system, the last two sections involve objective and subjective evaluation concepts. In Sec. 9.4, the efficiency of the graphical human-machine in-

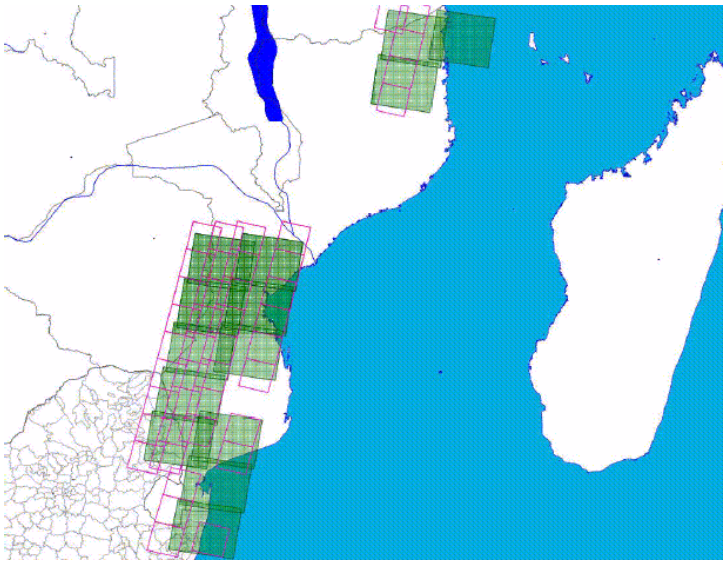


Figure 9.1: Coverage of the Mozambique test site by multi-mission data. The covered surface is larger than 400km × 800km. ERS1 scenes are displayed by small red quadrangles and Landsat TM scenes by larger filled quadrangles.

terface is analyzed in terms of interactive learning and probabilistic search, and in Sec. 9.5 an overall evaluation of the semantic image content is performed.

9.1 Overview of Inserted Datasets

For system verification, we inserted several datasets in the mining system as summarized in (Tab. 5.1). The main set consists of multi-mission — optical and SAR — data of two sites of Mozambique as depicted in (Fig. 9.1).

The complexity of Landsat TM images (resolution 30m, georeferenced, IR channel not used) covering Mozambique is rather high, both from the point of view of image content and subjective understanding by users. The images indicate a huge diversity of spectral signatures and a very broad variety of structural information at different scales. Most of the image structures are natural and have intricated shapes and textures. For many users, the visual understanding of the scenes is not easy because of different geological, climatic, cultural and technological environments. Owing to these reasons, the selection of these data was a challenge and a typical task for information mining.

As opposed to optical image data, the information content of ERS1 SAR images (resolution 30m, georeferenced, GTC processing level) of Mozambique is quite small. The low diversity of several of the image contents is due to various factors, among them are: the look angle of ∼23°, the C-band response for humid land cover vegetation and the type of materials used for man-made structures. Due to the SAR

sensitivity to the surface geometry, however, large scale structures like rivers and geomorphology are well visible in the images.

Besides Landsat TM and ERS1 scenes, we also used nine Ikonos scenes (resolution 1m, georeferenced, radiometrically corrected) covering urban terrain of Mozambique. The Ikonos images indicate a broad diversity of both natural and man-made structures. This dataset is challenging since it allows us to verify our algorithms in terms of primitive feature extraction at high resolution.

Additionally to the Mozambique data, we inserted Landsat TM images covering Switzerland at a surface larger than 400km×400km. The images indicate a quite high information content that is characteristic of alpine areas.

Particularly for the evaluation of the classification performance of the mining system, one Landsat ETM+ scene of Nepal covering an area of about 100km×100km was used. Next to the multispectral information we included the panchromatic data, too.

These datasets were completed by Ikonos scenes covering Germany and Daedalus ATM and E-SAR images covering the Kosovo.

Type of implemented mining function

In order to build a system that is free of application-specificity and manages large amounts of remote sensing image data, several requirements, such as the automatic extraction of relevant features, feature reduction and compression, transfer via Internet and usability have to be fulfilled. With the new generation of high resolution optical and radar sensors, an efficient data management is of the utmost importance. To eliminate these limitations, an information retrieval system has to be scalable to different levels of detail (Seidel and Datcu 1999):

- **Content-based image retrieval:** In this mode, the compression factor is very high ($\sim$ 100) and therefore allows the search of very large quantities of data. However, the search accuracy is restricted and proportional to the amount of information for on-line search. Content-based image retrieval systems mostly work with global image features.

- **Information mining:** A moderate compression factor is used and, consequently, a more detailed search of information for on-line mining is possible. However, the requirements for storage devices increase.

- **Scene understanding:** In this mode, the entire image content is used. Thus, it allows an accurate exploration and interpretation of physical image properties at pixel level.

In (Tab. 9.1), where we outline the implemented mining functions, we achieved compression factors for ingested datasets.

Site	Sensor	No. of scenes	Archive	Mining	Compr.	Function
Switzerland	Landsat TM	5	4,4 GB	52 MB	88	**CBIR**
Mozambique	Landsat TM	14	10,5 GB	616 MB	17	**Information mining**
	ERS-1	30	3.5 GB	808 MB	4,3	
	Ikonos	2	560 MB	124 MB	4,5	**Scene understanding**
Nepal	Landsat TM	1	216 MB	254 MB	1	
	Landsat TM	1	120 MB			
Total	all	53	20 GB	1,8 GB	12	

Table 9.1: Datasets inserted in the I^{2}M system: archive size, size of condensed information for 'mining', achieved overall compression factors and the implemented scalability functions CBIR, information mining and scene understanding.

9.2 Efficiency of Data Ingestion Chain

A further operational application of the image information mining system, like a pipeline in a ground segment system or the synchronization with periodic update (refresh) of data in large robotic archives, requires knowledge about the computational demands for off-line data processing. The extraction of primitive image features and both their reduction and compression by an unsupervised clustering is by far the most time-consuming part in the I^{2}M data processing chain. That's why we will first focus on these system modules.

Primitive feature extraction

As outlined in Sec. 5.1, we extract spectral and spatial parameters from optical images. The latter are extracted at multiple scales based on Gibbs random field (GRF) texture models to achieve a quasi-complete description of the image content. For radar data we apply an enhanced model-based despeckle filter (EMBD) to obtain a cleaned intensity image and textural information related to the Gauss-Markov random field (GMRF) model.

In (Tab. 9.2), we summarize the computation-time for primitive feature extraction for different sizes of the estimation window, model order (neighbourhood size), number of estimation points, overall computation-time and the time needed per estimation point. The calculation of spatial features from optical images (GRF) with a typical number of $466 \times 466 = 217.156$ estimation points with model order 3 and an estimation window of size 35×35 requires about 52 min. on a 500MHz SUN workstation, for instance. It is obvious that the computational demands per estimation point depend on the selected model order and estimation window (processing complexity). Although we applied a fast Linux PC for extracting the radar features, the processing was a computationally intensive task. For the chosen processing parameters (GRF: 35, 3 and GMRF+EMBD: 17, 5) with different estimation steps, the extraction of radar image attributes takes about twice as much time as the extrac-

Feature	Est. win.	Model order	Time (sec.)	Est. points	$\frac{\text{Time}}{\text{Est. point}}$ (msec.)
GRF	35 × 35	3	3173	466×466	14.6
GRF	20 × 20	3	839	481×481	3.6
GRF	50 × 50	3	6597	451×451	32.4
GRF	35 × 35	1	1266	466×466	5.8
GRF	35 × 35	5	6440	466×466	29.7
GRF	17 × 17	5	141	162×162	5.4
GMRF + EMBD	11 × 11	3	2783	166×166	101
GMRF + EMBD	11 × 11	5	4174	166×166	151
GMRF + EMBD	11 × 11	7	6538	166×166	237
GMRF + EMBD	17 × 11	5	8125	166×166	295
GMRF + EMBD	5 × 5	5	1347	166×166	48.9

Table 9.2: Summary of computation time for primitive feature extraction. For the determination of textural parameters (GRF) from optical imagery we applied a 500MHz Sun workstation and for radar features (EMBD and GMRF) a 800MHz Linux PC.

tion of texture features from optical data. The computation of Landsat TM image texture attributes (at a certain scale) for the Mozambique site (438 tiles) required about 48h using a 6 CPU SUN cluster and for the same amount of ERS1 image tiles about 42h using a 8 CPU Linux machine, for example. Notice that spectral features were not included in this validation since they can be directly obtained from raw image data (after a normalization).

Unsupervised clustering

Having extracted primitive image attributes, the samples in a multi-dimensional feature space are first normalized according to feature space dimensions and then clustered in a certain number of characteristic groups. As we see in (Tab. 9.3), the normalization of each feature dimension to form a uniform space is not a limiting factor of the data ingestion chain. As the measurements indicate, the more complicated task is the clustering process. Like for all clustering algorithms, the processing time depends on several parameters, such as the number of clusters and iterations, for example. We performed the global grouping across all images with typical parameters as 128 clusters and 30 iterations. As we can further see from the measurements, the computation time depends on the selected normalization method and the dimension of the feature space. The latter is evident since the algorithmic complexity of dyadic k-means is proportional to the number of dimensions.

Feature extraction and clustering is generally a complicated and time-intensive task. But because of the latest technologies, which make clusters of hundreds of CPU — each of which with more than 2GHz speed — available at low price, feature extraction is not a difficult task anymore. Additionally, the processing can be

Feature	scale	dim.	Norm. method	Norm. (s)	'0' clust.	Clust. (h)	$\frac{\text{Clust.}}{\text{tile}}$ (min.)
Spectral	30m	6	Gaussian	178.1	0	27.1	3.7
Spectral	30m	6	linear	65.9	16	26.4	3.6
GRF	30m	4	Gaussian	141.9	3	12.0	1.6
GRF	60m	4	Gaussian	121.9	7	12.0	1.7
GRF	120m	4	Gaussian	131.2	1	12.0	1.6
EMBD	60m	1	Gaussian	32.6	15	3.1	0.4
EMBD	120m	1	Gaussian	36.6	15	2.9	0.4
GMRF	60m	1	Gaussian	35.7	15	2.4	0.3
GMRF	120m	1	Gaussian	34.7	15	2.0	0.3

Table 9.3: Time requirements for feature normalization and clustering. All measurements are related to the processing of global feature spaces, each consisting of about 100 Mio. data samples from 438 Landsat TM and ERS1 tiles. '0' clust. indicates the number of clusters to which no data samples are assigned.

optimally adapted to an application due to a selected subsampling factor. The algorithms for feature extraction and clustering do not require parallelization; datasets are distributed in single instruction multiple data (SIMD) strategy.

Catalogue entry generation

After analyzing the time requirements for feature extraction and unsupervised clustering, we complete the verification of the off-line data ingestion chain by catalogue entry generation. As stated in Sec. 5.2, from the clustering results we compute — for each image tile — as many classification maps as different features models were applied. From these maps we derive the posterior probabilities $p(\omega_i|I_\zeta)$ as the frequency of signal classes ω_i given a certain image I_ζ. These probabilities can be easily derived from the normalized histogram of each classification map. In addition to classification maps and probabilities, the catalogue entry that is ingested in the DBMS includes quick looks and thumbnails, both in JPEG format. Most of the time in the off-line processing chain is required for feature extraction and unsupervised clustering. The other parts are less computationally intensive; only the generation of quick looks and thumbnails is time-demanding because of a performed Gaussian color histogram modification. All in all, even if the data ingestion chain seems to be quite time-intensive, it can be adapted and scaled to an existing hardware configuration without losing many information details as will be shown in Sec. 9.3.

Condensing information for 'mining'

Having identified the computational demands of the algorithms for primitive feature extraction, clustering and catalogue entry generation, we will now analyze the compression performance of the system. Unlike (Tab. 9.1), where we demonstrated

	Archive size		I²M size			
	# tiles	Size of tiles	Visual tiles	Thumbnails	Size of 'classfiles'	# models
I.	438	2000×2000	1864×1864	125×125	466×466, byte	4
	6 bands, byte		3 bands, c=23	3 bands, c=15		
	Archive size: 10.5 GB		232 MB (36.9%)	3.5 MB (0.6%)	392 MB (62.5%)	
II.	438	2000×2000	1864×1864	125×125	466×466, byte	4
	1 band, integer		1 band, c=9	1 bands, c=6		
	Archive size: 3.5 GB		413 MB (51.1%)	2.8 MB (0.4%)	392 MB (48.5%)	

Table 9.4: I²M data compression performance for the Mozambique Landsat TM (I.) and ERS1 (II.) datasets. The share of visual tiles, thumbnails and classfiles in the entire 'mining' size are outlined in brackets; c denotes the compression factor.

the implemented scalability functions for various datasets, we will now point out the compression factors for quick looks, thumbnails and classification maps (Tab. 9.4). For Landsat TM visual tiles (RGB color quick looks) a rather high compression factor could be achieved, c=23, whereas this value is much lower for ERS1, c=9. Differences in the compression rates resulted in storage requirements of 232 MB for the Landsat TM and of 413 MB for the ERS1 archive. Recent experiments proved that the overall 'mining' size can be further reduced by compressing the unsupervised classification maps. With publicly available tools, these maps can be compressed by factor 3 to 4.

9.3 Image Archive Complexity

In the algorithmic protocol of Chap. 7, we demonstrated a method that allows us to determine the complexity of an image in the entire archive. The approach is based on the Kullback-Leibler divergence $D(p_i; q_i)$ between two probability distributions $p_i = p(\omega_i)$ and $q_i = p(\omega_i|I_\zeta)$ indicating the prior probability of a certain class ω_i in the global class space and the posterior probability of ω_i conditioned on a particular image I_ζ. Divergence can be used to determine the diversity of image structures in a single image in relation to the whole database.

In (Fig. 9.2), we depict the KL divergence for optical and radar feature models computed for the multi-mission dataset. As we see, for optical Landsat TM data the highest complexity values are given for the texture model at a scale of 30m and decrease with increasing scale. The spectral feature model shows minimum complexity for this dataset. For signal classes computed from radar data, the maximum divergence appears for texture at a scale of 60m and the minimum divergence for the filtered intensity at a scale of 120m. Unlike optical data, the difference between the complexity of radar feature models is rather low.

In addition to Kullback-Leibler divergence, we regarded image and index space as the input and the output of a communication channel view and computed the mutual

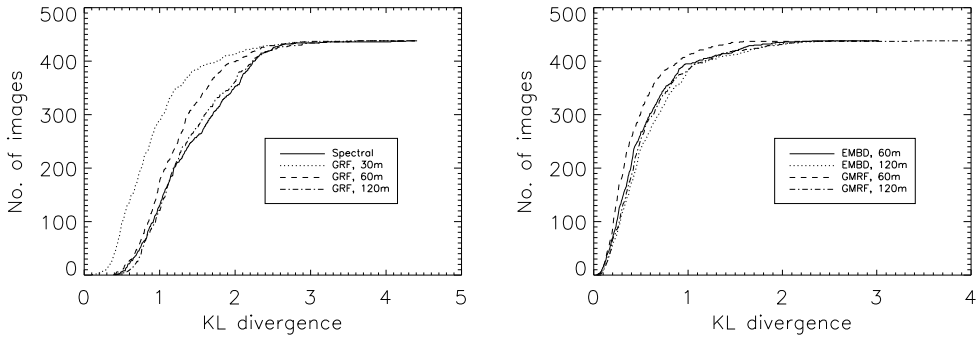


Figure 9.2: Agglomerated histogram of the complexity of images in the database. Kullback-Leibler divergences are computed for feature models of optical (left) and radar (right) images of the Mozambique multi-mission dataset.

information $I(\boldsymbol{I};\boldsymbol{\omega})$ between them. This quantity reflects how much information the characteristic set of global signal classes contain about images in the database. The mutual information computed for optical and radar feature models is illustrated in (Tab. 9.5). The obtained quantities correspond to the KL divergences as depicted in (Fig. 9.2). Whereas Kullback-Leibler divergences indicate the complexity of certain images in the archive, mutual information relates to the association between all images and signal classes.

Subsampling factor

A quantity that is of interest in terms of image archive complexity is the loss of information due to a certain subsampling factor S. In the current version of the system, this factor can be selected in the data ingestion chain to achieve a certain

Sensor	Signal models	scale	I($\boldsymbol{I}$; $\boldsymbol{\omega}$)
Landsat TM	spectral	30m	1,41
Landsat TM	GRF	30m	0,92
Landsat TM	GRF	60m	1,23
Landsat TM	GRF	120m	1.39
ERS1	MBD	60m	0,53
ERS1	MBD	120m	0,59
ERS1	GMRF	60m	0,43
ERS1	GMRF	120m	0,56

Table 9.5: Mutual information $I(\boldsymbol{I};\boldsymbol{\omega})$ between image space and class (content-index) space. The class space $\boldsymbol{\omega}$ was separately generated for 438 co-registered Landsat TM and ERS1 images based on different signal models.

Sensor	Model	Scale	Subsampling factor S						
			4^2	8^2	16^2	32^2	64^2	128^2	256^2
Landsat TM	Spectr.	30m	1.409	1.409	1.411	1.419	1.444	1.539	1.788
	GRF	30m	0.921	0.921	0.923	0.931	0.961	1.088	1.434
	GRF	60m	1.233	1.233	1.236	1.241	1.262	1.369	1.669
	GRF	120m	1.389	1.389	1.391	1.394	1.412	1.503	1.797
ERS1	EMBD	60m	0.527	0.527	0.531	0.543	0.591	0.787	1.328
	EMBD	120m	0.594	0.594	0.598	0.609	0.656	0.844	1.365
	GMRF	60m	0.430	0.430	0.433	0.444	0.486	0.645	1.093
	GMRF	120m	0.563	0.563	0.564	0.574	0.613	0.771	1.204

Table 9.6: Mutual Information $I(\boldsymbol{I};\boldsymbol{\omega})$ for the Mozambique dataset and different subsampling rates. For a subsampling rate of up to 16^2, only small changes for all feature models are visible.

data compression and to optimally adapt to the users' application. However, with the choice of factor S a loss of information has to be considered. With the help of information theory, we can measure this quantity as outlined in (Tab. 9.6). We computed the mutual information between image and signal class space for various feature models of the Mozambique dataset, this time in dependence on S. For an increasing data reduction, mutual information increases. However, the loss of information details is negligible to a certain extent. For the Mozambique dataset we could apply a subsampling rate of $S = \frac{1}{16^2}$ without risking to 'lose' image details in the archive.

9.4 Human-Machine Interface

The I^{2}M system is based on human-centered concepts in order to fully exploit the synergy of human and computer: the user guides the interactive learning process and the system continuously gives the operator relevance feedback about the performed training actions and searches the archive for relevant images. The implemented mining functions — training, image content interpretation and probabilistic search — require interactive operation in "real time" relative to the reaction of the user. Thus, we compare the following aspects for the evaluation: the average number of training actions per image, the interactive learning applet loading time, the duration of an interactive learning session and, additionally, the time requirements for the probabilistic search and presentation of the queried top-ranked images (Fig. 9.3). The average number of training samples per image is about 4 in relation to the complete number of 7 to 8 training samples for cover-type labels. These two quantities reflect the capability of I^{2}M to define semantic labels with just a few training samples on a small number of images. The loading average time of the applet is

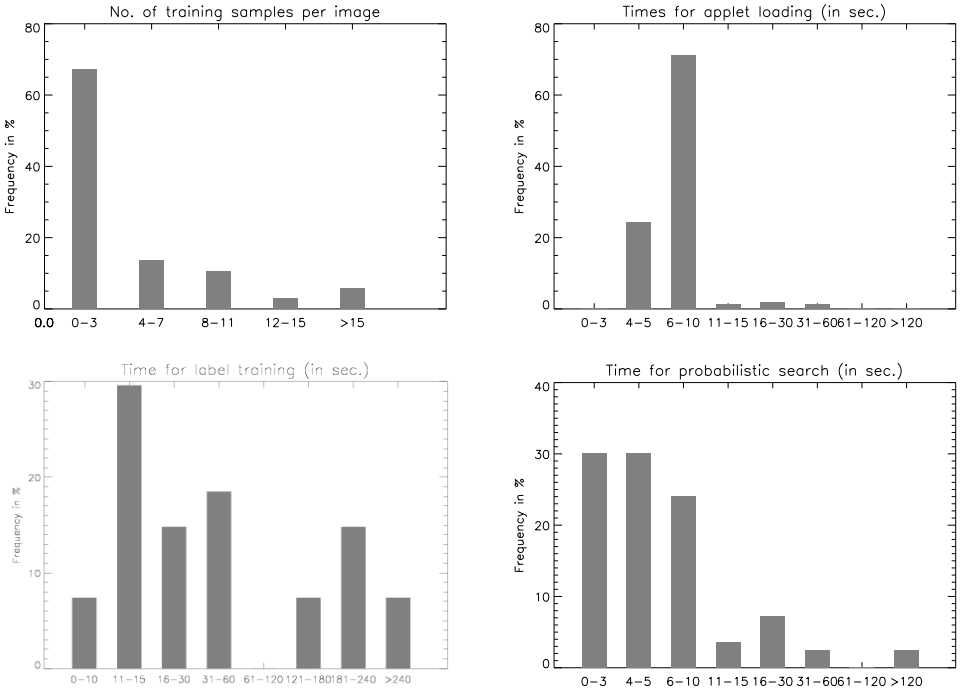


Figure 9.3: Evaluation of the image information mining system on-line graphical user interface. Plots from left to right: 1st row: number of training samples per image for training a semantic cover-type and the time required for loading the interactive learning applet. 2nd row: time required for training user-specific semantic labels and time for performing the probabilistic search and thumbnail gallery presentation.

about 8 seconds, relative to the average duration of the interactive learning session of about 70 seconds. The results are satisfactory and the ratio is maintained also in the over-the-net operation for normal network speed. The requirements for performing the probabilistic search depend on the number of selected signal models and the database size reflected by a large variance. Considering the fact that both probabilistic search and data transfer via Internet can be accelerated from an algorithmic and technical point of view, the obtained average time of 7 to 8 seconds is rather promising.

During the one-week evaluation test phase we counted 3 errors caused by the I^2M database management system and by local network problems. Although the number of operational faults is not equal to zero (as one expects from an operational retrieval system), the obtained results are rather promising due to the fact that the system was tested at the same time by up to four users between six to eight hours a day.

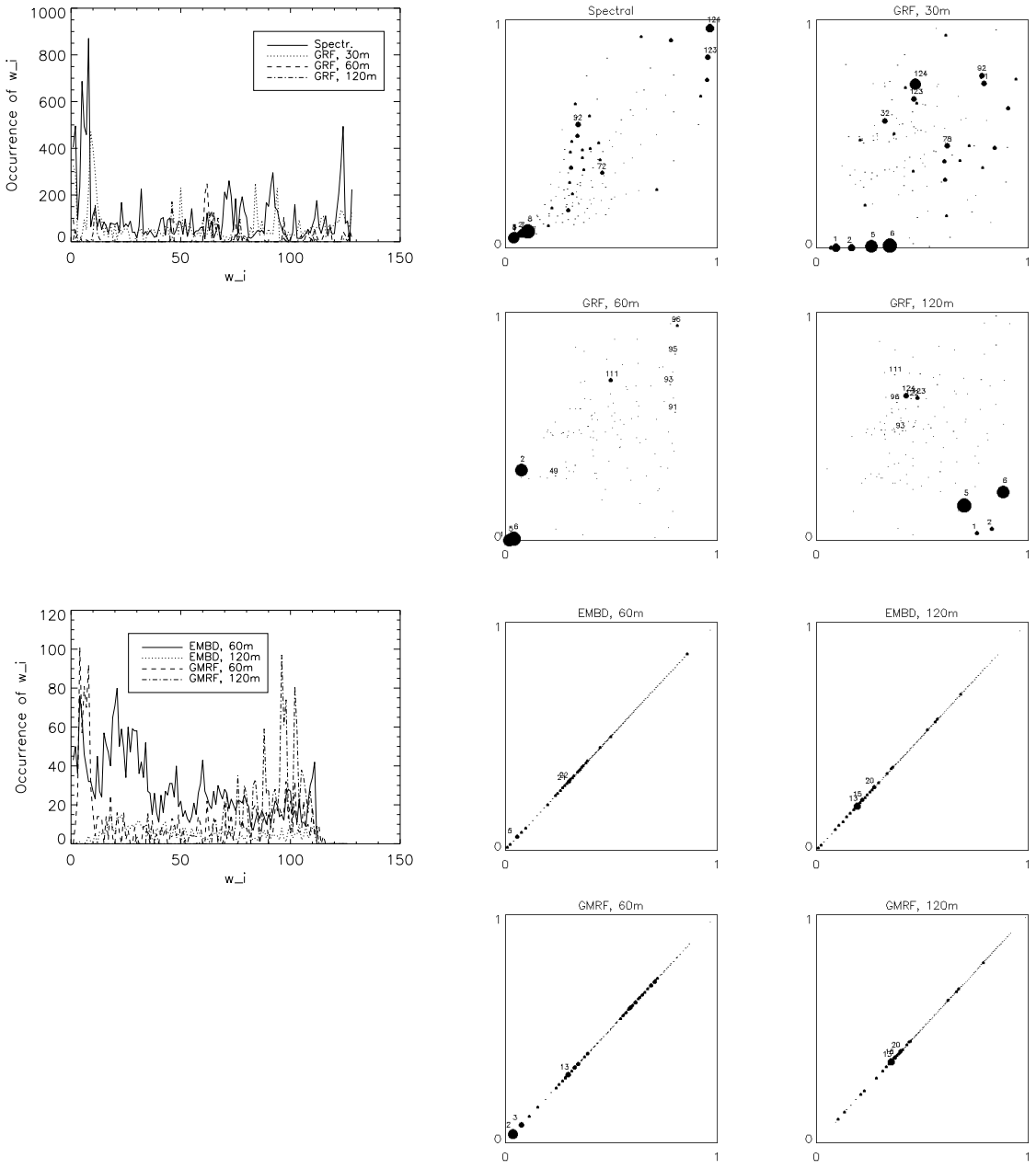


Figure 9.4: Occupation of different feature spaces of the Mozambique Landsat TM (1st. and 2nd. row) and ERS1 (3rd. and 4th. row) dataset by semantic cover-type labels. The most occupied clusters ω_i are attributed by the corresponding cluster number i.

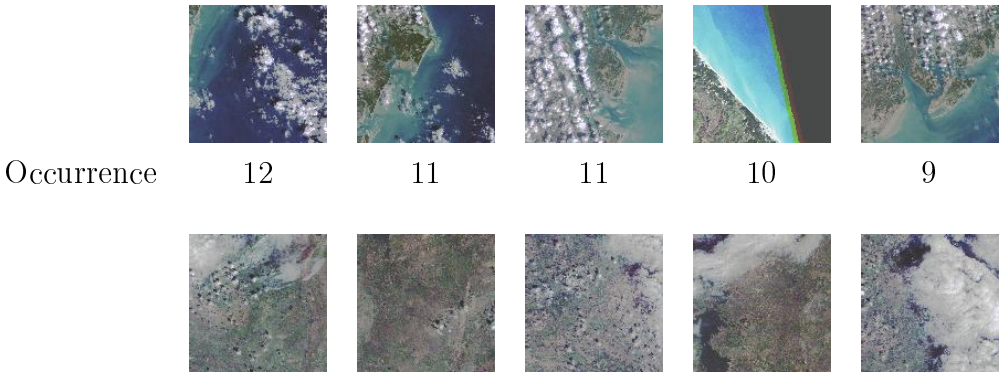


Figure 9.5: Most frequently queried images (top row) and images that never appeared in the resulting gallery (lower row).

9.5 Semantic Image Content

After having outlined the effectiveness of the off-line data processing chain and on-line system operation, we will focus on the semantic image content evaluation in this section. During the evaluation period, 31 labels for Landsat TM data, 7 labels for ERS1 data and 10 labels for the multi-mission dataset were defined. First, we identify the performance of the generated content-index to represent semantic cover-type labels. Unlike in previous chapters, where we outlined the quality of semantic labels in terms of stochastic link, classification accuracy and probabilistic search results, we now analyze the occupation of clusters by semantic cover-types as depicted in (Fig. 9.4). Note that the dimension of the different feature spaces varies from 1 for radar (EMBD and GMRF) to 4 (GRF) and 6 (spectral), respectively.

The 31 labels defined for the site of Mozambique are assigned to ∼75% of the images in the archive. Images that were most often queried and displayed in the resulting gallery are depicted in (Fig. 9.5) together with images that had never appeared in the top-ranked images. The analysis of objective and subjective criteria resulted in the evaluation of the defined labels as follows: 10% very good, 60% good, 20% acceptable and 10% not satisfying. However, for the overall evaluation one should consider the very high complexity of the Mozambique and Nepal images as described in Sec. 9.1.

9.6 Conclusions

In this chapter, we have discussed the following items:

- An overview of datasets ingested in the image information mining system was presented. In addition to the properties of the different volumes of data we showed the implemented mining functions and the achieved compression

factors. These characteristics are relevant for the level of detail the data can be searched for.

- The efficiency of the off-line data ingestion chain was analyzed. Although the extraction of primitive image features and their compression by clustering are time-consuming operations, they can be managed in an adequate period with existing computer technology.

- The complexity of the different datasets and how this quantity is affected by an increasing subsampling factor were presented. The main outcomes are that complexity measures depend on the applied signal models and a moderate subsampling rate does not significantly impair the complexity.

- During the evaluation test week we analyzed the performance of the graphical man-machine interface. The provided mining functions for interactive learning and probabilistic retrieval allow interaction in real time relative to the reaction of the user. Since the data transfer via Internet can be accelerated, the evaluation results are quite promising.

- In order to verify the semantic image content, we analyzed objective and subjective evaluation measures and the filling of the different feature space clusters by cover-type labels in the database inventory.

10

Conclusions

This dissertation concentrates on the enhancement and evaluation of a content-based remote sensing image information mining system. Content-based image retrieval has originally been developed to query for pictures with a particular content from large multimedia databases. Typical queries are "show me all images similar to this one" or "show me all images that contain cars", for example. The implemented methods and retrieval algorithms perform well for this kind of data, but turned out to be of limited use for remote sensing applications. The state-of-the-art systems for accessing remote sensing images allow only query by meta-information such as geographical coordinates, time of acquisition and sensor type. However, this information is often less relevant for the user than the actual content of the scene, e.g. structures, objects or scattering properties. In addition to the operational archives and database systems, we have implemented and evaluated a knowledge-driven information mining system. The system supports the human-machine interaction via Internet and adaptively incorporates application-specific interests by linking the user-defined semantic image content interpretation with Bayesian networks to a completely unsupervised content-index. Based on the stochastic link, the user can query the archive for relevant images and obtains a probabilistic classification of the entire archive as an intuitive information representation.

10.1 Summary of the Dissertation

The implemented mining system aims at providing both novice and experienced users with direct access to the content of remotely sensed images and therefore belongs to content-based image retrieval. Consequently, we started this thesis with a review of basic methods and algorithms applied in content-based image retrieval to search images in large archives according to their visual properties. We discussed how the visual content of images can be described by attributes like color, texture and shape and how these features can be reduced and indexed to obtain computationally manageable data quantities. Since the information content of a single image feature is not sufficient to distinguish the structures and objects of all images in the database, systems were equipped with methods to fuse various sources

of information at the level of features, content-index and semantic image content interpretation. Not only the fusion of information, but also the implementation of relevance feedback techniques resulted in functions that return images similar to the user's conjecture. As in other scientific fields, the further development and enhancement of content-based image retrieval depends on the capability to evaluate and compare the image understanding and mining functions — a subject which has not been dealt with much so far yet.

The information mining system that is in focus throughout this thesis models the image content from raw data pixel values up to user-related semantic interpretation in a stochastic Bayesian way using a hierarchy of levels with different semantic abstractions. Thus, we outlined the definition of probability, how information about unknown parameters can be inferred from observations in the Bayesian way and how the information content and accuracy of estimated parameters can be measured. Probability, Bayesian inference and information theory constitute the basic tools that are applied throughout this thesis.

Based on the above findings, we illustrated the implemented scheme of hierarchical Bayesian image content modeling in the next part of the thesis. We explained how the image characteristics at a certain level in the hierarchical representation are obtained from the elements of level(s) below in a step of Bayesian inference: primitive image features and meta-features are estimated from image data in a Bayesian way by applying parametric signal models, clouds of primitive features in the different feature spaces are replaced by a more compact representation using parametric models of the unsupervised clusters, user-specific interpretation of image content in form of semantic cover-types is linked to the unsupervised clusters with simple Bayesian networks, and the same Bayesian learning model is applied to aggregate cover-type labels to complex semantics. Then, we implemented the theoretical concept of hierarchically modeling the image content at multiple levels on remote sensing data. We described how primitive image features are extracted from optical and radar scenes, how an unsupervised content-index is generated using a dyadic k-means clustering algorithm, how user-specific semantic cover-types are linked to the index and how their definition can be used to search the archive in a probabilistic way.

Both for optical, radar and multi-mission data the signal-oriented way of image content modeling and retrieval was limited because complex semantic labels cannot be discriminated on the basis of the generated content-index. Consequently, a new level of image content abstraction was introduced: semantic grouping. Just like a cover-type label is defined through various man-machine interactions and linked to characteristic signal classes of the content-index, a user can group labels to higher-level semantics by weighting individual labels. A limitation of the traditional scheme of unsupervised image content description and supervised semantic labeling is that only images that are assigned to clusters in the global feature spaces of the different feature models can be queried. Images from other (non multi-mission) collections

cannot be included in the retrieval loop by using the standard way of image content modeling. We showed that the new level of aggregated semantics can be extended to search for images across sensors and image collections. However, as cover-types are defined by users with different background knowledge about sensor, feature models and scene classes and objects, the users have to share a certain ontology to avoid semantic ambiguities in the DBMS inventory.

After having described the basic theoretic concepts and the application of the image information mining system, we further dealt with its evaluation and verification. Since no single quantity is able to represent the overall system performance, we outlined an algorithmic protocol with functions specially designed for the evaluation of each system module. This protocol is organized in the same hierarchical way as previously shown for image content modeling. We started the evaluation presenting techniques to determine the information content of extracted primitive image parameters. Therefore, we did not determine the features' performance in terms of retrieval quality. Instead, we directly measured the information content of the features. For spectral features we obtained the distortion by noise and for spatial features at multiple scales we calculated the accuracy of the estimates using the Cramér-Rao lower bound. The information content of the unsupervised feature clusters is mainly reflected by measurements like isolation and compactness. For their determination we computed scatter matrices, divergence, Bayes' probability of error and compared the location and shape of clusters with the feature space density. In addition to analyzing clusters in multi-dimensional feature spaces, we determined the accuracy of two-dimensional unsupervised classification maps by using error matrices. To analyze the classification accuracy and selectivity of the system according to a trained user-specific semantic label, we measured the quality of the stochastic link between unsupervised content-index and semantic cover-type, and the separation between semantic labels for a given combination of feature models. Then, we evaluated the system retrieval function using standard measurements like precision/recall and the coverage of the entire archive by the defined semantic image content. While the validation of primitive features and unsupervised clusters focused on objective criteria and measurements, the validation of interactive learning and probabilistic search included objective and subjective concepts. This combination of objective and subjective evaluation aspects was even more relevant for the verification of system operation and human-computer interaction. We analyzed the complexity of actions between user and system, the timing of actions and the combination of various action types to form certain action classes in an information-theoretic way. Additionally, we evaluated the graphical man-machine interface, the capability of users to "learn" the system and communication and information representation aspects. The outlined algorithmic protocol did not only concentrate on the verification of individual system modules but also on the association of system modules and their interactions. We illuminated the mining system from the communication channel view and measured the information flow between

the basic levels of different semantic abstraction.

The overall validation of the mining system required a special organization and the development of appropriate tools. For the evaluation of the objective performance of the system, we used a tool to trace the interactions between operator and system and developed functions for the statistical analysis of user-tracing parameters. With these methods we had a number of evaluation measurements at hand that allowed us to compare the relevance of objective measurements with the subjective level of the protocol. The latter was expressed as a questionnaire and included the user's degree of satisfaction. In order to obtain reliable and statistically correct results, various persons from different working fields and with different background knowledge participated in the system evaluation procedure.

In the final chapter dealing with system evaluation we summarized the main evaluation results, gave an overview of processed and inserted datasets, showed the capability of the system to add and scale large data quantities, measured the information-theoretic complexity of different datasets, illustrated the efficiency of the on-line human-machine interface and showed the evaluation of the overall semantic image content.

10.2 Outlook

The enhancement and evaluation of the image information mining system that we addressed in this thesis resulted in the identification of topics that should be the subject of further development and optimization.

System upgrade

First, the applied set of primitive image features has to be extended to include geometrical attributes. While spectral and textural parameters demonstrated their capability to describe the content of remote sensing images at medium resolution, they are of limited importance for data with increasing geometrical resolution, e.g. provided by the Ikonos and the TerraSAR sensor. For a geometrical characterization of the scene content, shape descriptors and elements of topology derived from the outcomes of an unsupervised segmentation are the key issues to be solved. In the current mining system, we ingested and tested geometrical features for both Landsat TM and Ikonos scenes. However, the results did not meet the expectations.

As stated in Sec. 5.2, the reduction and compression of extracted primitive features by an unsupervised clustering assumes that the features obtained from all images in the archive build a global space for different feature models. What is needed to efficiently manage and process large volumes of data (as it is the case in a satellite ground segment system) is an incremental clustering method that allows the unsupervised grouping of extracted features "on-the-fly". A short experiment in that direction demonstrated that the classification of images added to the database

according to the existing vocabulary performed well and did not significantly affect the retrieval performance.

In terms of information fusion for interactive learning and probabilistic search, the current version of the system allows the combination of up to two different signal models. In order to discover and understand high complex scenes, particularly multi-mission datasets, the combination of more than two models is required. In the off-line version of the mining system, we successfully implemented and tested the semantic labeling of cover-type labels with up to eight feature models. What hinders the fast implementation of the probabilistic search is the computational complexity of the search algorithm. A simplification by applying a dyadic scheme on the full set of models showed good results but needs more testing and validation.

Other factors that should be considered to achieve a further system upgrade are visualization techniques for multi-mission and temporal remote sensing datasets, and the interoperability with GIS systems.

System evaluation

In this thesis, we mainly addressed the evaluation of an existing image information mining system. Although we described an extensive algorithmic protocol consisting of methods and tools to perform an overall validation with objective and subjective components, further tests should be performed.

First and foremost, the mining system should be verified in a large-scale test by including huge datasets as they are relevant for satellite ground segments or robotic archiving systems. A question that arises about mining huge volumes of data is how much value can be added to the system, how an adequate compression factor can be selected and how much information details are lost due to a selected compression factor. We demonstrated that — although it takes some time and is computationally intensive — the data ingestion chain using a cluster of 100s of CPUs can manage the extraction and clustering of primitive image features as needed for a pipeline with a SAR processor, for instance. The required data storage, for a scenario using a moderate compression factor of up to 400 that guarantees a good image content preservation according to information-theoretic complexity, enables the on-line exploration of more than 10 Terra Bytes of data. In order to access very large volumes of data, the system can be operated in two steps. First, a classical query using meta-information like geographical coordinates in combination with semantic data grouping restricts the search to a subspace of the large query archive. Second, a 'cluster archive' from the large archive is produced and the interactive learning and probabilistic search is performed for this dataset. Interesting quantities that are relevant for the mining of such large data quantities are the completeness of search results, the time for executing the query and the technical limitations due to disk storage and computing power.

The presented evaluation procedure was characterized by a limited number of

participants and the fact that all of them were familiar with remote sensing data. In order to test the usability of the system and its adaptivity and suitability for various kinds of applications, the system has to be tested by a larger number of evaluators from heterogenous working fields. What should be verified along with testing the mining system by many users is the stability of the DBMS for several thousands of parallel Internet accesses.

A

Karhunen-Loéve Transform

A method to cope with the "curse of dimensionality" in statistical data analysis is to reduce the dimensionality by applying a Karhunen-Loéve transform [1]. This transform projects high-dimensional data (observations) onto a lower dimensional subspace that is optimal in the sum-squared error sense (Fukunaga 1990). The efficiency of the Karhunen-Loéve transform is due to its low computational complexity and analytical tractability.

If we assume a set of data samples in a d-dimensional space with each sample described by its appropriate vector $\boldsymbol{x}$, we can compute the mean vector

$$\boldsymbol{\mu} = E\{\boldsymbol{x}\} \tag{A.1}$$

and covariance matrix

$$\boldsymbol{\Sigma} = E\left\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^t\right\} \tag{A.2}$$

for the full dataset. For the Karhunen-Loéve transform it is fundamental to find a linear transformation $\boldsymbol{G}$ of the original coordinate system in the multi-dimensional space, so that

$$\boldsymbol{y} = \boldsymbol{G}\,\boldsymbol{x} \tag{A.3}$$

represents the samples in the $\boldsymbol{y}$-space and the covariance matrix for $\boldsymbol{y}$ is diagonal. The covariance matrix for data samples in the transformed space can be expressed as

$$\boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{G}\,\boldsymbol{\Sigma}_{\boldsymbol{x}}\,\boldsymbol{G}^t \tag{A.4}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ denoting the covariance matrix in the original space. As $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ is assumed to be a diagonal matrix, $\boldsymbol{G}$ can be considered as the transposed matrix of normalized eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. The eigenvectors $\boldsymbol{g}_i$ of the transformation matrix $\boldsymbol{G}$ are composed of result from

$$(\boldsymbol{\Sigma}_{\boldsymbol{x}} - \lambda_i \boldsymbol{I})\,\boldsymbol{g}_i = \boldsymbol{0} \tag{A.5}$$

[1] The Karhunen-Loéve transform is also called principal component analysis (PCA) or Hotelling transform.

and the corresponding eigenvalues λ_i (elements of the diagonal matrix $\boldsymbol{\Sigma_y}$) have to fulfill

$$|\boldsymbol{\Sigma_x} - \boldsymbol{\lambda I}| = \mathbf{0} \tag{A.6}$$

with identity matrix $\boldsymbol{I}$. Since $\boldsymbol{\Sigma_y}$ is defined as a diagonal covariance matrix, its elements indicate the variances of the samples in the transformed space. The advantage of the Karhunen-Loéve transform is to reduce the original data space to a certain subspace by analyzing the variances.

B

Notation

B.1 Variables

The following table lists all important variables used in this thesis. Owing to the diversity of covered topics, some symbols may have several meanings. However, the correct denotation can be gathered from the context.

A_j^i	user action with consecutive number j and type i
a	auto-interaction in the scalar term η
$\boldsymbol{\alpha}$	vector of hyper-parameter to describe $p(\omega_i\|L_\nu)$ and $p(L_\nu\|\Lambda_\tau)$
B	Bhattacharyya distance
$\varphi(\cdot)$	Parzen window function
α_i	hyper-parameter
C	coverage for a certain label
$\boldsymbol{\phi}$	parameter vector to model $p(\omega_i\|L_\nu)$
ϕ_i	element of $\boldsymbol{\phi}$
$\boldsymbol{\psi}$	parameter vector to model $p(L_\nu\|\Lambda_\tau)$
ψ	element of $\boldsymbol{\psi}$
N_i	occurrence of ω_i in T
N_ν	occurrence of L_ν in T
D	data/observations, main element of level 0
$D(\cdot)$	Kullback-Leibler divergence
D_{kl}	Kullback-Leibler divergence between cluster ω_k and and ω_l
J_{kl}	Jeffries-Matusita distance between cluster ω_k and and ω_l
$\langle D \rangle$	average divergence
E	cost function
$E\{\cdot\}$	expectation value
G	maximum grey value
$H(\cdot)$	Shannon's entropy measure

	chap. 3: hypothesis
	chap. 5: energy function
$I(X,Y)$	mutual information between X and Y
$\boldsymbol{I}(\boldsymbol{\theta})$	Fisher information matrix
I_ζ	a certain image in the database
L	equivalent number of looks
L_ν	user-specific semantic cover-type, main element of level 4
Λ_τ	aggregated semantic label, main element of level 5
M	stochastic parametric model
N, n	Number of observations, data samples or actions
$\mathcal{N}(\mu, \sigma^2)$	Gaussian noise process with mean μ and variance σ^2
$Pr(\cdot)$	probability
$p(\cdot)$	probability, probability distribution or pdf
P_o	probability to over-retrieve images
P_f	probability to forget relevant images
ω_i	signal class (cluster), main element of level 3
Q_o	overall classification accuracy
C_ν	average classification accuracy of cover-type L_ν
θ	scalar model parameter
$\hat{\theta}$	estimated scalar model parameter
$\boldsymbol{\theta}$	model parameter vector, main element of level 1
$\lvert\hat{\boldsymbol{\theta}}\rvert$	norm of estimated model parameter vector
$\boldsymbol{\theta}_j$	parameter vector of a certain point
$\hat{\boldsymbol{\theta}}$	estimated model parameter vector
X	random variable
x	state of random variable X
	chap. 5: noise-free pixel
∂x_s	neighbourhood of x_s
x_s	image pixel at a certain site s
ξ_i	measurement for clustering performance
x_{ij}	neighbouring image pixels around x_s
Y	random variable
y	state of random variable Y
	chap. 5: pixel of a noisy observation
$\sigma^2_{\hat{\theta}}$	conditional variance of an estimated scalar parameter
$S(L_\nu \lvert I_\zeta)$	separability of cover-type L_ν in image I_ζ
$\sigma^2_{\hat{\boldsymbol{\theta}}}$	conditional variance of an estimated parameter vector
$\hat{\sigma}^2_M$	variance of model M to describe data

θ	scalar parameter
θ'	transformed scalar parameter
$\boldsymbol{\mu}_i$	mean vector of cluster ω_i
$\boldsymbol{\Sigma}_i$	co-variance matrix of cluster ω_i
d	dimension of feature space
r	total number of clusters
R	relevant images in retrieval set
$\Gamma(\cdot)$	Gamma function
η	scalar quantity denoting the joint influence by neighbouring pixels
$\boldsymbol{S}_i$	scatter matrix of cluster ω_i
T, T'	training samples provided by the user
	Sec. 5.2: mathematical cluster model
	Sec. 7.4: set of queried images
	Sec. 7.6: time for a complete learning session
V	classification and including parameters
Z_s	Normalization constant of Gibbs distribution

B.2 Acronyms

Throughout this thesis, the following acronyms have been used:

ACS	Advanced Computer Systems
CBIR	content-based image retrieval
CLS	conditional least-squares
CRB	Cramér-Rao bound
DBMS	database management system
DFD	Deutsches Fernerkundungsdatenzentrum (at DLR)
	(German Remote Sensing Data Center)
DLR	Deutsches Zentrum für Luft- und Raumfahrt
	German Aerospace Center
ENL	equivalent number of looks
ERS	European Remote-Sensing Satellite
ESA	European Space Agency
ETHZ	Eidgenössische Technische Hochschule in Zürich
	(Swiss Federal Institute of Technology in Zurich)
GIS	geographical information system
GMRF	Gauss-Markov random field
GRF	Gibbs-Markov random field (auto-binomial model)
I^2M	image information mining

IMF	Institut für Methodik der Fernerkundung (Remote Sensing Technology Institute)
JPEG	Joint Photographic Expert Group
KES	Knowledge Enabled Services
KIM	Knowledge-driven Information Mining
KL	Kullback-Leibler
KLT	Karhunen-Loéve transform
MAP	maximum a posteriori
MIT	Massachusetts Institute of Technology
ML	maximum likelihood
NASA	National Aeronautics and Space Agency
PCA	principal component analysis
pdf	probability density function
SAR	synthetic aperture radar
SIMD	single instruction multiple data
TM	Thematic Mapper

Bibliography

Aksoy, S. and Haralick, R. M. (1998a): Content-based image database retrieval using variances of gray level spatial dependencies, *Proc. of IAPR Int. Workshop on Multimedia Information Analysis and Retrieval*, Hong Kong, 3–19.

Aksoy, S. and Haralick, R. M. (1998b): Textural features for image database retrieval, *Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries, in conjunction with CVPR'98*, Santa Barbara, CA, 45–49.

Aksoy, S. and Haralick, R. M. (2001): Feature normalization and likelihood-based similarity measures for image retrieval, *Pattern Recognition Letters* 22(5): 563–582.

Aksoy, S., Haralick, R. M., Cheikh, F. A. and Gabbouj, M. (2000): A weighted distance approach to relevance feedback, *Proc. IAPR Int. Conference on Pattern Recognition*, 812–815.

Alsabti, K., Ranka, S. and Singh, V. (1998): An efficient k-means clustering algorithm, *Proc. 11th Int. Parallel Processing Symposium (IPPS)*.

Arabie, P., Hubert, J. and De Soete, G. (1996): *Clustering and classification*, World Scientific.

Bach, J. R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R. and Shu, C.-F. (1996): The Virage search engine: an open framework for image management, *Proc. Storage and Retrieval of Image and Video Databases (SPIE)*.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999): *Modern Information Retrieval*, Addison-Wesley.

Barros, J., French, J. and Martin, W. (1995): System for indexing multispectral satellite images for efficient content-based search, *Proc. Storage and Retrieval for Image and Video Databases (SPIE)*, 228–237.

Bayes, T. (1763): An essay towards solving a problem in the doctrine of chances, *Philosophical Trans. of the Royal Society* 53: 370–418.

Benitez, A. B., Beigi, M. and Chang, S.-F. (1998): Using relevance feedback in content-based image metasearch, *IEEE Internet Computing* 2(4): 58–69.

Berman, A. P. and Shapiro, L. G. (1999): A flexible image database system for content-based retrieval, *Computer Vision and Image Understanding* 75(1/2): 175–195.

Bernardo, J. M. and Smith, A. F. M. (2001): *Bayesian theory*, Wiley.

Bernoulli, J. (1713): Ars conjectandi, *Thurnisiorum*.

Bertoia, C. and Ramsay, B. (1998): Sea ice analysis and products: cooperative work at the US and Candadian ice centers, *Proc. of the IEEE Int. Conference on Geoscience and Remote Sensing (IGARSS'98)*.

Blahut, R. E. (1987): *Principles and practice of information theory*, Addison-Wesley.

Bretschneider, T., Cavet, R. and Kao, O. (2002): A retrieval system for remotely sensed imagery, *Proc. of the Int. Geoscience and Remote Sensing Symposium (IGARSS'02)*, 2253–2256.

Carson, C., Belongie, S., Greenspan, H. and Malik, J. (1997): Region-based image querying, *Proc. CVPR '97 Workshop on Content-Based Access of Image and Video Libraries*.

Carson, C., Thomas, M., Belongie, S., Hellerstein, J. M. and Malik, J. (1999): Blobworld: a system for region-based image indexing and retrieval, *in* D. P. Huijsmans and A. W. M. Smeulders (editor), *Visual information and information systems*, Springer.

Casasola, E. (1998): *ProFusion personal assistent: an agent for personalized information filtering on the WWW*, Master's thesis, University of Kansas, Lawrence.

Castelli, V., Bergman, L. D., Kontoyiannis, I., Li, C.-S., Robinson, J. T. and Turek, J. J. (1998): Progressive search and retrieval in large image archives, *IBM Journal of Research and Development* 253–268.

Chandrasekaran, S., Manjunath, B. S., Wang, Y. F., Winkler, J. and Zhang, H. (1997): An eigenspace update algorithm for image analysis, *Computer Vision, Graphics and Image Processing* 59(5): 321–332.

Chang, T. and Kuo, C.-C. J. (1993): Texture analysis and classification with tree-structured wavelet transforms, *IEEE Trans. on Image Processing* 2(4): 429–441.

Charikar, M., Chekuri, C., Feder, T. and Motwani, R. (1997): Incremental clustering and dynamic information retrieval, *Proc. of the 29th Annual ACM Symposium on Theory of Computing*, 626–635.

Cheeseman, P. and Stutz, J. (1995): Bayesian classification (autoclass): theory and results, *in* U. Fayyad, G. Shapiro, P. Smyth and R. Uthu-

RUSAMY (editor), *Advances in Knowledge Discovery and Data Mining*, MIT Press, 153–180.

COLOMBO, C., DEL BIMBO, A. and PALA, P. (1999): Semantics in visual information retrieval, *IEEE Trans. on Multimedia* 6(3): 38–53.

COMPUTER VISION GROUP (2003): Image retrieval systems: `http://viper.unige.ch/other_systems/`.

CONGALTON, R. G. (1991): A review of assessing the accuracy of classification of remotly sensed data, *Remote Sensing of Environment* 37: 35–46.

COVER, T. M. and THOMAS, J. A. (1991): *Elements of information theory*, Wiley.

COX, I. J., MILLER, M. L., MINKA, T. P., PAPATHOMAS, T. V. and YIANILOS, P. N. (2000): The bayesian image retrieval system PicHunter: theory, implementation, and psychophysical experiments, *IEEE Trans. on Image Processing* 9(1): 20–37.

COX, I. J., MILLER, M. L., OMOHUNDRO, S. M. and YIANILOS, P. N. (1996): PicHunter: bayesian relevance feedback for image retrieval, *Proc. of the 13th IAPR Int. Conference on Pattern Recognition*, volume 13, 361–369.

COX, R. T. (1946): Probability, frequency, and reasonable expectation, *American Journal of Physics* 14: 1–13.

DASCHIEL, H. and DATCU, M. (2002a): Cluster structure evaluation of a dyadic k-means algorithm for mining large image archives, *Proc. SPIE Remote Sensing Symposium*.

DASCHIEL, H. and DATCU, M. (2002b): Exploration information content for remote sensing image mining application, *Workshop on Image Information Mining*, Zürich.

DASCHIEL, H. and DATCU, M. (2003a): Classification, semantic grouping and category learning of image content, *IEEE Trans. on Image Processing*: in review.

DASCHIEL, H. and DATCU, M. (2003b): Evaluation of man-machine interaction for image information mining, *IEEE Trans. on Multimedia*: in review.

DASCHIEL, H. and DATCU, M. (2003c): Image information mining system evaluation using information-theoretic measures, *Proc. of Advanced Concepts for Intelligent Vision Systems (ACIVS) 2003*, Ghent, 109–115.

DASCHIEL, H., DATCU, M. and PELIZZARI, A. (2003): Information mining in remote sensing image archives—part B: system evaluation, *IEEE Trans. on Geoscience and Remote Sensing*: in review.

DATCU, M. and SEIDEL, K. (1999): New concepts for remote sensing information dissemination: query by image content and information mining, *Proc. IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS'99)*, volume 3, 1335–1337.

Datcu, M., Daschiel, H., Pelizzari, A., Quartulli, M., Galoppo, A., Colapicchioni, A., Pastori, M., Seidel, K., Marchetti, P. G. and D'Elia, S. (2003): Information mining in remote sensing image archives—part A: system concepts, *IEEE Trans. on Geoscience and Remote Sensing* 41(12): 2923–2936.

Datcu, M., Pelizzari, A., Daschiel, H. and Quartulli, M. (2002): Advanced value adding to metric resolution SAR data: information mining, *Proc. of the 4th European Conference on Synthetic Aperture Radar (EUSAR 2002)*.

Datcu, M., Seidel, K. and Schwarz, G. (1999): Information mining in remote sensing image archives, *in* I. Kanellopoulos, G. Wilkinson and T. Moons (editor), *Machine Vision and Advanced Image Processing in Remote Sensing (MAVIRIC)*, Springer, 199–212.

Datcu, M., Seidel, K. and Walessa, M. (1998): Spatial information retrieval from remote sensing images. part I: information theoretical perspective, *IEEE Trans. on Geoscience and Remote Sensing* 36: 1431–1445.

Daubechies, J. (1990): The wavelet transform, time-frequency localization and signal analysis, *IEEE Trans. on Information Theory* 36: 961–1005.

Del Bimbo, A. and Pala, P. (1997): Visual image retrieval by elastic matching of user sketches, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 19(2): 121–132.

Dimai, A. (1999a): Assessment of effectiveness of content based image retrieval systems, *in* A. Smeulders and D. Huijsmans (editor), *Visual Information and Information Systems*, Springer.

Dimai, A. (1999b): Rotation invariant texture description using general moment invariants and Gabor filters, *Proc. of the 11th Scandinavian Conference on Image Analysis*.

Dimai, A. (1999c): Unsupervised extraction of salient region-descriptors for content based image retrieval, *Proc. of the 10th Int. Conference on Image Analysis and Processing*.

Dowe, J. (1993): Content-based retrieval in multimedia imaging, *Proc. SPIE Storage and Retrieval for Image and Video Databases*.

du Buf, J. M. H., Kardan, M. and Spann, M. (1990): Texture feature performance for image segmentation, *Pattern Recognition* 23(3/4): 291–309.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001): *Pattern Classification*, Wiley.

Egan, D. E. (1988): Individual differences in human-computer interaction, *in* M. Helander (editor), *Handbook of Human-Computer Interaction*, Springer, 543–568.

Faloutsos, C. and Lin, K. (1995): Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets, *Proc. SIGMOID*, 163–174.

Fisher, R. A. (1925): Theory of statistical estimation, *Proc. Cambridge Phil. Society*, volume 22, 700–725.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafine, J., Lee, D., Petkovic, D., Steele, D. and Yanker, P. (1995): Query by image and video content: the QBIC system, *IEEE Computer*.

Fu, K. S. (1982): *Synthactic pattern recognition and applications*, Prentice Hall.

Fukunaga, K. (1990): *Introduction to statistical pattern recognition*, Academic Press, San Diego.

Gee, A. and Cipolla, R. (1999): Tracking faces, *in* R. Cipolla and A. Pentland (editor), *Computer vision for human-machine interaction*, Cambridge University Press.

Gevers, T. and Smeulders, A. W. M. (1999): The PicToSeek image search system, *Proc. IEEE Int. Conference of Multimedia Computing and Systems*, volume 1.

Gevers, T. and Smeulders, A. W. M. (2000): PicToSeek: combining color and shape invariant features for image retrieval, *IEEE Trans. on Image Processing* 9(1): 102–119.

Gimel'farb, G. L. and Jain, A. K. (1996): On retrieving textured images from an image database, *Pattern Recognition* 29(9): 1461–1483.

Gokhale, M., Frigo, J., McCabe, K., Theiler, J. and Lavenier, D. (2001): Early experience with a hybrid processor: K-means clustering, *First Int. Conference on Engineering of Reconfigurable Systems and Algorithms (ERSA '01)*.

Goodman, J. W. (1975): Statistical properties of laser speckle patterns, *in* J. C. Dainty (editor), *Laser Speckle and related Phenomena*, Springer.

Gruber, T. R. (1993): A translation approach to portable ontologies, *Knowledge Acquisition* 5(2): 199–220.

Guida, G. and Tassa, C. (1994): *Design and development of knowledge-based systems — from life cycle to methodology*, Wiley, Chichester.

Günsel, B. and Tekalp, A. M. (1998): Shape similarity matching for query-by-example, *Pattern Recognition* 31(7): 931–944.

Haering, N. and da Vitoria Lobo, N. (1999): Features and classification methods to locate deciduous trees in images, *Computer Vision and Image Understanding* 75(1/2): 133–149.

Haralick, R. M., Shanmugam, K. and Dinstein, I. (1973): Textual features for image classification, *IEEE Trans. on Systems, Man, and Cybernetics* 3(6): 610–621.

Harman, D. (1992): Overview of the first text retrieval conference (TREC-1), *Proc. Text Retrieval Conference (TREC)*.

He, X. (2002): *Laplacian eigenmap for image retrieval*, Master's thesis, The University of Chicago.

Healey, J. and Picard, R. (1997): Digital processing of affective signals, *Technical Report 444*, MIT Media Laboratory.

Heckermann, D., Geiger, D. and Chickering, D. (1994): Learning bayesian networks: the combination of knowledge and statistical data, *Technical Report MSR-TR-94-09*, Microsoft Research.

Huang, J., Kumar, S., Mitra, M., Zhu, W. and Zabih, R. (1997): Image indexing using color correlograms, *Proc. IEEE Comp. Soc. Conf. Comp. Vis. and Patt. Rec.*, 762–768.

Jacobs, D. W., Weinshall, D. and Gdalyahu, Y. M. (1998): Condensing image databases when retrieval is based on non-metric distances, *Proc. ICCV'98*, 596–601.

Jain, A. K. and Dubes, R. C. (1988): *Algorithms for clustering data*, Prentice Hall.

Jain, A. K. and Vailaya, A. (1996): Image retrieval using color and shape, *Pattern Recognition* 29(8): 1233–1244.

Jermyn, I. H., Shaffrey, C. W. and Kingsbury, N. G. (2002): Evaluation methodologies for image retrieval systems, *Proc. of Advanced Concepts for Intelligent Vision Systems (ACIVS 2002)*.

Jose, J. M., Furner, J. and Harper, D. J. (1998): Spatial querying for image retrieval: a user-oriented evaluation, *Proc. 21st Annual Int. ACM/SIGIR Conference on Research and Development in Information retrieval*.

Jumarie, G. (1990): *Relative information*, Springer.

Kailath, T. (1967): The divergence and Bhattacharyya distance measures in signal selection, *IEEE Trans. on Communication Technology* 15(1): 52–60.

Kay, S. M. (1993): *Statistical signal processing*, volume 1, Prentice Hall.

Kelly, P. M. and Cannon, T. M. (1994): CANDID: Comparison algorithm for navigating digital image databases, *Statistical and Scientific Database Management*.

Koh, W. and Mui, L. (2001): An information theoretic approach for ontology-based interest matching, *Workshop on Ontology Learning*.

Kohonen, T. (1989): Self-organization and associative memory, *Proc. IEEE*, volume 78, 1464–1480.

Korfhage, R. R. (1997): *Information storage and retrieval*, Wiley.

Kullback, S. (1997): *Information theory and statistics*, Dover Publications.

Kumar, V. P. and Desai, U. B. (1996): Image interpretation using bayesian networks, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 18(1): 74–77.

Latecki, L. J. and Lakämpfer, R. (2000): Shape similarity measure based on correspondence of visual parts, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22(10): 1185–1190.

Leung, T. K. and Malik, J. (1999): Recognizing surfaces using three-dimensional textons, *IEEE Int. Conf. on Computer Vision (ICCV)*.

Lee, C. S., Ma, W. Y. and Zhang, H. J. (1999): Information embedding based on user's relevance feedback for image retrieval, *SPIE Symposium on Voice, Video and Data Communications*.

Lehtonen, R. and Pahkinen, E. (2003): *Practical methods for design and analysis of complex surveys*, Wiley.

Lele, S. R. and Ord, J. K. (1986): Conditional least squares estimation for spatial processes: some asymptotics results, *Tech. Rep. 65*, Dept. Stat., The Pennsylvania State Univ.

Li, C.-S. and Chen, M.-S. (1996): Progressive texture matching for earth observing satellite image databases, *Proc. Multimedia Storage and Archiving Systems (SPIE)*.

Liu, F. and Picard, R. (1996): Periodicity, directionality, and randomness: wold features for image modeling and retrieval, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 18(7): 722–733.

Lorenz, O. (1996): Automatic indexing of line drawings for content-based image retrieval: Unpublished doctoral thesis.

Lu, Y., Hu, C., Zhu, X., Zhang, H. and Yang, Q. (2000): A unified framework for semantics and feature based relevance feedback in image retrieval systems, *ACM Multimedia*, 31–37.

Ma, W. and Manjunath, B. S. (1998): A texture thesaurus for browsing large aerial photographs, *Journal of the American Society for Information Science* 49(7): 633–648.

Ma, W. and Manjunath, B. S. (1999): NeTra: a toolbox for navigating large image databases, *Multimedia Systems* 7(3): 184–198.

Ma, W. and Zhang, H. (1998): Benchmarking of image features for content-based retrieval, *Proc. of the 32nd Asilomar Conference on Signals, Systems and Computers*.

MacKay, D. (1991): *Bayesian methods for adaptive models*, PhD thesis, California Institute of Technology.

Mandelbrot, B. B. (1982): *The fractal geometry of nature*, Freeman.

Manjunath, B. S. and Ma, W. Y. (1996): Texture features for browsing and retrieval of image data, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 18(8): 837–842.

Mao, J. and Jain, A. K. (1992): Texture classification and segmentation using multiresolution simultaneous autoregressive models, *Pattern Recognition* 25(2): 173–188.

Marchisio, G. B. and Cornelison, J. (1999): Content-based search and clustering of remote sensing imagery, *Proc. of the IEEE Int. Conference on Geoscience and Remote Sensing (IGARSS'99)*, 290–292.

McAulay, R. J. and Hofstetter, E. M. (1971): Barankin bounds parameter estimation, *IEEE Trans. on Information Theory* 17(6): 669–676.

McLean, G. F. (1993): Vector quantization for texture classification, *IEEE Trans. on Systems, Man and Cybernetics* 23(3): 637–649.

Minka, T. P. and Picard, R. W. (1997): Interactive learning with a "society of models", *Pattern Recognition* 30(4): 565–581.

Müller, H., Müller, W., McG. Squire, D., Marchand-Maillet, S. and Pun, T. (2001): Performance evaluation in content-based image retrieval: overview and proposals, *Pattern Recognition Letters* 22(5): 593–601.

Mumford, D. (1987): The problem with robust shape descriptions, *First Int. Conference on Computer Vision*.

Nastar, C. (1997): The image shape spectrum for image retrieval, *Technical Report 3206*, Institut National de Recherche en Informatique et en Automatique (INRIA), France.

Neuman, W. R. (1991): *The Future of Mass Audience*, Cambridge University Press.

Ng, R. and Sedighian, A. (1998): Evaluating multi-dimensional indexing structures for images transformed by principal component analysis, *Proc. SPIE Storage and Retrieval for Image and Video Databases*.

Niblack, W., Barber, R., Equitz, W., Flickner, M., Glasman, E., Petkovic, D., Yanker, P., Faloutsos, C. and Taubin, G. (1993): The QBIC project: querying images by content using color, texture and shape, *Proc. SPIE Storage Retrieval for Image and Video Databases*, 173–187.

Oh, I.-S., Lee, J.-S. and Suen, C. Y. (1999): Analysis of class separation and combination of class-dependent features for handwriting recognition, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 21(10): 1089–1094.

Olsen, S. I. (1993a): Estimation of noise in images: an evaluation, *CVGIP: Graphical Models and Image Processing* 55(4): 319–323.

Olsen, S. I. (1993b): Noise variance estimation in images, *The 8'th Scandinavian Conference on Image Analysis*, 989–996.

Ortega, M., Rui, Y., Chakrabarti, K., Mehrotra, S. and Huang, T. S. (1997): Supporting similarity queries in MARS, *Proc. of the 5th ACM Int. Multimedia Conference*, 403–413.

Pao, M. L. and Lee, M. (1989): *Concepts of information retrieval*, Libraries Unlimited.

Papathomas, T. V., Conway, E., Cox, I. J., Ghosn, J., Miller, M. L., Minka, T. P. and Yianilos, P. N. (1998): Psychophysical studies of the performance of an image database retrieval system, *Proc. Symp. Electronic Imaging: Conf. Human Vision and Electronic Imaging III*, 591–602.

Papoulis, A. (1984): *Probability, random variables and stochastic processes*, McGraw-Hill.

Parzen, E. (1962): On estimation of a probability density function and mode, *Annals of Mathematical Statistics* 33(3): 1065–1076.

Pass, G., Zabih, R. and Miller, J. (1996): Comparing images using color coherence vectors, *ACM Multimedia*, 65–73.

Patrick, E. A. and Fischer, F. P. (1970): A generalized k-nearest neighbor rule, *Information and control* 16(2): 128–152.

Pentland, A., Picard, R. W. and Sclaroff, S. (1996): Photobook: content-based manipulation of image databases, *Int. Journal of Computer Vision* 18(3): 233–254.

Picard, R. W. (1996): Computer learning of subjectivity, *Technical Report 359*, MIT Media Laboratory.

Picard, R. W. and Kabir, T. (1993): Finding similar patterns in large image databases, *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing*, 161–164.

Picard, R. W. and Lui, F. (1994): A new wold ordering for image similarity, *Proc. ICASSP*, volume 5, Adelaide, Australia, 129–132.

Rao, A. R. and Lohse, G. L. (1993): Towards a texture naming system: identifying relevant dimensions of texture, *Vision Research*, volume 36, 1649–1669.

Rao, A., Srihari, R. K., Zhu, L. and Zhang, A. (2002): A method for measuring the complexity of image databases, *IEEE Trans. on Multimedia* 40(2): 160–173.

Reed, T. R. (1993): A review of recent texture segmentation and feature extraction techniques, *Computer Vision, Graphics and Image Processing* 57(3): 359–372.

Resnick, P., Suchak, N. I., Bergstrom, P. and Riedl, J. (1994): Grouplens: an open architecture for collaborative filtering of netnews, *Proc. of ACM 1994 Conference on Computer Supported Cooperative Work*, Chapel Hill, North Carolina, 175–186.

Richards, J. A. and Jia, X. (1999): *Remote sensing digital image analysis*, Springer.

Ruanaidh, J. J. O. and Fitzgerald, W. J. (1996): *Numerical bayesian methods applied to signal processing*, Springer.

Rui, Y., Chakrabarti, K., Mehrotra, S., Zhao, Y. and Huang, T. S. (1997a): Dynamic clustering for optimal retrieval in high dimensional multimedia databases, *Technical Report TR-MARS-10-97*, University of Illinois at Urbana-Champaign.

Rui, Y., Huang, T. S. and Mehrotra, S. (1997b): Content-based image retrieval with relevance feedback in MARS, *Proc. IEEE Int. Conference on Image Processing*.

Rui, Y., Huang, T. S., Ortega, M. and Mehrotra, S. (1998a): Relevance feedback: a power tool for interactive content-based image retrieval, *IEEE Trans. on Circuits and Video Technology* 8(5): 644–655.

Rui, Y., Huang, T. and Chang, S. (1999): Image retrieval: current techniques, promising directions and open issues, *Journal of Visual Communication and Image Representation* 10(4): 39–62.

Rui, Y., She, A. C. and Huang, T. S. (1998b): A modified fourier descriptor for shape matching in MARS, *in* S. K. Chang (editor), *Image Databases and Multimedia Search, Series on Software Engineering and Knowledge Engineering*.

Schäuble, P. (1997): *Multimedia information retrieval*, Kluwer Academic Publisher.

Schröder, M., Rehrauer, H., Seidel, K. and Datcu, M. (2000): Interactive learning and probabilistic retrieval in remote sensing image archives, *IEEE Trans. on Geoscience and Remote Sensing* 38(5): 2288–2298.

Schröder-Brzosniowsky, M. (2000): *Stochastic modeling of image content in remote sensing image archives*, PhD thesis, ETH Zürich.

Schröder, M. and Dimai, A. (1998): Texture information in remote sensing images: a case study, *Workshop on Texture Analysis (WTA '98)*.

Schröder, M., Rehrauer, H., Seidel, K. and Datcu, M. (1998): Spatial information retrieval from remote sensing images. part II: Gibbs-Markov random fields, *IEEE Trans. on Geoscience and Remote Sensing* 36: 1446–1455.

Sclaroff, S. (1997): Deformable prototypes for encoding shape categories in image databases, *Pattern Recognition* 30(4): 627–641.

Seidel, K. and Datcu, M. (1999): Architecture of a new generation of remote sensing ground segments, *in* J. L. Casanova (editor), *Proc. of the 19th EARSeL Symposium on Remote Sensing in the 21st Century, Valladolid*, A. A. Balkema Rotterdam/Brookfield, 223–228.

Shaffrey, C. W., Jermyn, I. H. and Kingsbury, N. G. (2002): Psychophysical evaluation of image segmentation algorithms, *Proc. of Advanced Concepts for Intelligent Vision Systems (ACIVS) 2002*.

Shannon, C. E. (1948): A mathematical theory of communication, *Bell System Technical Journal* 27: 379–423.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. and Jain, R. (2000): Content-based image retrieval at the end of the early years, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22(12): 1349–1379.

Smith, J. R. (1998): Image retrieval evaluation, *IEEE Workshop on Content-based Access of Image and Video Libraries*, Santa Barbara, California.

Song, Y. and Zhang, A. (2003): SceneryAnalyzer: a system supporting semantics-based image retrieval, *in* C. Djeraba (editor), *Intelligent Multimedia Documents*, Kluwer Academic Publishers.

Stone, H. S. and Li, C. S. (1997): Image matching by means of intensity and texture matching in the Fourier domain, *Proc. IEEE Int. Conf. Image Processing*.

Stricker, M. A. and Dimai, A. (1997): Spectral covariance and fuzzy regions for image indexing, *Machine Vision and Applications* 10(2): 66–73.

Stricker, M. A. and Orengo, M. (1995): Similarity of color images, *Proc. SPIE Storage and Retrieval for Image and Video Databases*.

Su, Z., Zhang, H., Li, S. and Ma, S. (2003): Relevance feedback in content-based image retrieval: bayesian framework, feature subspaces, and progressive learning, *IEEE Trans. on Image Processing* 12(8): 924–937.

Swain, M. J. and Ballard, D. H. (1991): Color indexing, *Int. Journal of Computer Vision* 7(1): 11–32.

Swain, P. H. and Davis, S. M. (editor) (1978): *Remote sensing: the quantitative approach*, McGraw-Hill.

Tagare, H. (1997): Increasing retrieval efficiency by index tree adaptation, *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*.

Tamura, H., Mori, S. and Yamawaki, T. (1978): Textural features corresponding to visual perception, *IEEE Tran. on Systems, Man, and Cybernetcis* 8(6): 460–473.

Tan, T. N. (1998): Rotation invariant texture features and their use in automatic script identification, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 20(7): 751–756.

Therrien, C. W. (1989): *Decision, estimation and classification*, Wiley.

Tuceryan, M. and Jain, A. K. (1990): Texture segmentation using Voronoi polygons, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 12: 211–216.

Tuceryan, M. and Jain, A. K. (1998): Texture analysis, *in* C. H. Chen, L. F. Pau and P. S. P. Wang (editor), *Handbook of Pattern Recognition and Computer Vision*, World Scientific Publishing, 207–248.

Vailaya, A., Figueiredo, M. A. T., Jain, A. K. and Zhang, H.-J. (2001): Image classification for content-based indexing, *IEEE Trans. on Image Processing* 10(1): 117–130.

Vailaya, A., Jain, A. K. and Zhang, H. J. (1998): On image classification: city images vs. landscapes, *Pattern Recognition* 31(12): 1921–1935.

van Trees, H. L. (1968): *Detection, estimation and modulation theory. Part I*, Wiley, New York.

Vasconcelos, N. and Lippman, A. (2000): Learning from user feedback in image retrieval systems, *in* S. Solla, T. Leen and K. Müller (editor), *Advances in Neural Information Processing Systems 12*, MIT Press.

Veltkamp, C. R., Burkhardt, H. and Kriegel, H.-P. (2001): *State-of-the-art in content-based image and video retrieval*, Kluwer.

Veltkamp, R. and Hagedoorn, M. (1999): State-of-the-art in shape matching, *Technical Report UU-CS-1999-27*, Utrecht University, the Netherlands.

Walessa, M. and Datcu, M. (2000): Model-based despeckling and information extraction from SAR images, *IEEE Trans. on Geoscience and Remote Sensing* 38(5): 2258–2269.

Wong, Y.-F. (1993): Clustering data by melting, *Neural Computation* 89–104.

Zhang, H. and Zhong, D. (1995): A scheme for visual feature based images indexing, *Proc SPIE Storage and Retrieval for Image and Video Databases*.

Zhang, J., Hsu, W. and Lee, M. L. (2001): Image mining: issues, frameworks and techniques, *Proc. of the Second Int. Workshop on Multimedia Data Mining (MDM/KDD'2001)*, San Franzisco, CA, USA.

Zhang, T., Ramakrishnan, R. and Livny, M. (1996): Birch: An efficient data clustering method for very large databases, *Proc. of the ACM SIGMOD Conference on Management of Data*, 103–114.

Ziv, J. and Zakai, M. (1969): Some lower bounds on signal parameter estimation, *IEEE Trans. on Information Theory* 15(3): 386–391.

Acknowledgments

Since I started my Ph.D. period at DLR in spring 2001, many people have helped and inspired me. Without their support, it would not have been possible to write this dissertation.

Fist, I want to express my particular thank to my supervisor at DLR Prof. Dr.-Ing. Mihai Datcu. He offered me the opportunity to extend and continue the work of my diploma in remote sensing image data mining. Moreover, he introduced me to the signal- and information-theoretic way of data analysis. His special interest in data mining has resulted in fruitful discussions and critical comments on the performed experiments and presented articles. I feel fortunate to having had the possibility to work together with such a competent supervisor and researcher.

I am also very grateful to Prof. Dr.-Ing. Olaf Hellwich of the chair of Computer Vision & Remote Sensing at the Technical University of Berlin for giving me the opportunity to carry out the thesis at DLR and for the good and friendly cooperation. I also want to thank Prof. Dr.-Ing. Lothar Gründig of the chair of Geodesy at the Technical University of Berlin, the second co-examiner of my dissertation.

In terms of projects and experiments on image information mining, I want to thank Andrea Colapicchioni and Annalisa Galoppo from Advanced Computer Systems (ACS) for the good cooperation and for maintaining the mining system. Furthermore, I want to thank the researchers from the European Union Satellite Center (EU-SC) and Nansen Environmental and Remote Sensing Center (NERSC) for the good cooperation during the system evaluation procedure and their detailed feedback about the mining system.

I would like to thank the members of the image analysis group at DLR/IMF for their support on the Ph.D. work. In particular, I want to thank my former and current roommates Andrea Pelizzari and Ines Gomez for the very friendly and convenient atmosphere in our office. Special thanks also go to Marco Quartulli and Mariana Ciucu. Our discussions and 4 o'clock coffee breaks contributed much to the good working atmosphere in the image analysis group. Whenever I had problems, they had an open ear and took the time for discussions. Besides, I want to thank Dr. Gintautas Palubinskas, Cyrille Maire and Patrick Heas.

Additional thanks goes to our system manager Rolf Konjack for his technical support and to our secretary Mrs. Hantel. She kept the 'unpleasant' administrative matters from us as far a possible.

I want to thank M. A. Stefan Fromholzer for proofreading this dissertation. Very special thanks go to my wife Angelika for her understanding and encouragement during this three years of Ph.D. work and for proofreading all my articles and this dissertation. She showed me that life is more than just studying and programming. Finally, I want to thank my parents, Franziska and Herbert Daschiel, for their love, support and patience during my study and Ph.D.

Curriculum Vitae

Personal data:

Name:	Herbert Andreas Daschiel
Date of birth:	18. August 1973
Place of birth:	Trostberg, Germany
Nationality:	German
Marital status:	married

Secondary education

09.1986 - 08.1990:	Staatl. Realschule Trostberg, Mittlere Reife
09.1990 - 08.1992:	Staatl. Fachoberschule Traunstein, Fachhochschulreife

Studies:

10.1994 - 09.1996:	Vermessungswesen, FH München
10.1996 - 01.2001:	Vermessungswesen, TU München
10.2000 - 01.2001:	Diploma thesis at German Aerospace Center (DLR) Topic: Bayesian Texture Extraction from High Resolution SAR Images

Military service:

10.1992 - 09.1994:	Hochgebirgsjägerzug, Berchtesgaden

Practical training:

10.1994 - 02.1995:	Vermessungsamt, Traunstein
03.1999 - 04.1999:	DLR, Topic: Texture Extraction and Classification
04.2000 - 07.2000:	DLR, JRC Project "Database for the City of Tomorrow"

Occupational activity:

03.2001 - 02.2004:	Ph.D. at German Aerospace Center (DLR)
03.2004 - this day:	Project Scientist at DLR