Philipps Universität Marburg

# Visual Concept Detection in Images and Videos

Dissertation
zur Erlangung des Doktorgrades der Naturwissenschaften
(Dr. rer. nat.)

dem Fachbereich Mathematik und Informatik
der Philipps-Universität Marburg
vorgelegt von

Markus Mühling
geboren in Schwalmstadt

Marburg, 2014
Hochschulkennziffer 1180

Vom Fachbereich Mathematik und Informatik der
Philipps-Universität Marburg als Dissertation am
*24.03.2014*
angenommen.

Erstgutachter: Prof. Dr. Bernd Freisleben

Zweitgutachter: Prof. Dr. Wolfgang Effelsberg

Tag der mündlichen Prüfung am *18.09.2014*.

# Abstract

The rapidly increasing proliferation of digital images and videos leads to a situation where content-based search in multimedia databases becomes more and more important. A prerequisite for effective image and video search is to analyze and index media content automatically. Current approaches in the field of image and video retrieval focus on semantic concepts serving as an intermediate description to bridge the "semantic gap" between the data representation and the human interpretation. Due to the large complexity and variability in the appearance of visual concepts, the detection of arbitrary concepts represents a very challenging task. In this thesis, the following aspects of visual concept detection systems are addressed:

First, enhanced local descriptors for mid-level feature coding are presented. Based on the observation that scale-invariant feature transform (SIFT) descriptors with different spatial extents yield large performance differences, a novel concept detection system is proposed that combines feature representations for different spatial extents using multiple kernel learning (MKL). A multi-modal video concept detection system is presented that relies on Bag-of-Words representations for visual and in particular for audio features. Furthermore, a method for the SIFT-based integration of color information, called color moment SIFT, is introduced. Comparative experimental results demonstrate the superior performance of the proposed systems on the *Mediamill* and on the *VOC* Challenge.

Second, an approach is presented that systematically utilizes results of object detectors. Novel object-based features are generated based on object detection results using different pooling strategies. For videos, detection results are assembled to object sequences and a shot-based confidence score as well as further features, such as position, frame coverage or movement, are computed for each object class. These features are used as additional input for the support vector machine (SVM)-based concept classifiers. Thus, other related concepts can also profit from object-based features. Extensive experiments on the *Mediamill*, *VOC* and *TRECVid* Challenge show significant improvements in terms of retrieval performance not only for the object classes, but also in particular for a large number of indirectly related concepts. Moreover, it has been demonstrated that a few object-based features are beneficial for a large number of concept classes. On the *VOC* Challenge, the additional use of object-based features led to a superior performance for the image classification task of 63.8% mean average precision

(AP). Furthermore, the generalization capabilities of concept models are investigated. It is shown that different source and target domains lead to a severe loss in concept detection performance. In these cross-domain settings, object-based features achieve a significant performance improvement. Since it is inefficient to run a large number of single-class object detectors, it is additionally demonstrated how a concurrent multi-class object detection system can be constructed to speed up the detection of many object classes in images.

Third, a novel, purely web-supervised learning approach for modeling heterogeneous concept classes in images is proposed. Tags and annotations of multimedia data in the WWW are rich sources of information that can be employed for learning visual concepts. The presented approach is aimed at continuous long-term learning of appearance models and improving these models periodically. For this purpose, several components have been developed: a crawling component, a multi-modal clustering component for spam detection and subclass identification, a novel learning component, called “random savanna”, a validation component, an updating component, and a scalability manager. Only a single word describing the visual concept is required to initiate the learning process. Experimental results demonstrate the capabilities of the individual components.

Finally, a generic concept detection system is applied to support interdisciplinary research efforts in the field of psychology and media science. The psychological research question addressed in the field of behavioral sciences is, whether and how playing violent content in computer games may induce aggression. Therefore, novel semantic concepts most notably “violence” are detected in computer game videos to gain insights into the interrelationship of violent game events and the brain activity of a player. Experimental results demonstrate the excellent performance of the proposed automatic concept detection approach for such interdisciplinary research.

# Zusammenfassung

Aufgrund der starken Verbreitung digitaler Bild- und Videodaten wächst der Bedarf an Möglichkeiten zur inhaltsbasierten Suche. Hierzu sind Verfahren, welche Bilder und Videos automatisch mit entsprechenden Annotationen versehen, erforderlich. Aktuelle Forschungsansätze im Bereich Bild- und Videoretrieval basieren auf semantischen Konzepten, die als intermediäre Beschreibung dienen, um die "semantische Lücke" zwischen der Datenrepräsentation und der menschlichen Interpretation des Inhalts zu schließen. Die Detektion beliebiger visueller Konzepte stellt aufgund der hohen Komplexität und Variabilität im Erscheinungsbild eine große Herausforderung dar. Die vorliegende Arbeit befasst sich mit folgenden Aspekten zur visuellen Konzeptdetektion in Bildern und Videos:

Erstens werden verbesserte lokale Deskriptoren für die Kodierung von Mid-Level Merkmalen präsentiert. Basierend auf der Beobachtung, dass Scale-Invariant Feature Transform (SIFT) Deskriptoren mit unterschiedlichen räumlichen Ausdehnungen große Performanceunterschiede pro Konzept erzielen, wird ein neuer Ansatz zur Konzeptdetektion vorgestellt. Dieser kombiniert Merkmalsrepräsentationen für unterschiedliche räumliche Ausdehnungen mittels Multiple Kernel Learning (MKL). Zur Konzeptdetektion in Videos wird ein multimodales System vorgeschlagen, welches das Bag-of-Words Verfahren auf visuelle Merkmale und insbesondere auf Audiomerkmale anwendet. Zudem wird eine Methode zur SIFT-basierten Integration von Farbinformationen, Color Moment SIFT genannt, eingeführt. Experimentelle Ergebnisse demonstrieren die hervorragende Performanz der vorgeschlagenen Ansätze auf der *Mediamill* sowie auf der *VOC* Challenge.

Zweitens wird ein Ansatz präsentiert, der systematisch die Ergebnisse von Objektdetektoren nutzt. Neuartige Objekt-basierte Merkmale werden basierend auf Objektdetektionsergebnissen unter Verwendung unterschiedlicher Pooling-Strategien generiert. Für Videos werden die Detektionsergebnisse zu Objektsequenzen zusammengefasst und ein Shot-basierter Konfidenzwert sowie weitere Merkmale wie z.B. Position, Videoframe-Abdeckung oder Bewegung für jede Objektklasse berechnet. Diese Merkmale werden als zusätzliche Eingabe für die Support Vector Machine (SVM)-basierten Konzeptklassifizierer verwendet. Somit können auch indirekte Konzepte von Objekt-basierten Merkmalen profitieren. Umfangreiche Experimente auf der *Mediamill*, *VOC* und *TRECVid* Challenge zeigen signifikante Verbesserungen der Retrieval-Performance, nicht nur für die Objektklassen selbst sondern insbesondere für eine große Anzahl indi-

rekt im Zusammenhang mit den Objektklassen stehender Konzepte. Es konnte gezeigt werden, dass bereits eine geringe Anzahl Objekt-basierter Merkmale für eine große Anzahl an Konzeptklassen vorteilhaft ist. Auf der *VOC* Challenge wurde durch das Hinzunehmen Objekt-basierter Merkmale eine herausragende Performanz von im Schnitt 63.8% Average Precision (AP) erzielt. Desweiteren wurde die Generalisierungsfähigkeit von Konzeptmodellen untersucht. So wurde gezeigt, dass unterschiedliche Quell- und Zieldomänen zu einem schwerwiegenden Performanzverlust führen und dass Objekt-basierte Merkmale die Domänenübergreifende Performanz in der Konzeptdetektion deutlich verbessern. Da es sich als ineffizient erwiesen hat, eine große Anzahl von Single-Class Objektdetektoren anzuwenden, wurde ferner ein Multi-Class System basierend auf Hough-Forests zur Beschleunigung der Detektion mehrerer Objektklassen in Bildern vorgeschlagen.

Drittens wird ein neuartiger rein Web-überwachter Lernansatz zur Modellierung heterogener Konzeptklassen in Bildern vorgestellt. Annotationen von Multimediadaten im WWW sind ergiebige Informationsquellen, welche zum Lernen visueller Konzepte herangezogen werden können. Der präsentierte Ansatz zielt auf ein kontinuierliches langfristiges Lernen von Modellen und deren periodischer Verbesserung ab. Zu diesem Zweck wurden mehrere Komponeneten entwickelt, darunter ein Webcrawler, eine multi-modale Clustering-Komponente zur Spamdetektion und zur Erkennung von Unterklassen, ein neuartiges Lernverfahren, das sog. “Random Savanna” sowie eine Validierungs-, eine Aktualisierungs- und eine Skalierbarkeitskomponente. Ein einziges Wort zur Beschreibung des visuellen Konzepts reicht aus, um den Lernprozess zu starten. Experimentelle Ergebnisse demonstrieren die Leistungsfähigkeit der einzelnen Komponenten.

Schließlich wird ein generisches System zur visuellen Konzeptdetektion vorgestellt, welches zur Unterstützung interdisziplinärer Forschung im Bereich Psychologie und Medienwissenschaften eingesetzt wird. Um Einblick in die Wechselbeziehung von gewalttätigen Spielereignissen und der Gehirnaktivität des Spielers zu gewinnen, wurden neuartige semantische Konzepte, allen voran “Gewalt”, in Computerspielvideos detektiert. Experimentelle Ergebnisse zeigen die exzellente Performanz des vorgeschlagenen generischen Ansatzes zur automatischen Konzeptdetektion für derartige interdisziplinäre Forschung.

# Acknowledgements

First, I like to express my gratitude to my supervisor Prof. Dr. Bernd Freisleben for his support and the opportunity to pursue scientific research. I also want to thank Prof. Dr. Ralph Ewerth for the fruitful discussions and helpful comments. Furthermore, I would like to thank Prof. Dr. Wolfgang Effelsberg of the University of Mannheim for acting as the second reviewer.

This work is partly supported by the German Research Foundation and by the German Ministry of Education and Research. The funding in the context of the collaborative research center "Media Upheavals" (DFG SFB/FK 615) and the project "MediaGrid" (BMBF D-Grid Initiative) is gratefully acknowledged.

In addition, I want to thank my colleagues from the Distributed Systems Group for the great time at the University of Marburg.

Finally, I would like to thank my family, most of all, my son Maximilian for giving me a great deal of motivation and my wife Miriam for her encouragement and patience.

# Contents

*"You can never solve a problem on the level on which it was created."*

Albert Einstein

# 1 Introduction

## 1.1 Motivation

Content-based search in multimedia databases and archives becomes more and more important due to the rapidly increasing proliferation of digital images and videos. In Germany, almost 10 billion digital pictures are taken per year. Billions of images and videos are hosted on web portals such as Flickr or YouTube. At Flickr, more than 3,000 images are uploaded every minute.

While humans easily understand the content of images and videos within split seconds, current search engines have no or only a very limited ability to recognize the image or scene content. Instead, the search is primarily based on subjective and partly erroneous, scarce and incomplete, manually created annotations and comments. Google's image search, for example, does not find images if important tags are missing in the filename or in the surrounding HyperText Markup Language (HTML) text. In practice, missing tags or annotations make images unfindable. To enable content-based search, the scene content of images and videos needs to be automatically recognized and annotated with semantic concepts.

This thesis on image and video content analysis is additionally motivated by the project *Methods and Tools for Computer-Assisted Scientific Media Research* (*MT*), which was part of the collaborative research center *Media Upheavals* (SFB/FK 615) conducted at the Universities of Siegen and Marburg and funded by the German Research Foundation (DFG).

The interdisciplinary media project focused on the media upheavals at the beginning and the end of the 20th century. While the emergence of cinema marked

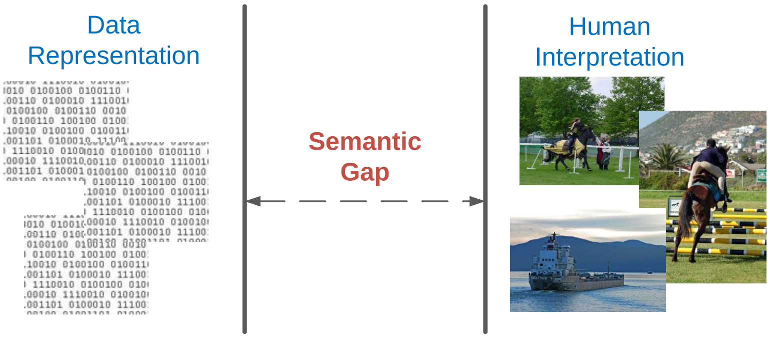


Figure 1.1: “Semantic gap” between the data representation (low-level features) and the human interpretation of the semantic content.

the first media upheaval, the second one was triggered by the introduction of digital media and the internet. The relevance of media upheavals, the genesis and mutation of media cultures and the development of media aesthetics were investigated within this project.

The research project $MT$ supported media scientists in applying film analysis. The goal was to provide a video content analysis system called *Videana* to support the scholarly analysis of audio-visual material and to relieve media scholars from the time-consuming task of tagging images and films manually. This includes the annotation of images and videos with semantic concepts like “indoor/outdoor”, “studio”, “anchor”, “politician” or concepts rather related to media science such as “bullet time”, “top view” or “duel”. Additionally, concept detection offers a more objective annotation of multimedia databases and archives instead of inhomogeneous manually assigned user tags.

Altogether, the need for efficient retrieval techniques to support search and navigation in multimedia collections is rapidly growing. Visual concept detection, also known as high-level feature extraction or semantic indexing, plays the key role in the field of image and video retrieval and is thus the research focus of the present thesis.

## 1.2 Problem Statement

The challenging task to be solved is the automatic assignment of semantic tags to images and videos for the purpose of facilitating content-based search and navigation. The fundamental problem of content-based search is to overcome the discrepancy between the extracted (low-level) features and the human interpretation of the (audio-)visual data. In the literature, this discrepancy is also known as “semantic gap” (Figure 1.1). Smeulders et al. [2000] describe the semantic gap as “the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a

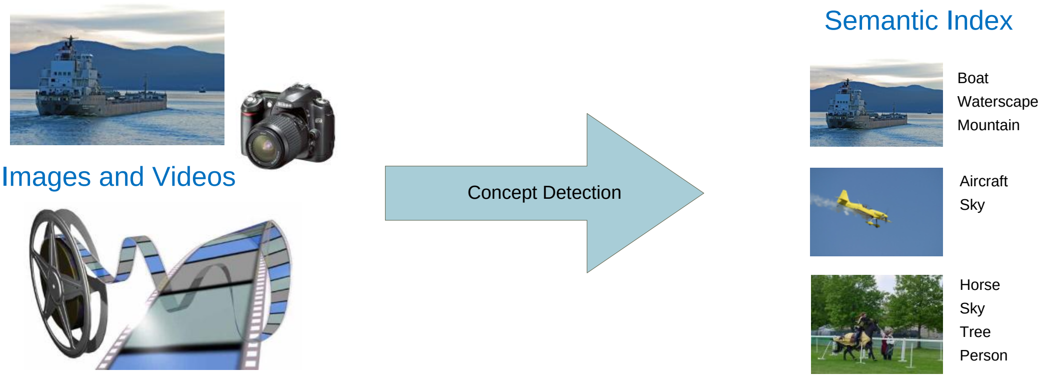


Figure 1.2: Concept detection represents the task of automatically assigning semantic concepts to images and video shots.

given situation". Query-by-content based on low-level features turned out to be insufficient to search successfully in large-scale multimedia databases [Naphade and Smith 2004]. Thus, state-of-the-art approaches in the field of image and video retrieval focus on semantic concepts serving as an intermediate description to bridge the "semantic gap" between the data representation and the human interpretation. Semantic concepts, also known as high-level features, can be, for example, objects, sites, scenes, personalities, events or activities. Concept detection, as the enrichment of images and videos with semantic tags (Figure 1.2), is the key to facilitate access to multimedia databases. Given the semantic index, search queries on large-scale multimedia databases can be processed very efficiently. Based on the resulting index and given that the concept lexicon is large enough, arbitrary search queries can be responded by mapping the queries to the semantic concepts. Hauptmann et al. [2007] stated that approximately 5,000 concepts, detected with a minimum accuracy of 10% mean AP, are sufficient to provide search results comparable to text retrieval in the World Wide Web.

The detection of arbitrary concepts is a very challenging task due to the large complexity and variability in the appearance of visual concepts. Especially the generalization capabilities of learned concept models applied to foreign target domains are a severe problem in the field of concept detection, because in some cases the visual appearance of semantic concepts strongly depends on the domain of the respective image or video source. This can be easily observed, for example, in the difference of television news and user-generated YouTube videos.

Altogether, a major problem in the field of concept detection is to find robust features being the fundament for successful concept detection systems.

Besides feature extraction and machine learning issues, another problem consists in the acquisition of concept-specific training data in sufficient quantity and quality. This manual, very time-consuming step is a prerequisite for learning classification rules. Reducing the manual annotation effort would facilitate the construction of new concept models.

## 1.3 Contributions

The contributions of this thesis to the scientific state-of-the-art are as follows:

- State-of-the-art systems for visual concept detection typically rely on the Bag-of-Visual-Words (BoVW) feature representation. While several aspects of this representation have been investigated, such as keypoint sampling strategy, vocabulary size, projection method or the weighting scheme, the impact of the spatial extents of local descriptors has not been studied in previous work. In this thesis, the effect of different spatial extents in an up-to-date system for visual concept detection is investigated. Based on the observation that SIFT descriptors with different spatial extents yield large performance differences, a novel concept detection system is proposed that combines feature representations for different spatial extents using MKL. Superior performance is achieved on the *Mediamill* as well as on the *PASCAL Visual Object Classes* (*VOC*) Challenge using the proposed system. This contribution is presented in Section 3.2.

- A novel approach to multi-modal video concept detection is proposed in Section 3.3 that combines visual and audio features. A Bag-of-Auditory-Words (BoAW) approach is investigated that models mel-frequency cepstral coefficients (MFCC) features in an auditory vocabulary. Furthermore, MKL is proposed as appropriate fusion scheme to combine the resulting BoAW features with state-of-the-art visual features. Experimental results show the effectiveness of using BoAW features: The system using BoAW features and a SVM with a $\chi^2$-kernel is superior to a current audio approach relying on probabilistic latent semantic indexing. Furthermore, it is shown that an early fusion scheme degrades detection performance, whereas the combination of auditory and visual Bag-of-Words features via MKL yields a clear performance improvement.

- Color information can be very helpful to classify semantic concepts, like for example “sunset”, “meadow” or “sky”. It contributes to improve the discriminative power of concept classifiers. Existing approaches combine SIFT descriptors with local color histograms or concatenate SIFT descriptors extracted from different color channels. In Section 3.4, a method for the SIFT-based integration of color information, called color moment SIFT (CMSIFT), is presented. CMSIFT achieves the same concept detection performance as the state-of-the-art transformed color SIFT descriptor, but it is considerably faster.

- The latest systems for generic concept detection mainly rely on BoVW features and, in some cases, additionally on features based on face detection, optical character recognition and/or speech recognition in the case of videos.

Based on the observation that the use of face detection results improved the performance of several face related concepts, further object detectors trained on separate data sets are incorporated. Novel features based upon object detection results are used as additional input for SVM-based concept classifiers. Additionally, MKL is proposed to combine object-based and BoVW features. Extensive experiments on the *Mediamill*, *VOC* and *TRECVid Challenge* show significant improvements in terms of retrieval performance, not only for the object classes, but also in particular for a large number of indirectly related concepts. Furthermore, it is shown that the use of object-based features clearly improves the generalization capabilities of the learned concept models in cross-domain settings, for example, from broadcast news videos to documentary films and vice versa. These contributions are presented in Chapter 4.

- To reduce the processing time of the object recognition task, an extension of random Hough forests for the purpose of multi-class object detection is proposed. Experimental results for the *Caltech-101* test set demonstrate that the performance of the proposed approach is almost as good as the performance of a single-class object detector, even when detecting a large number of 24 object classes at a time. This contribution is presented in Section 4.4.

- A novel incremental and scalable web-supervised learning system that continuously learns appearance models for image categories with heterogeneous appearances and that improves these models periodically is presented in Chapter 5. Simply specifying the name of the concept that has to be learned initializes the proposed system, and there is no further supervision afterwards. Textual and visual information on websites are used to filter out irrelevant and misleading training images. To obtain a robust, flexible, and updatable way of learning, a novel learning framework is presented that relies on clustering in order to identify visual subclasses before using an ensemble of random forests, called “random savanna”, for subclass learning. Experimental results demonstrate that the proposed web-supervised learning approach outperforms a SVM, while at the same time being simply parallelizable in the training and testing phase.

- An automatic video content analysis system is built to support interdisciplinary research efforts in the field of psychology and media sciences. The psychological research question studied is whether and how violent content in computer games may induce aggression. Therefore, novel semantic concepts, most notably “violence”, are detected in computer game videos to gain insights into the interrelationship of violent game events and the underlying neurophysiologic basis (brain activity) of a player. The application of semantic concept detection to novel domains and novel concepts is presented in Chapter 6.

## 1.4 Publications

In the context of the research leading to this thesis, the following papers have been published:

- M. Mühling, R. Ewerth, J. Zhou, and B. Freisleben. Multimodal Video Concept Detection via Bag of Auditory Words and Multiple Kernel Learning. In *Proceedings of the 18th International Conference on Advances in Multimedia Modeling (MMM'12)*, pages 40–50, Klagenfurt, Austria, 2012. Springer

- R. Ewerth, K. Ballafkir, M. Mühling, D. Seiler, and B. Freisleben. Long-Term Incremental Web-Supervised Learning of Visual Concepts via Random Savannas. *IEEE Transactions on Multimedia*, 14(4):1008–1020, 2012

- M. Mühling, R. Ewerth, and B. Freisleben. On the Spatial Extents of SIFT Descriptors for Visual Concept Detection. In *Proceedings of the 8th International Conference on Computer Vision Systems (ICVS'11)*, pages 71–80, Sophia Antipolis, France, 2011b. Springer

- M. Mühling, R. Ewerth, B. Shi, and B. Freisleben. Multi-Class Object Detection with Hough Forests Using Local Histograms of Visual Words. In *Proceedings of 14th International Conference on Computer Analysis of Images and Patterns (CAIP'11)*, pages 386–393, Seville, Spain, 2011c. Springer

- R. Ewerth, M. Mühling, and B. Freisleben. Robust Video Content Analysis via Transductive Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):1–26, 2011

- P. Abend, T. Thielmann, R. Ewerth, D. Seiler, M. Mühling, J. Döring, M. Grauer, and B. Freisleben. Geobrowsing the Globe: A Geovisual Analysis of Google Earth Usage. *Linking GeoVisualization with Spatial Analysis and Modeling (GeoViz)*, 2011

- M. Mühling, K. Ballafkir, R. Ewerth, and B. Freisleben. University of Marburg at TRECVID 2011: Semantic Indexing. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'11)*, Gaithersburg, Maryland, USA, 2011a. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm

- M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, and B. Freisleben. University of Marburg at TRECVID 2010: Semantic Indexing. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, Gaithersburg, Maryland, USA, 2010. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm

- M. Mühling, R. Ewerth, and B. Freisleben. Improving Semantic Video Retrieval via Object-Based Features. In *Proceedings of the 3$^{rd}$ IEEE International Conference on Semantic Computing (ICSC'09)*, pages 109–115, Berkeley, CA, USA, 2009a. IEEE

- M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, and B. Freisleben. University of Marburg at TRECVID 2009: High-Level Feature Extraction. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'09)*, Gaithersburg, Maryland, USA, 2009b. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm

- D. Seiler, R. Ewerth, S. Heinzl, T. Stadelmann, M. Mühling, B. Freisleben, and M. Grauer. Eine Service-Orientierte Grid-Infrastruktur zur Unterstützung medienwissenschaftlicher Filmanalyse. In *Proceedings of the Workshop on Gemeinschaften in Neuen Medien (GeNeMe'09)*, pages 79–89, Dresden, Germany, Sept. 2009

- M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, and B. Freisleben. University of Marburg at TRECVID 2008: High-Level Feature Extraction. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'08)*, Gaithersburg, Maryland, USA, 2008. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm

- M. Mühling, R. Ewerth, T. Stadelmann, B. Freisleben, R. Weber, and K. Mathiak. Semantic Video Analysis for Psychological Research on Violence in Computer Games. In *Proceedings of the 6$^{th}$ ACM International Conference on Image and Video Retrieval (CIVR'07)*, pages 611–618, Amsterdam, The Netherlands, July 2007a. ACM

- M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, C. Zöfel, and B. Freisleben. University of Marburg at TRECVID 2007: Shot Boundary Detection and High-Level Feature Extraction. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'07)*, Gaithersburg, Maryland, USA, 2007b. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm

- R. Ewerth, M. Mühling, and B. Freisleben. Self-Supervised Learning of Face Appearances in TV Casts and Movies. *International Journal of Semantic Computing*, 1(2):185–204, 2007a

- R. Ewerth, M. Mühling, T. Stadelmann, J. Gllavata, M. Grauer, and B. Freisleben. Videana: A Software Toolkit for Scientific Film Studies. In *Proceedings of the International Workshop on Digital Tools in Film Studies*, pages 1–16, Siegen, Germany, 2007b. Transcript Verlag

- R. Ewerth, M. Mühling, and B. Freisleben. Self-Supervised Learning of Face Appearances in TV Casts and Movies. In *Proceedings of the 8th IEEE International Symposium on Multimedia (ISM'06)*, pages 78–85, Washington, DC, USA, 2006a. IEEE

- R. Ewerth, M. Mühling, T. Stadelmann, E. Qeli, B. Agel, D. Seiler, and B. Freisleben. University of Marburg at TRECVID 2006: Shot Boundary Detection and Rushes Task Results. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'06)*, Gaithersburg, Maryland, USA, 2006b. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm

## 1.5 Organization of this Thesis

This thesis is organized as follows:

Chapter 2 introduces the fundamentals of semantic concept detection and presents a general overview of important and related work from the literature. The question what is a semantic concept is addressed and existing concept lexicons (ontologies) for image and video retrieval are introduced. A general scheme of concept detection systems is described, just as important visual feature representations and predominant machine learning algorithms. Furthermore, performance measures and international image and video retrieval benchmarks are presented and their importance for the progress in semantic concept detection is emphasized.

In Chapter 3, different extensions to the BoVW concept detection approach are presented. First, the impact of the spatial extents of SIFT descriptors is investigated. Second, a multi-modal concept detection system is introduced, whereby the Bag-of-Words approach is leveraged for audio features and MKL is proposed as the appropriate fusion scheme for these BoAW and state-of-the-art BoVW features. Finally, a method for the SIFT-based integration of color information, called color moment SIFT (CMSIFT), is suggested.

Chapter 4 deals with novel features based on object detection results. Extensive experiments on the *Mediamill*, *VOC* and *TRECVid Challenge* show the usefulness of the proposed object-based feature representations. Furthermore, the generalization capabilities of concept models based on object-based features are investigated in a cross-domain setting and a method for multi-class object detection based on Hough forests is suggested to speedup the concurrent detection of multiple object classes.

Chapter 5 pays attention to the idea of using image and video data available in the World Wide Web (WWW) as training data for classifier construction. A novel incremental and scalable web-supervised learning system is presented that continuously learns concept models for image categories with heterogeneous appearances. New concept models are learned by just passing the name of the concept to the proposed system.

Chapter 6 deals with the application of concept detection approaches in the field of psychology and media sciences. A computer games video content analysis system is built to support psychological research by automatically detecting violent game events. Furthermore, the software toolkit `Videana` has been extended by a video retrieval component to support efficient search and navigation in large video databases.

Chapter 7 concludes this thesis and outlines areas of future work.

*"What makes us trouble is not what we don't know*
*it's just that, what we know for sure."*

Mark Twain

# 2 Fundamentals

## 2.1 Introduction

The task of visual concept detection belongs to the field of image and video analysis, which is rooted in the computer vision community. Traditionally, computer vision deals with image and video processing and includes algorithms like edge detection, image normalization or compression techniques. Due to the large amount of data that has to be processed and the associated high hardware requirements, only in recent years the computer vision community payed great attention to image and video analysis and retrieval.

Concept detection systems automatically analyze and tag images and videos with semantic information on the scene content, so that the resulting index can serve as a basis for content-based search. In this way, users are enabled to search images and videos via textual queries. To support arbitrary search queries, the concept lexicon has to cover a wide range of categories. Therefore, researchers have shifted their attention to generic concept detection systems, since the development of specialized detectors for hundreds or thousands of concepts seems to be infeasible.

Besides facilitated search and navigation, the annotation of videos with semantic concepts could be, together with the results of scene segmentation and person indexing, the basis for video abstracting [Lienhart et al. 1997] or automatic storyline extraction.

In recent years, the main progress in the area of visual concept detection is driven by enhancements in the feature extraction stage. Due to the large visual

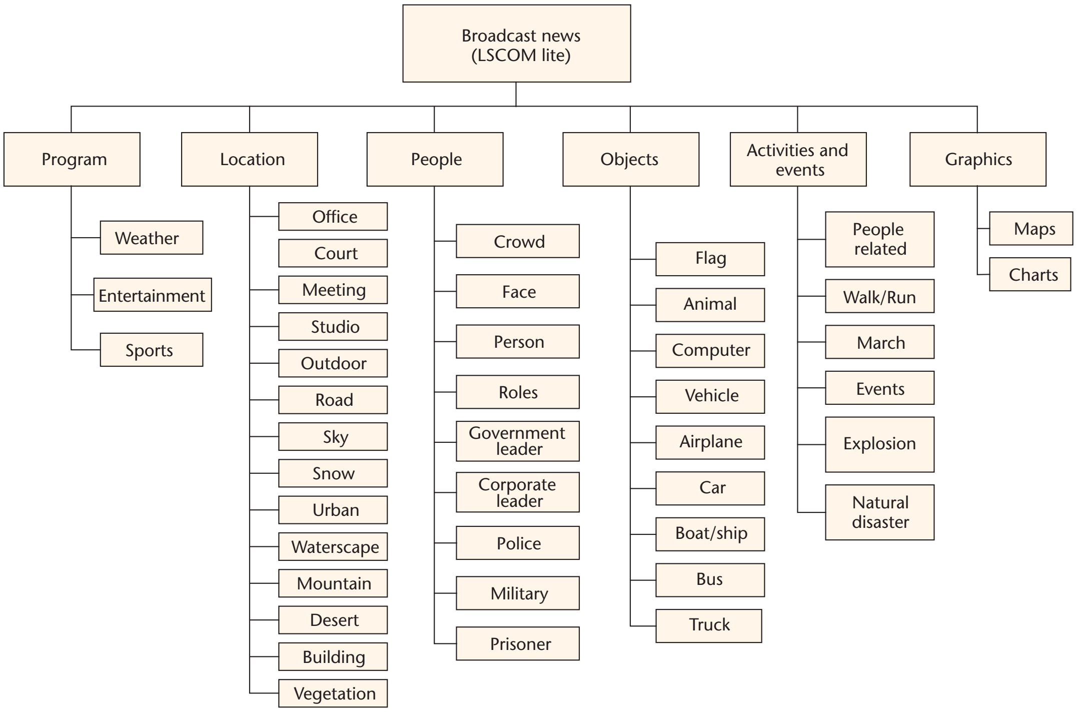


Figure 2.1: LSCOM-lite concepts, source [Naphade et al. 2006].

variations in the appearance of semantic concepts, current approaches mainly focus on local visual features based on keypoints, like SIFT descriptors or speeded-up robust features (SURF) [Bay et al. 2008; Lowe 2004]. The commonly used SIFT descriptors achieve top performances in visual recognition tasks. Based on these local descriptors, an image or a video shot is represented, similar to the representation of documents in the field of text retrieval, as a BoVW.

In this chapter, a survey of fundamental methods and research work related to generic concept detection in images and videos is provided. The remainder of this chapter is organized as follows: Section 2.2 focuses on the definition of semantic concepts. In Section 2.3 the general procedure for the task of concept detection is introduced. The focus of Section 2.4 is directed to BoVW approaches describing its main components, including SIFT as the most prominent local image descriptor, vocabulary generation, and coding and pooling strategies. The predominant SVM machine learning algorithm as well as MKL are described in Section 2.5. The difference between inductive and transductive learning is explained in Section 2.6. The publicly available image and video benchmarks used within this thesis are presented in Section 2.7. Last but not least, performance metrics for evaluation purposes of concept detection systems are defined in Section 2.8. Finally, Section 2.9 summarizes the chapter.

## 2.2 Semantic Concepts

Generic concept detection systems are able to build models for arbitrary concept classes such as objects, sites, scenes or events. The visual complexity and variability of concepts ranges from concepts like “face” or “sunset” to very complex concepts like “violence” or “baking a cake”. There has been a fruitful discussion in the research community about reasonable semantic concepts for tagging images and videos. As already mentioned in previous sections, the underlying concept lexicon has to cover a wide range of categories to finally support arbitrary search queries. The choice of meaningful concepts for multimedia data has been exhaustively discussed by the *Large-Scale Concept Ontology for Multimedia* (*LSCOM*) project [Naphade et al. 2006]. People from many different communities, such as computer vision researchers, ontology experts as well as end users, have been involved in the concept selection process. The final aim of this initiative was the development of a standard concept lexicon of about 1,000 semantic concepts. An intermediate result of these efforts for broadcast news videos was the *light concept ontology for multimedia* (*LSCOM-lite*) concept set [Naphade et al. 2005], which was developed in conjunction with the *TRECVid* benchmark [Smeaton et al. 2006]. Based on a case study on news videos [Gans 1980], the semantic concept space was divided into seven categories: program, setting/scene/site, people, objects, activity, event, and graphics. The choice of the concepts was influenced by analyzing and mapping search terms of BBC query logs to the WordNet hierarchy. To cover a preferably wide range of semantic space, the concepts have been chosen from different categories. Finally, the LSCOM-lite concept set comprises 39 semantic concepts (Figure 2.1) and is the most commonly applied visual concept lexicon in the literature.

In the further course of the project, concept selection criteria have been determined. The main criteria are the utility or usefulness for search queries, the observability by humans and the feasibility in the sense of automatic detection performance. Not all LSCOM-lite concepts were carried over into the LSCOM vocabulary. The current version of the LSCOM lexicon contains 856 concept definitions.

## 2.3 General Concept Detection Scheme

An early discussion about generic concept detection approaches with respect to news videos has been provided by Naphade and Smith [2004]. The authors state that in most approaches concept detection is considered as a supervised pattern recognition problem. The general concept detection scheme is illustrated in Figure 2.2. The concept models are learned from a manually labeled set of training images or video shots. An image cannot be just categorized into one of the concept classes, instead several concepts may co-occur in an image or video

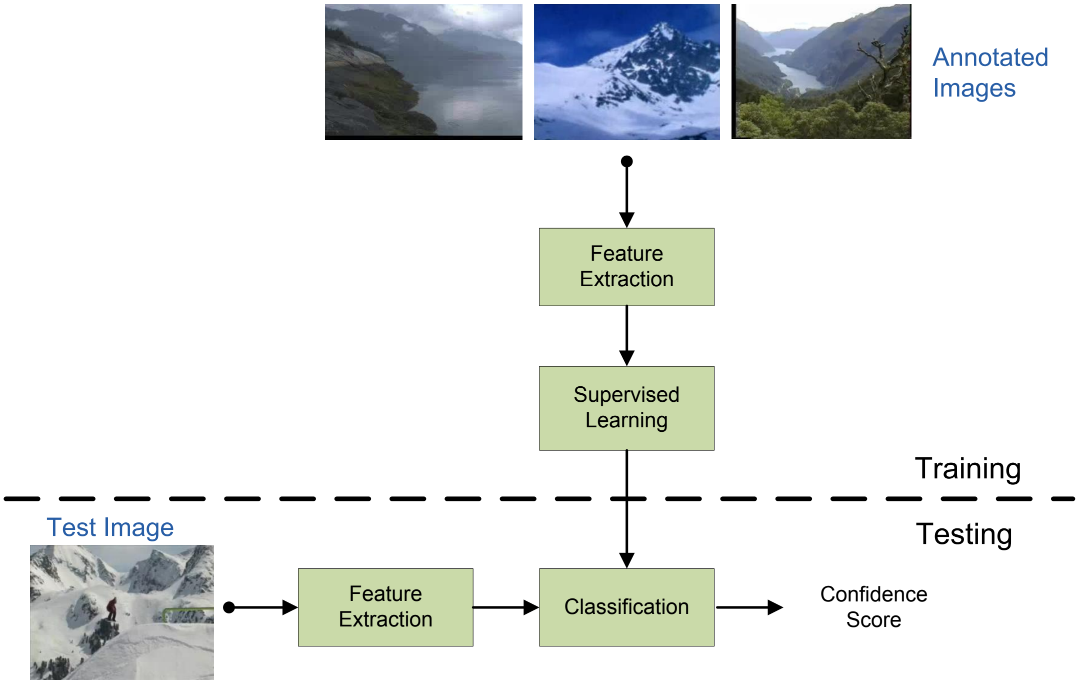


Figure 2.2: Semantic concept detection as a supervised classification problem.

shot. Thus, concept detection is generally regarded as a multi-class multi-label problem. This problem is typically broken down into binary classification tasks. Therefore, a classification model is build for each concept class in a discriminative setting by using positive and negative training instances. In a first step, features, for example, color histograms or BoVW features, are extracted from the labeled training images. These feature representations are fed together with the corresponding class labels into a machine learning algorithm. The most prominent and successful classifier in the field of concept detection is the SVM algorithm. The result of the learning process is a concept model that is used to classify unknown test instances. In the classification stage, a test image passes through the same feature extraction process as the training images. The resulting feature vector is classified using the previously trained concept model. SVM-based classifiers return a confidence score, indicating the presence of the corresponding concept class in an image or video shot.

For videos, different levels of temporal resolution can be used for indexing. In an early automatic movie content analysis project (the MoCA project [Lienhart et al. 1996]), the levels “frame”, “shot”, “scene” and “video” were identified [Lienhart et al. 1999a]. A (camera) shot is a sequence of frames of a non-interrupted camera recording. The transitions between consecutive shots are realized by cuts and dissolves, where cuts are the most frequently used transition. Today, the task of automatic shot segmentation is considered to be solved [Smeaton et al. 2006]. The next level of temporal resolution are “scenes”. A scene consists of one or

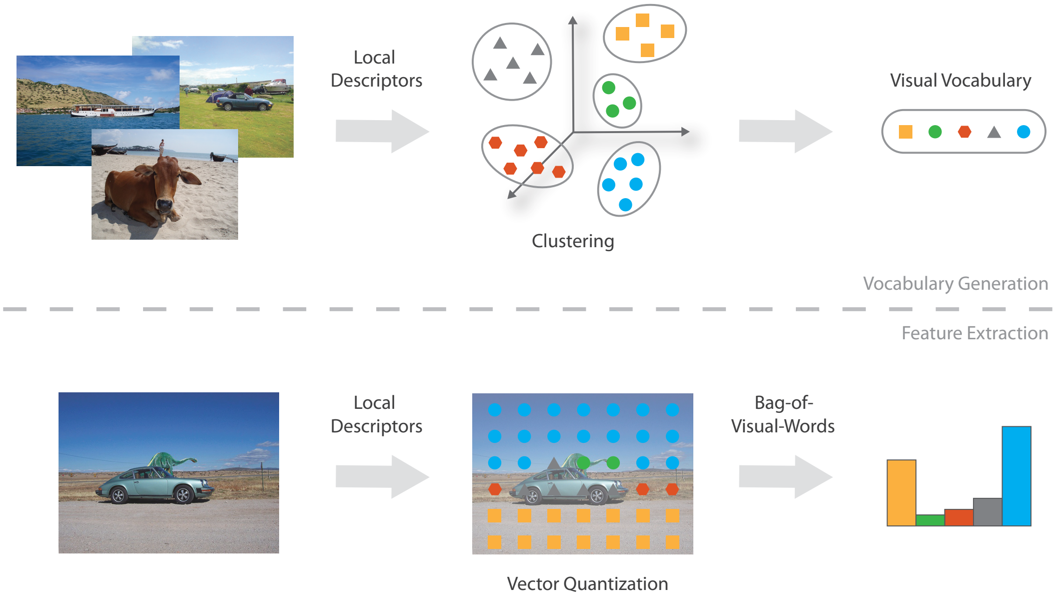


Figure 2.3: Bag-of-Visual-Words image representation.

more shots related in terms of time, space and/or content. Lienhart et al. [1999b] revealed dialogs, consecutive shots of the same setting and continuing sounds as reliable indicators for scene boundaries. However, the task of scene segmentation is rather difficult due to high-level semantic relations, subjectivity and a missing precise definition of the term "scene". Therefore, concept detection is generally based on shot segmentation which is used as a preprocessing step for semantic video indexing.

Video retrieval systems mostly rely on the analysis of so called keyframes, which are the representative images of a video shot. The frame in the middle of a shot is often used as the representative image for analysis and visualization purposes. Hence, image classification approaches are easily extensible to the video domain. Most of the features used for video retrieval are related to the visual modality. Popular and commonly used features are, e.g., in the context of MPEG-7: global and local color histograms, texture features, edge features, color moments, Gabor wavelet features [Manjunath et al. 2001], and motion features [Jeannin and Mory 2000]. Except for the last feature type, these features can be used for video as well as for image analysis and retrieval purposes.

## 2.4 Bag-of-Visual-Words

The main progress in recent years in the field of visual concept detection is due to improvements in the feature extraction stage. These improvements are inspired by text retrieval and rely on the extraction of region-based image descriptors. Similar to the representation of documents in the field of text retrieval, an image or a video shot can be represented as a bag of visual words. These visual words are the result of a vector quantization process. Therefore, local image features, like for example SIFT descriptors, are extracted and assigned to a given visual vocabulary. In a preprocessing step this vocabulary, also called codebook, is generated from a large set of local image descriptors from a wide choice of training images by clustering the extracted keypoint descriptors in their feature space. The resulting cluster centers are regarded as visual words.

Using this vocabulary of visual words, an image or a video shot can be represented as a BoVW by mapping the local descriptors to the vocabulary. Finally, the visual words are assembled to Histograms of Visual Words (HoVW) by counting the frequency of each visual word in the image. The terms BoVW and HoVW are often used interchangeably in the literature. The process of vocabulary as well as histogram generation are visualized in Figure 2.3.

Using the BoVW approach, continuous progress has been reported in recent years. The top runs at the semantic indexing task of the *TRECVid Challenge* from 2010 to 2012 heavily rely on BoVW representations [Over et al. 2011, 2012, 2013].

Comprehensive comparisons concerning representation choices of keypoint-based concept detection approaches have been provided by Jiang et al. [2007, 2010a]. They evaluated various factors of the BoVW representation for semantic video retrieval including the choice of keypoint detector, kernel function, vocabulary size as well as the integration of spatial information. In another study, Yang et al. [2007a] applied techniques commonly used in text categorization, including term weighting, stop word removal or feature selection to generate image representations for scene classification that differ in dimension, selection, and weighting of visual words.

The main components of the BoVW approach are discussed in the following subsections: Local image features, especially SIFT as the most prominent local descriptor, are introduced in Section 2.4.1. The codebook generation process is described in Section 2.4.2. Section 2.4.3 presents coding and pooling strategies. Super-vector coding is explained in Section 2.4.4. Section 2.4.5 addresses the issue of how to integrate spatial information.

### 2.4.1 Local Image Features

Recently, there is a tendency towards using scale- and rotation-invariant interest point representations, such as SIFT [Lowe 2004] or SURF features [Bay et al. 2008]. The extraction of local image features can be divided into two main components: the image sampling strategy and the construction of the keypoint descriptors. The SIFT algorithm published by Lowe [2004], as the most prominent and successful local image descriptor, describes both: an interest point detection

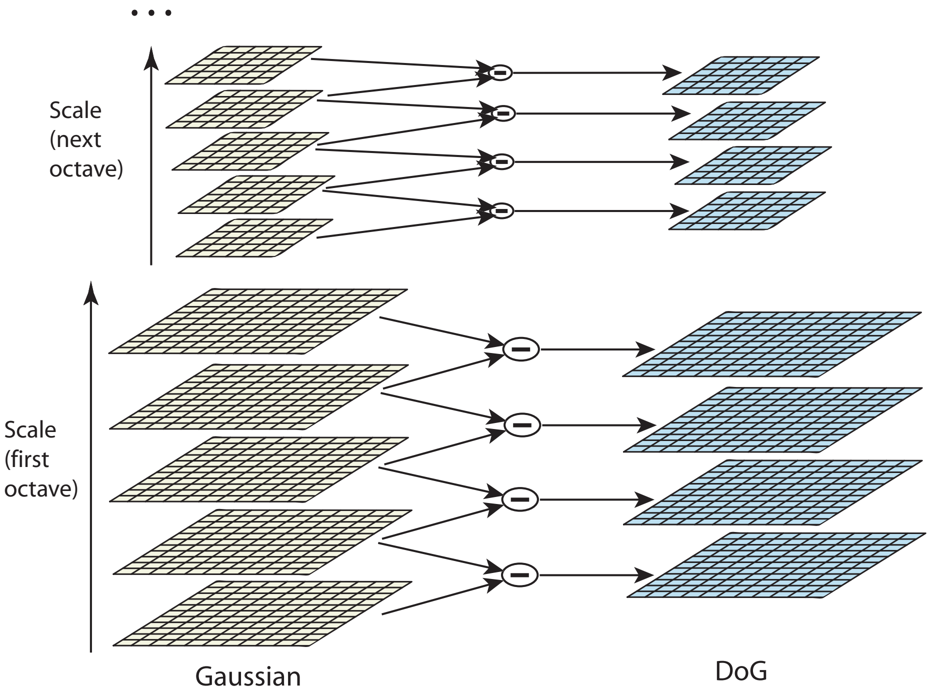


Figure 2.4: Image scale-space and Difference of Gaussian pyramid, source [Lowe 2004].

and a local feature extraction method. Besides scale- and rotation invariance, SIFT features are robust against noise, illumination changes and small geometric deformations. The underlying ideas of the keypoint detection approach as well as the local descriptor extraction method of the SIFT algorithm are introduced in the following paragraphs in more detail.

### Image Sampling Strategy

Interest point detection is often used as a preprocessing step in computer vision algorithms, like epipolar geometry, object tracking or BoVW approaches. It is closely related to the traditional field of corner detection. The terms “interest point”, “salient point” and “keypoint” are often used interchangeably in the literature. A survey of local scale- and rotation-invariant interest point detectors is given by Tuytelaars and Mikolajczyk [2008]. The problem of interest point detection is somehow similar to the problem of salient region detection. In the case of stereo videos, even depth information can be used for saliency detection [Dittrich et al. 2013].

In this thesis, “interest points” are considered as robust and well-defined points of an image, which are stable under affine or even perspective image transformations, while “keypoints” are used as a generic term including sampled or randomly chosen points.

Interest points describe salient image regions, which contain the “most important” content of an image. In order to provide reproducibility with respect to scale changes, images are analyzed at multiple scales. Therefore, the scale-space of an image is constructed and interest points are selected at locations where the image gradient magnitude yields a local maximum in the three-dimensional scale-space. Often used interest point detectors are the Harris-Laplace and Hessian-Laplace detector [Mikolajczyk and Schmid 2001], which use a Laplacian of Gaussians (LoG) to find stable interest point locations in the image scale-space. The SIFT algorithm suggests a Difference of Gaussian (DoG) detector to approximate the LoG algorithm and to accelerate the computations. For this purpose, a DoG pyramid (Figure 2.4) is constructed

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \tag{2.1}$$

where $G$ is the Gaussian scale function

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp^{-(x^2+y^2)/2\sigma^2}, \tag{2.2}$$

$\sigma$ and $k\sigma$ are two nearby scales separated by a constant factor $k$ and $*$ is the image convolution operator.

In a first step, scale-space extrema are detected in the DoG pyramid by searching over all image locations and scales. Second, the rough coordinates of these interest point candidates are further refined by fitting a three-dimensional quadratic function to determine the exact interpolated location and scale. Due to stability criteria, interest points with low contrast or along edges are eliminated. While the scale-invariance is realized by searching for scale-space extrema in the DoG pyramid, the rotation-invariance of a subsequently extracted local interest point descriptor is achieved by determining the dominant direction of the local image gradient. Hence, an interest point is given by its image coordinates, scale and orientation.

While early BoVW approaches mainly extracted local descriptors at interest points using detectors like Harris-Laplace or DoG, today it seems that this representation is often insufficient to describe natural images. For scene classification, random or dense sampling strategies have outperformed the previously mentioned scale- and rotation-invariant interest point detectors [Bosch et al. 2008; Jurie and Triggs 2005; Nowak et al. 2006]. The dense sampling strategy uses keypoints on a rectangular grid, where the image is sampled at a fixed step size.

### Keypoint Descriptors

Keypoint descriptors are used to describe localized image regions. The extraction of a local interest point descriptor is performed relative to the orientation, scale and location of an interest point. While local descriptors for interest points are invariant to these transformations, local image descriptors for dense or randomly

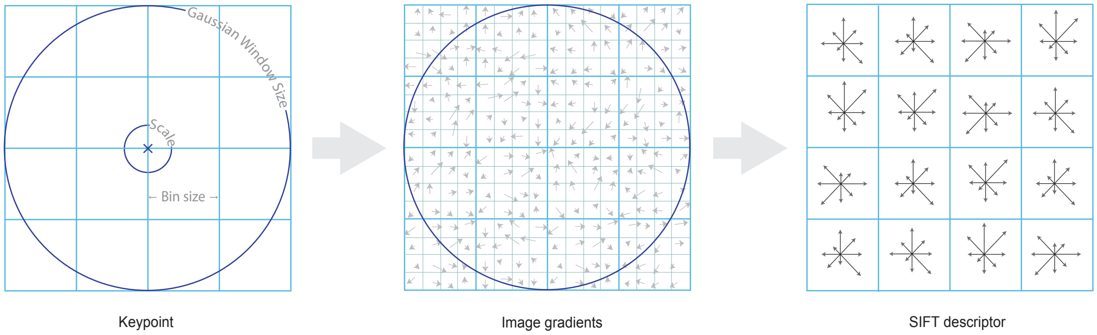


Figure 2.5: SIFT descriptor extraction, based on Lowe [2004].

sampled keypoints are extracted at fixed scales and orientations. In this case the scale and rotation invariance is partly abandoned. It seems that for scene classification these descriptors using a fixed orientation contain valuable information about the image content.

SIFT, as the most successful local descriptor, describes the appearance of a keypoint using a spatial histogram of image gradients, where a Gaussian weighting function is applied to reduce the influence of gradients further away from the keypoint location (Figure 2.5). The SIFT descriptor geometry is specified by the number and size of the spatial bins and the number of orientation bins. The size of the spatial bins results from the multiplication of the keypoint scale with a predefined magnification factor. The final keypoint descriptor consists of the concatenated gradient histograms for the spatial bins. Using eight orientation bins and 4x4 spatial bins results in a 128-dimensional feature vector. These local histograms are visualized in Figure 2.5 as “stars”, whereby the length of the arrows characterizes the amount of gradients in the corresponding direction.

In contrast to SIFT, the SURF algorithm uses distributions of Haar wavelet filter responses instead of gradient histograms. The use of integral images further speeds up the computation of SURF descriptors and clearly improves the runtime performance in comparison to the SIFT algorithm.

### 2.4.2 Visual Vocabulary

In the feature extraction stage of the BoVW approach, the local keypoint descriptors of an image are assigned to the visual words of a predefined vocabulary. An image or video shot is represented by the distribution of these words. The visual vocabulary, also called codebook, has to be constructed in a preprocessing step. Its visual words are a set of prototype vectors from the continuous feature space of local keypoint descriptors.

In the literature, different approaches for the construction of visual vocabularies have been investigated. These approaches can be roughly divided into generative and discriminative methods.

### Generative Codebooks

In general, local keypoint descriptors from a set of randomly selected training images are clustered in their feature space using the K-means algorithm, regarding the resulting cluster centers as visual words. To speed up the codebook generation process, hierarchical clustering algorithms, like hierarchical K-means have been used. Also, self-organizing maps (SOM) have been applied for codebook construction. But despite a clearly smaller quantization error, SOM-based codebooks yielded clearly worse classification results [Viitaniemi and Laaksonen 2008].

Instead of an universal codebook, it is also possible to build concept or class-specific codebooks. But, the weak improvement using class-specific codebooks is not worth the enormous additional memory and runtime requirements for building seperate codebooks and class-specific feature representations [Viitaniemi and Laaksonen 2008]. Slight performance improvements are achieved by merging small class-wise clustered codebooks [Wojcikiewicz et al. 2010], but this strategy does not scale for a large number of concept classes.

Jurie and Triggs [2005] stated that for densely sampled image descriptors K-means-like cluster algorithms lead to a concentration of cluster centers in high density regions. Discriminative codewords in low density regions are therefore under-represented. They proposed a simple alternative based on mean shift, called fixed radius clustering. Like K-means, the codebook is built upon a random subset of local keypoint descriptors from a wide choice of training images. The fixed radius clustering algorithm produces an ordered list of cluster centers, whereby the mean shift algorithm is used to find maximum density positions. The algorithm starts with the maximum density position of all descriptors as the first cluster center. Then, all descriptors within a fixed radius of this center are removed and the next codeword is determined as the maximum density region of the remaining descriptors. The procedure continues until the desired size of the codebook is reached. Jurie and Triggs [2005] showed that codebooks based on the fixed radius clustering algorithm outperformed K-means-based codebooks on an object detection task.

Nowak et al. [2006] revealed that even randomly selected codewords produce very respectable results. Although the K-means algorithm produced the best performing codebooks, Viitaniemi and Laaksonen [2008] confirmed that in the case of large codebook sizes a simple random selection of codewords achieves similar results.

### Discriminative Codebooks

Besides unsupervised codebook generation methods, discriminatively trained codebooks have also been investigated. Moosmann et al. [2006] used a random forest extension, called *extremely randomized clustering forest*, to construct a fast discriminative codebook. Based on the class label purity, the *Shannon entropy* is used as optimization criterion during tree construction. The leaf nodes of the decision trees in the random forest represent the visual words. To yield the desired vocabulary size the decision trees are pruned by recursively removing the leaf node with the lowest information gain. During histogram generation, local feature descriptors are propagated through the trees of the random forest and increase the corresponding word count at the leaf node. In their experiments, the random forest based codebook outperforms the K-means algorithm. But the experiments are conducted on a very small dataset with only four object classes and the results are compared to a traditional (non-state-of-the-art) BoVW approach using K-means in combination with hard vector quantization.

Viitaniemi and Laaksonen [2008] used a learning vector quantization algorithm to include class-label information into the codebook generation process. But the resulting discriminative codebook performed even worse than randomly selected codewords.

Semantic meaningful codewords, representing, for example, grass or sky, have been used by Gemert et al. [2010a], who investigated codeword selection methods for compact codebooks. However, semantic codebooks did not lead to an improved concept detection performance and missed the expected success.

### 2.4.3 Coding and Pooling

Traditional BoVW approaches assigned continuous local image features to discrete visual words by mapping the keypoint descriptors to their nearest neighbors in the visual vocabulary. Especially in the case of a small vocabulary, this procedure is susceptible for quantization loss, because similar keypoints can be assigned to different visual words. Furthermore, two keypoints assigned to the same visual word do not need to be equally similar to that word due to their distances to the cluster center. Ignoring these distances discards valuable information about the importance of a visual word. Jiang et al. [2007] proposed a *soft-weighting scheme* where a keypoint is assigned to multiple visual words and the importance is determined by the similarity of the keypoint to the visual word. Instead of mapping a keypoint only to its nearest neighbor, the top-$k$ nearest visual words are selected. Using a vocabulary of $N$ visual words, an image is represented by the weights of a histogram $w = [w_1, \dots, w_t, \dots, w_N]$ where the importance of a visual word $t$ is given by

$$w_t = \sum_{i=1}^{k} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(j,t) \tag{2.3}$$

with $M_i$ being the number of keypoints whose $i$-th nearest neighbor is the visual word $t$.

In an in-depth analysis of this visual word ambiguity, Gemert et al. [2010b] distinguished between *kernel weighted* and *constant weighted approaches*. Kernel weighted approaches use a similarity function, also called *kernel.* This function computes the similarity between a local descriptor and the codebook candidates based on the Euclidean distance $d$:

$$sim(i,j) = exp(-\gamma \cdot d(i,j)) \tag{2.4}$$

The previously mentioned soft-weighting scheme of Jiang et al. [2007] falls into this category.

Constant weighted approaches ensure an equal contribution of each local descriptor to the resulting HoVW. This is achieved by normalizing the sum of the weights to the codebook candidates to 1. The traditional BoVW approach using hard vector quantization is a constant weighted approach, because for each descriptor the corresponding codeword accumulator is incremented by a constant value.

Furthermore, a distinction is made between approaches using only the best codebook candidate for a local keypoint descriptor and approaches where multiple candidates are considered.

Let N be the size of the visual vocabulary, $w = [w_1, \ldots, w_t, \ldots, w_N]$ the resulting feature vector and K the number of local image descriptors, the traditional hard vector quantization approach can be expressed as

$$w_t = \sum_{i=1}^{K} \begin{cases} 1 & \text{if } t = \operatorname*{arg\,min}_{j=1,\ldots,N} d(i,j) \\ 0 & \text{otherwise} \end{cases} . \tag{2.5}$$

The extension of the traditional vector quantization method to multiple codeword candidates is called *codeword uncertainty* approach, whereby the constant weight is proportionately distributed among the codebook entries:

$$w_t = \sum_{i=1}^{K} \frac{sim(i,t)}{\sum_{j=1}^{N} sim(i,j)} . \tag{2.6}$$

Kernel weighted approaches are divided into *codeword plausability* approaches using only the best candidate and *kernel codebook* approaches for multiple codebook candidates [Gemert et al. 2008]. While the codeword plausibility method is expressed as

$$w_t = \sum_{i=1}^{K} \begin{cases} sim(i,t) & \text{if } t = \operatorname*{arg\,min}_{j=1,\ldots,N} d(i,j) \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

the kernel codebook formulation considers the similarity of a local descriptor to all codebook entries:

$$w_t = \sum_{i=1}^{K} sim(i,t). \tag{2.8}$$

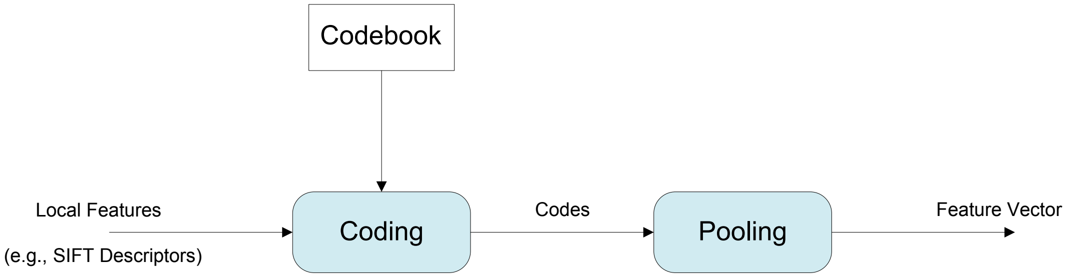


Figure 2.6: Main BoVW components: feature coding and pooling.

It should be noticed that in contrast to the soft weighting scheme of Jiang et al. [2007], Gemert et al. [2010b] accumulated similarities from a local descriptor to all codebook entries. A disadvantage of this strategy is that the accumulation of many small similarities can lead to codeword activations that are misleading for the classification process. It has been shown, that the restriction to the top-$k$ nearest neighbors achieves superior concept detection performance [Liu et al. 2011].

Yang et al. [2009] identified two main components of the BoVW feature extraction method: coding and pooling. While the coding step transforms the local keypoint descriptors of an image to codes, which in the case of hard vector quantization correspond to the visual words, the pooling step summarizes the bag of codes to a final feature vector.

Yang et al. [2009] proposed an alternative method using *sparse coding* instead of vector quantization and *max-pooling* instead of average pooling. While the generally used average pooling strategy sums up the weights for each codebook entry, *max-pooling* remembers only the largest weight (similarity) per codeword. In this case, the resulting feature vector indicates for each visual word the probability of occurring at least once in the image. Yang et al. [2009] showed that this strategy in combination with a spatial pyramid representation is clearly better suited for learning and classification with linear SVMs. In the case of sparse coding, local features like SIFT descriptors are reconstructed using a linear combination of codebook entries, whereby the number of nonzero coefficients is forced to be small. Thus, each local feature descriptor is represented by a sparse code with a small number of weights assigned to codebook entries.

Yu et al. [2009] observed that sparse coding algorithms mostly select codebook entries from the local feature space neighborhood of a given example for its reconstruction. Since a locality constraint automatically leads to sparsity, they introduced a modified sparse coding approach called *Local Coordinate Coding* and replaced the sparsity constraint of the underlying optimization problem by a penalty term for distant codebook entries. Despite these improvements, the feature extraction process based on sparse coding is computationally more complex than vector quantization. To speed up the sparse coding process, Wang et al. [2010] suggested a fast implementation of Local Coordinate Coding called

*Locality-constrained Linear Coding.* In an approximated locality-constrained linear coding method, a k-nearest-neighbor search is performed. Then, these neighbors are used as a reduced set of local base vectors for the reconstruction. Thus, a much smaller linear system has to be solved and the runtime complexity for the reconstruction of a local descriptor is reduced from $O(M^2)$ to $O(M + K^2)$, where $M$ is the number of codebook entries. Together with an incremental codebook optimization and a max-pooling strategy, this approach achieved top image classification performance even with linear SVMs.

Boureau et al. [2010] identified the max-pooling strategy, which selects the largest coefficient for a visual word, as responsible for superior image classification performance especially in the case of linear SVMs.

As an extension to the max-pooling strategy Liu et al. [2011] proposed a *mix-order max-pooling* strategy, which estimates the probability of a visual word of being k-times present in an image. The mix-order max-pooling strategy achieved similar performance compared to the max-pooling approach.

The kernel weighted approach, also called soft assignment coding, which has a high computational efficiency, has been revisited by Liu et al. [2011]. They argue that the inferior performance of soft-assignment coding is often caused by employing the membership to all codewords. In contrast, it is beneficial to only use the k-nearest visual words for coding a local descriptor.

### 2.4.4 Super-Vector Coding

Beyond the coding schemes, as described in the previous section, super-vectors [Inoue and Shinoda 2012; Zhou et al. 2010] and the closely related Fisher vectors [Perronnin et al. 2010] can be used to encode local image descriptors. The idea of super-vectors was born in the field of speaker identification [Campbell et al. 2006] and first applied to image classification by Zhou et al. [2010].

Using the expectation maximization algorithm a *Gaussian mixture model* (GMM) is build from randomly selected local keypoint descriptors of the training set. This model that can be regarded as a visual vocabulary describes the distribution of the overall data and is also called *universal background model* (UBM). A GMM consists of $K$ Gaussian components $\lambda_k = \{\omega_k, \mu_k, \Sigma_k\}$, where $\lambda_k$ is the $k$-th component with the weight $\omega_k$, the mean vector $\mu_k$ and the diagonal covariance matrix $\Sigma_k$. The probability of a local descriptor $x_i$ belonging to the $k$-th Gaussian component is given by:

$$c_{ik} = \frac{\omega_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \omega_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}. \tag{2.9}$$

where $\mathcal{N}$ is the normal distribution. Compared to soft assignment coding, super and Fisher vector approaches try to capture information about the fitting error between local keypoint descriptors and its nearest neighbors. Let $X =$

$\{x_1, \ldots, x_N\}$ a set of $N$ local image descriptors. Instead of summing similarities (probabilities) per cluster center to build a histogram, Zhou et al. [2010] encode the fitting error by calculating a weighted sum of difference vectors per GMM component:

$$\nu_k = \sum_{i=1}^{N} c_{ik} \left( x_i - \mu_i \right). \tag{2.10}$$

Since opposite deviations (difference vectors) assigned to the same cluster center (GMM component) cancel each other, Zhou et al. [2010] additionally encode the weights (similarities) $C_k = \sum_{i=1}^{N} c_{ik}$ according to soft assignment coding. Based on the UBM, the sum of weights as well as the weighted sum of difference vectors per component are normalized and combined as follows:

$$\phi(X) = \begin{pmatrix} S_1 \\ \nu_1 \\ \vdots \\ S_K \\ \nu_K \end{pmatrix} \quad \text{with} \quad S_k = \alpha \sqrt{C_k} \quad \text{and} \quad \hat{\nu}_k = \frac{1}{\sqrt{C_k}} \nu_k \tag{2.11}$$

where $\alpha$ is a balancing factor between the component-wise weights and the mean vectors. The combined vector can be regarded as an early fusion of a HoVW (component-wise weights) and the fitting error (mean difference vectors).

While Zhou et al. [2010] encoded the differences between local SIFT descriptors and GMM mean vectors, Fisher vectors additionally integrate the second order differences (squared difference vectors) to retain information that gets lost during mean vector calculation.

A slightly different strategy was followed by [Inoue and Shinoda 2012]. The underlying idea of their GMM super-vectors is to describe the local descriptors of an image or video shot by fitting a GMM. The similarity between two instances can then be determined based on the model parameters. They used a *maximum a posteriori* (MAP) adaption technique to fit the UBM to the local keypoint descriptors of an image or video shot. Instead of building differences between the mean vectors of the UBM model and the local descriptors, the local descriptors are assigned to the Gaussian mixtures in a soft manner and the mean vectors are adapted in the following way:

$$\hat{\mu}_k = \frac{\tau \mu_k + \sum_{i=1}^{N} c_{ik} x_i}{\tau + \sum_{i=1}^{N} c_{ik}} \quad \text{with} \quad c_{ik} = \frac{\omega_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \omega_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}. \tag{2.12}$$

The mean vectors of the adapted GMM are normalized $\tilde{\mu}_k = \sqrt{\omega_k} \left( \Sigma_k \right)^{-\frac{1}{2}} \hat{\mu}_k$ and concatenated to form the final feature vector, also called GMM super-vector. Here, the normalization is based on the weights and covariances of the UBM, while Zhou et al. [2010] weight the mean difference vectors based on the posterior probabilities. Using UBM and MAP adaption this approach permits, like the fisher vectors and the super vectors of Zhou et al. [2010] the comparison of the resulting vectors in the euclidean space. Thus, super-vectors can be efficiently compared without an expensive matching of Gaussian components, like for example by using the earth movers distance.

A disadvantage of super-vector approaches is that the resulting feature vectors are much larger and less sparse compared to the HoVW representation. The final size of the feature vectors is given by $D{\cdot}K$, where $D$ is the dimension of the local descriptors and $K$ is the codebook size, specifically the number of Gaussian components of the UBM. Applying spatial pyramids in addition the dimension is extraordinary high. Then, even for relatively small data sets like the *VOC Challenge* this leads to severe memory problems [Chatfield et al. 2011]. To counter this, *principal component analysis* (PCA) is typically applied to the local SIFT descriptors for dimensionality reduction. Additionally, the size of the codebook is kept relatively small. These steps using PCA and a relatively small codebook not only reduce the dimension of the final feature vector, but also result in an accelerated coding step. To further speedup the GMM super-vector coding approach, Inoue and Shinoda [2011] introduced a fast and accurate tree-structured MAP adaption, where a tree of Gaussian components is used to search efficiently for the approximate nearest neighbors (UBM components) of a local descriptor. The GMM super-vector approach using tree-structured MAP adaption leads to superior concept detection performance [Inoue et al. 2010b, 2012]. It is not mentioned but the performance boost is probably triggered by the restricted contribution of a local descriptor. Compared to the baseline GMM super-vector approach, a local SIFT descriptor in the tree-structured extension does not contribute to all mean vectors, but only to its k nearest Gaussian components. This strategy was already very successful in the field of soft assignment coding.

### 2.4.5 Spatial Information

In the following, two different strategies to integrate spatial information into BoVW image representations are presented: spatial pyramids and spatial coordinate coding. While the spatial pyramid representation concatenates HoVWs for image subregions, the second approach considers spatial coordinates already in the codebook generation process.

#### Spatial Pyramids

Lazebnik et al. [2006] suggested spatial pyramid representations for scene classification to integrate spatial information. This approach has been derived from the pyramid match kernel [Grauman and Darrell 2005], which is an efficient method to find a partial matching between two sets of keypoint descriptors. To build spatial pyramids BoVW representations are computed for equally sized image subregions of different partitioning levels. These regional representations are concatenated per partitioning level and finally fused using a weighted combination of kernel functions. It has been shown that the combination of multiple spatial layouts is helpful, whereas an image partitioning of more than 2x2 regions is ineffective [Sande et al. 2008; Yang et al. 2007a]. Some approaches additionally use horizontal or vertical divisions of 1x3 or 3x1 subregions for visual concept detection [Sande et al. 2010].

Another approach has been proposed by Krapac et al. [2011], where the spatial layout of an image is modeled by estimating the mean spatial location and the spatial variance for each visual word. Thus, for each visual word two additional feature values are considered. This representation has the advantage of being more compact compared to the spatial pyramid representation.

### Spatial Coordinate Coding

Recently, Koniusz and Mikolajczyk [2011] and Mbanya et al. [2011] proposed to append spatial coordinates at the descriptor level. Therefore, the local descriptor size is increased by two dimensions. This kind of representation limits the overhead introduced by considering spatial information and achieves almost comparable results to the spatial pyramid representation. The early fusion of a local visual descriptor and its spatial information, also called *spatial coordinate coding* [Koniusz and Mikolajczyk 2011], leads to an integration of spatial information into the visual dictionary, which is called *spatial codebook* by Mbanya et al. [2011]. A weighting factor in the distance function for the local descriptors determines the relevance the spatial component. Since an optimized weighting factor per concept did not improve the detection performance, Mbanya et al. [2011] used a global optimized weighting factor. Concept-specific weight optimization would result in concept-specific spatial codebooks and thus triggers large additional memory as well as computational requirements. Koniusz and Mikolajczyk [2011] introduced spatial coordinate coding for soft assignment as well as for sparse coding. They achieved the best results using the soft assignment approach with spatial coordinate coding.

## 2.5 Support Vector Machines

Visual concept detection is generally considered as a supervised learning task. SVMs, introduced by Vapnik [2000], are the predominant classifier in this field. They have proven to be very powerful for classifying semantic concepts [Smeaton et al. 2006]. In this section, the ideas and fundamentals of SVMs are introduced. Practical user guidelines for employing SVMs have been published by Hsu et al. [2010] and Ben-Hur and Weston [2011]. Further information about SVMs and

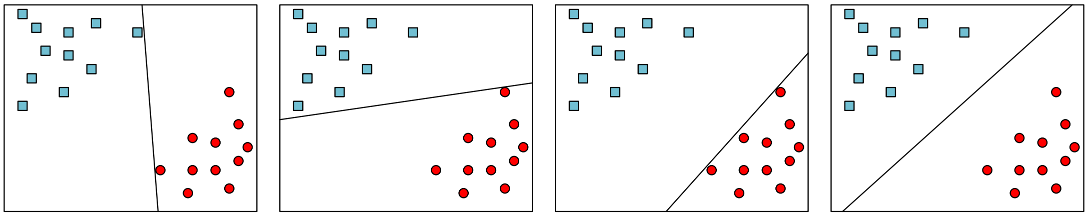


Figure 2.7: Potential separating hyperplanes for a given set of training instances.

the underlying statistical learning theory can be found in the publications of Cristianini and Shawe-Taylor [2000] and Bishop [2009].

Generally speaking, the idea of a SVM is to find a separating hyperplane between positive and negative training examples. In the following, the training data comprises $N$ labeled training examples given by $(x_1, y_1), \ldots, (x_N, y_N)$ where $x_i \in \mathbb{R}^d$ are the feature vectors and $y_i \in \{\pm 1\}$ are the given class labels. Based on the hyperplane, which is learned from the labeled training instances, unknown examples can be assigned to one of the two classes by the following discriminative function:

$$f(x) = \operatorname{sgn}(w^T \cdot x + b) \quad \text{with} \quad w \in \mathbb{R}^d \; and \; b \in \mathbb{R} \tag{2.13}$$

whereby the hyperplane is given by the equation $0 = w^T x + b$. The weight vector $w$ is perpendicular to the hyperplane and the bias $b$ describes the translation of the hyperplane to the origin. This representation is not unique because all pairs of $(a \cdot w, a \cdot b)$ with $a \in \mathbb{R} \setminus 0$ describe the same hyperplane.

Fortunately, the scaling of the parameters $w$ and $b$ does not change the distance of any point from the hyperplane, which is given by

$$\text{distance}_{w,b}(x_i) = \frac{|w^T x_i + b|}{\|w\|}. \tag{2.14}$$

This is important because it allows to scale the parameters $w$ and $b$ so that

$$\min_{i=1,\ldots,N} |w^T x_i + b| = 1 \tag{2.15}$$

which is called the *canonical representation* of the hyperplane.

First, the easiest case is considered where the training examples are linear separable. As shown in Figure 2.7 there are many possible solutions for a separating hyperplane. In order to minimize the generalization error, the idea is to maximize the distance of the closest point to the hyperplane. That is why SVMs are also called maximum-margin classifiers. The *margin* of a SVM is defined as the perpendicular distance of the closest point to the hyperplane. If the margin is maximized there will be at least two closest points and the margin of the canonical representation is given by $\frac{1}{\|w\|}$ (Figure 2.8). Thus, maximizing the margin is equivalent to the following optimization problem

$$\min_{w,b} \quad \frac{1}{2} w^T w \tag{2.16}$$

$$s.t. \quad y_i(w^T x_i + b) \geq 1 \quad \forall\, i = 1, \ldots, N.$$

In order to solve this problem, *Lagrange Multipliers* $\alpha = (\alpha_1, \ldots, \alpha_N)$ are introduced for the inequality constraints, leading to the following Lagrangian function:

$$L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i (y_i (w^T x_i + b) - 1) \tag{2.17}$$

$$s.t. \quad \alpha_i \geq 0 \quad \forall\, i = 1, \ldots, N.$$

By setting the derivatives of the *Lagrangian function* with respect to $w$ and $b$ to zero, the following two equations are obtained:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad \text{and} \quad 0 = \sum_{i=1}^{N} \alpha_i y_i. \tag{2.18}$$

These conditions allow the elimination of $w$ and $b$ from the Lagrangian function

$$\begin{aligned}
L(w, b, \alpha) &= \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i (y_i (w^T x_i + b) - 1) \\
&= \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i y_i w^T x_i - \sum_{i=1}^{N} \alpha_i y_i b + \sum_{i=1}^{N} \alpha_i \\
&= \frac{1}{2} w^T w - w^T \sum_{i=1}^{N} \alpha_i y_i x_i - b \sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i \\
&= \frac{1}{2} w^T w - w^T \sum_{i=1}^{N} \alpha_i y_i x_i + \sum_{i=1}^{N} \alpha_i \\
&= \frac{1}{2} (\sum_{i=1}^{N} \alpha_i y_i x_i)^T \sum_{i=1}^{N} \alpha_i y_i x_i - (\sum_{i=1}^{N} \alpha_i y_i x_i)^T \sum_{i=1}^{N} \alpha_i y_i x_i + \sum_{i=1}^{N} \alpha_i \\
&= -\frac{1}{2} (\sum_{i=1}^{N} \alpha_i y_i x_i)^T \sum_{i=1}^{N} \alpha_i y_i x_i + \sum_{i=1}^{N} \alpha_i \\
&= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \alpha_i
\end{aligned}$$

and yield the so called *dual problem*, which is given by

$$\max_{a \in \mathbb{R}^N} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j, \tag{2.19}$$

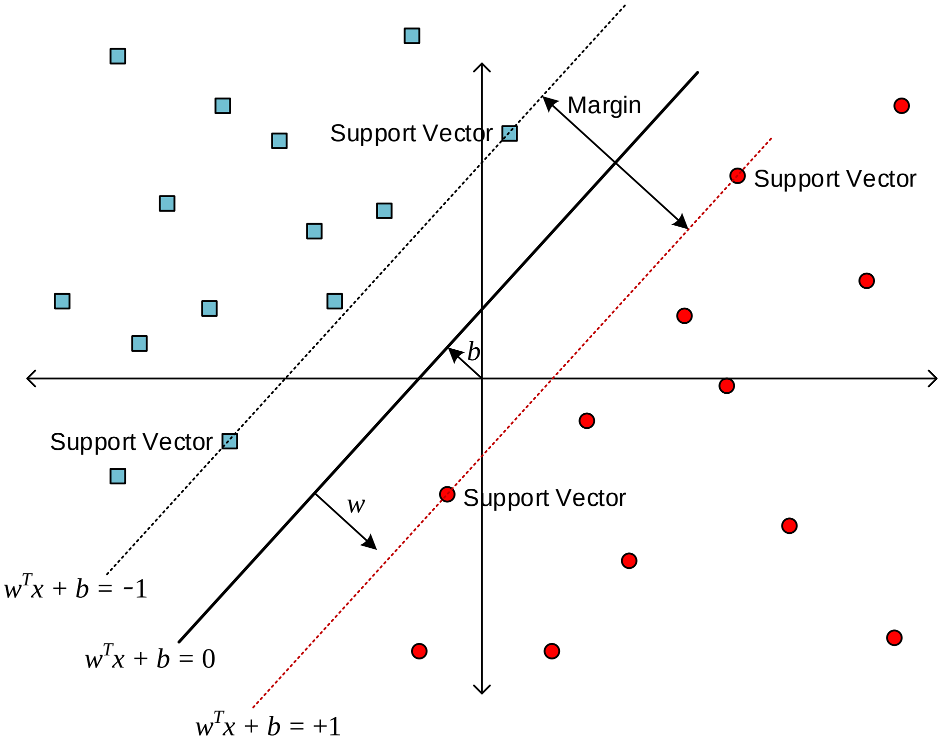


Figure 2.8: Canonical representation of the hyperplane.

subject to the following constraints:

$$\alpha_i \geq 0 \quad \forall\, i = 1, \ldots, N \quad \mathit{and} \quad \sum_{i=1}^{N} \alpha_i y_i = 0.$$

According to the *Karush-Kuhn-Tucker* conditions, the following properties hold for the solution of the dual representation:

$$\alpha_i(y_i(w^T x_i + b) - 1) = 0 \quad \forall\, i = 1, \ldots, N \tag{2.20}$$

which means that either $\alpha_i = 0$ or $y_i(w^T x_i + b) = 1$ is valid. The corresponding training examples with $\alpha_i > 0$ are the closest points to the hyperplane. These training instances are also called *support vectors*. Consequently, only the support vectors determine the separating hyperplane. This is important in the sense of structural risk minimization as generalization capabilities of a classifier depend not only on the training error but also on the complexity of the learned training model. The discriminative function can be reformulated in terms of the support vectors as

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{N} \alpha_i y_i x_i^T x + b). \tag{2.21}$$

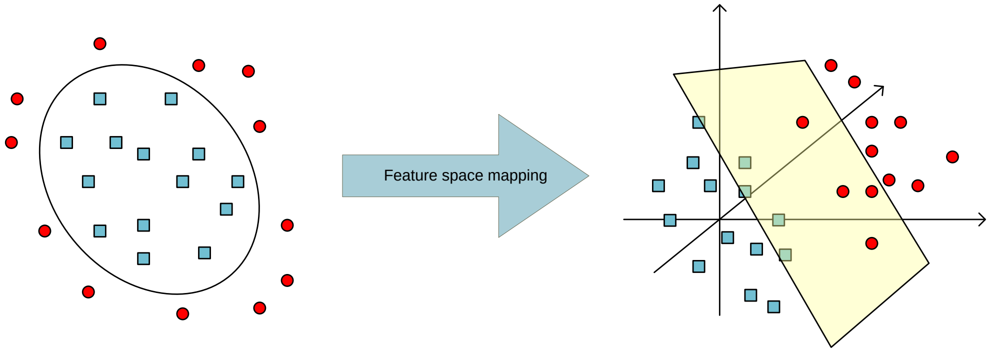


Figure 2.9: Feature space mapping into higher dimensional feature spaces provides linear separability of data points which are not linearly separable in the original feature space.

It should be noted, that for the dual optimization problem as well as for the discriminative function all operations in the feature space solely rely on dot products with the support vectors. This leads directly to another fundamental concept of SVMs, the *kernel trick*.

### 2.5.1 Kernel Trick

In the nonlinear separable case, the idea is to map the data points from the original feature space into a higher dimensional space where the data is linearly separable (Figure 2.9). In consideration of a feature space mapping $\phi(x)$, the discriminative function changes to

$$f(x) = \operatorname{sgn}(\sum_{i=1}^{N} \alpha_i y_i \phi(x_i)^T \phi(x) + b) \tag{2.22}$$

and the dual formulation of the optimization problem yields

$$\max_{a \in \mathbb{R}^N} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j), \tag{2.23}$$

subject to the constraints

$$\alpha_i \geq 0 \quad \forall\, i = 1, \ldots, N \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i y_i = 0.$$

Since computations in the high dimensional feature space are very expensive, the trick is to use a kernel function $k(x, y) = \phi(x)^T \phi(y)$ that simulates the dot product in the high dimensional feature space. The kernel function avoids the explicit mapping of data points into the high dimensional feature space, which can be even unknown. Nonetheless, the hyperplane is calculated in the high dimensional feature space. Despite the high dimensions, overfitting is generally avoided because of the maximum margin approach that leads to a sparse set of support vectors. Hence, the final decision function can be expressed as an $\alpha$-weighted linear combination of kernel responses with a bias $b$:

$$f(x) = \operatorname{sgn}(\sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b). \tag{2.24}$$

### 2.5.2 Kernel Functions

The kernel choice is a critical decision for the performance of a SVM. The kernel function measures intuitively the similarity between two data instances. According to *Mercer's theorem*, a positive definite kernel function has to be chosen to learn a SVM. A function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, where $\mathcal{X}$ is a nonempty set, is called a positive definite kernel function if the following conditions are met:

$$k(x_i, x_j) = k(x_j, x_i) \quad \forall x_i, x_j \in \mathcal{X}, \tag{2.25}$$

$$\sum_{i,j \in \{1,\dots,n\}} c_i c_j k(x_i, x_j) \geq 0 \quad \forall n \in \mathbb{Z}^+, c_1, \dots, c_n \in \mathbb{R}, x_1, \dots, x_n \in \mathcal{X}. \tag{2.26}$$

In general, it is difficult to check these conditions. Fortunately, there are general rules for constructing new kernels. If, for example, $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a distance function, then $k(x_i, x_j) = e^{-d(x_i, x_j)}$ is a kernel. Furthermore, positive definite kernels have the property that they are closed under sum, product, direct sum and multiplication by a scalar.

The most simple and fastest kernel function is the *linear kernel*

$$k_{linear}(x, y) = x^T y. \tag{2.27}$$

Further frequently used kernel functions are the *polynomial kernel*

$$k_{poly}(x, y) = (\gamma \cdot x^T y + c)^d, \tag{2.28}$$

the *radial basis function* (RBF) kernel

$$k_{rbf}(x, y) = e^{-\gamma \sum_i (x_i - y_i)^2}, \tag{2.29}$$

and the $\chi^2$-*kernel*

$$k_{\chi^2}(x, y) = e^{-\gamma \chi^2(x, y)} \tag{2.30}$$

which is based on the corresponding histogram distance:

$$\chi^2(x, y) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}. \tag{2.31}$$

Jiang et al. [2010a] used the $\chi^2$ kernel successfully for HoVW feature representations in the context of visual concept detection. They have shown that the $\chi^2$ kernel outperforms the linear and the RBF-kernel function. While the Gaussian kernel concentrates on the largest histogram differences due to the quadratic exponential decay, the $\chi^2$-kernel considers the bins more equally.

### 2.5.3 Soft Margin

Even in high dimensional feature spaces, linear separability cannot be taken for granted. Moreover, outliers, noise or mislabeled training samples can hinder the computation of a hyperplane. Therefore, slack variables are introduced to soften the constraints and to penalize data points on the wrong side of the hyperplane. This allows the solution for nonlinear separable data and keeps the separating hyperplane as simple as possible. These extensions lead to the following optimization problem

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \tag{2.32}$$

subject to

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \quad \forall\, i = 1, \ldots, N \tag{2.33}$$

where $\xi_i$ are the slack variables and $C$ is the penalty factor that controls the influence of misclassified training examples. Under these conditions the Lagrangian function changes to

$$L(w, b, \alpha, \beta) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (y_i(w^T \phi(x_i) + b) - (1 - \xi_i)) - \sum_{i=1}^{N} \beta_i \xi_i$$

$$\text{s.t.} \quad \alpha_i \geq 0 \quad \text{and} \quad \beta_i \geq 0 \quad \forall\, i = 1, \ldots, N.$$

The following equations can be derived from the Lagrangian function by setting its derivatives to zero:

$$w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i), \quad 0 = \sum_{i=1}^{N} \alpha_i y_i \quad \text{and} \quad \alpha_i = C - \beta_i. \tag{2.34}$$

Using the previous equations, the dual representation can be expressed as:

$$\max_{a \in \mathbb{R}^N} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) \tag{2.35}$$

subject to the constraints

$$0 \leq \alpha_i \leq C \quad \forall\, i = 1, \ldots, N \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i y_i = 0.$$

It can be observed that the slack variables completely disappear. Compared to the dual representation of equation 2.23, the only difference is the *penalty factor* $C$, that appears as an additional constraint of the Lagrange multipliers $\alpha_i$.

Due to the high memory requirements and the computational complexity, it is inefficient to directly solve the quadratic programming problem in form of the dual representation. The general idea of most optimization approaches is to break down the big problem into several smaller chunks, which can be solved more efficiently. One of the most popular training approaches is *sequential minimal optimization* SMO [Platt 1999]. This fast training method scales somewhere between linear and quadratic in the training set size. Apart from implementation techniques, it should be noted that SVMs basically do not provide probabilities. In order to obtain probabilistic results, Platt [1999] fits a sigmoid function to the training examples, that can be used in the classification stage to transform the confidence scores to posterior probabilities.

### 2.5.4 Multiple Kernel Learning

MKL is a SVM-based approach that considers a combination of multiple kernels. As already mentioned in Section 2.5.2, kernel functions are closed under scalar multiplication and sum. Thus, new kernel functions can be constructed as a linear combination of existing kernels $k = \sum_{j=1}^{K} \beta_j k_j$. Using a combined kernel, the SVM-based decision function from Section 2.5.1 changes to

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{N} \alpha_i y_i \sum_{j=1}^{K} \beta_j k_j(x_i, x) + b) \tag{2.36}$$

where $x_i$ are the support vectors with the corresponding class labels $y_i$. Additionally to the weights $\alpha_i$ and the bias $b$, the kernel weights $\beta_j$ have to be estimated. The idea of MKL is to find an optimal kernel combination. In general, this problem can be solved by cross-validation, but this approach additionally requires a cross-validation set and is limited to a small set of parameter combinations because of its computational complexity.

A more effective approach is yielded by plugging the combined kernel into the optimization problem of a single kernel SVM, whereby each kernel combination induces a different feature space. In other words, the aim of MKL is to find the feature space in which the largest margin between the classes can be obtained. Besides the combined feature space, each of the included kernel functions itself induces a feature space $\mathbb{R}^{d_k}$ and a corresponding mapping $\phi_k(x)$, where $d_k$ is the dimensionality of the $k$-th space. This leads to the following MKL primal optimization problem:

$$\min_{w,b,\xi,\beta} \quad \frac{1}{2}\left(\sum_{k=1}^{K} w_k^T w_k\right)^2 + C\sum_{i=1}^{N} \xi_i \tag{2.37}$$

subject to

$$y_i \left( \sum_{i=1}^{K} w_k^T \phi_k(x_i) + b \right) \geq 1 - \xi_i \quad \forall\, i = 1, \ldots, N \tag{2.38}$$

$$\xi_i \geq 0 \tag{2.39}$$

where $w_k$ can be written as $\beta_k w_k'$ with $\beta_k \geq 0\ \forall i = 1, \ldots, K$. In the literature, different MKL formulations are suggested, especially with regard to the regularization strategy [Bach et al. 2004; Kloft et al. 2009, 2010, 2011; Sonnenburg et al. 2006]. The most common regularization choice is the $L_1$-norm ($\sum_{i=1}^{K} \beta_k = 1$), which promotes sparse solutions. $L_1$-norm MKL allows to extract relevant knowledge about the underlying problem because the optimized kernel weights provide information about the relevance of the used feature representations. But sparse kernel combinations are not always preferable in practice [Kloft et al. 2009, 2011]. A more uniform distribution of kernel weights can be obtained by using the $L_2$-norm regularization $\|\beta_i\|_2^2 = \sum_{i=1}^{K} \beta_i^2 = 1$ [Kloft et al. 2009].

For solving the primal optimization problem of the MKL setting, it has to be transformed into its dual representation:

$$\min \gamma - \sum_{i=1}^{N} \alpha_i \tag{2.40}$$

subject to the following constraints

$$\begin{aligned} 0 \leq \alpha_i &\leq C \quad \forall\, i = 1, \ldots, N, \\ \sum_{i=1}^{N} \alpha_i y_i &= 0 \\ \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k_k(x_i, x_j) &\leq \gamma, \quad \forall\, i = 1, \ldots, K. \end{aligned}$$

Setting $K = 1$ this formulation corresponds to the SVM dual representation as described in Equation 2.23. Sonnenburg et al. [2006] solved the MKL problem by wrapping around a single kernel SVM and alternatingly optimizing the parameter values for $\alpha$ and $\beta$.

MKL can be applied to combine different feature representations, where each kernel takes a different feature representation into account. However, the fusion of heterogeneous data representations via MKL does not ensure a better classification performance, especially in the case of a few well-designed kernels [Bach and Obozinski 2010].

## 2.6 Transductive Learning

In comparison to inductive learning, transductive learning is not aimed at obtaining a general classification model for all possible test data items, but at obtaining an optimal classification model for the given test data only. The transductive setting is closely related to the problem of semi-supervised learning where the training set consists of labeled as well as unlabeled data. Semi-supervised and transductive learning have been researched for several years. The methods often rely on the semi-supervised smoothness assumption [Zhou et al. 2004], cluster assumption, or manifold assumption [Chapelle et al. 2006]. Typically, semi-supervised approaches can be divided into generative models (e.g., [Nigam et al. 2000]), low-density separation methods, and graph-based methods. An example for a low-density separation approach is Joachims' transductive SVM [Joachims 1999], while the approach of Bengio et al. [2006] using label propagation is an example of a graph-based approach. Zhou and Burges [2007] present a graph-based approach for transductive inference that exploits multiple views on the feature set, where each view (i.e., feature set) is represented by a separate graph.

In the field of video retrieval, the appearance of concepts is often related to a particular video sequence, episode, program, or broadcast. This observation motivates the idea of considering the content analysis task for a single video or episode as a transductive setting: the final classification model must be optimal for the given video only, and not in general, as expected for inductive learning. For this purpose, the unlabeled video test data has to be used in the learning process.

There are only a few approaches that propose model adaptation for video indexing and retrieval purposes, for example, by using semi-supervised or transductive learning. Wu et al. [2005] stated that the appearance of concepts changes over time and addressed the problem of concept drifting in videos. The authors used GMMs to model a concept and propose an incremental online learning framework to cope with concept drifting. For this purpose, videos are processed in a batch mode. The first batch of prelabeled data is used to learn a global GMM for each concept. The next batch of data is used to learn a set of locally optimized GMMs for each concept from the first unlabeled portion of the new data, aiming at an optimal classification performance on the current test batch. Recently, transductive learning has been considered in the field of video concept detection. Qi et al. [2007b] presented a transductive concept detection system for home videos relying on two clustering steps. The training and test samples are clustered together and an expectation maximization (EM)-based tuning of clusters aims at purifying the clusters. Tian et al. [2008] proposed a transductive approach for video annotation using local kernel classifiers. To check its applicability, a measure has been suggested to test whether a sample can be classified using the neighbored samples. Wang et al. [2008a] defined several subdomains for the domain of news videos to model and analyze videos using a transductive framework. Transductive inference is based on clustering and Vapnik's combined bound. Finally, Wang et al. [2008b] presented an approach for transductive multilabel learning for video concept detection. The authors proposed a discrete hidden Markov random field that also assures that multilabel interdependence over unlabeled data is coherent with interdependence of labeled data.

In contrast to the previously mentioned transductive approaches, our framework [Ewerth 2008; Ewerth et al. 2011] is designed in a general manner and not only for a single video analysis task. It has been applied, for example, for shot boundary detection, semantic video retrieval or person indexing [Ewerth et al. 2007a]. The goal of the proposed framework is to obtain a robust indexing or retrieval result for a given video. Therefore, an initial classification or clustering result of a baseline system is exploited to improve its quality for a particular video. The unlabeled data of the previously unseen video are incorporated into the learning and classification process. To adapt a concept model to its appearance in a particular test video, the baseline model is used either to classify the frames or to obtain probability scores for them. This result serves as the source for automatically generating the training set for the subsequent learning and model adaptation process. This training set consists only of samples of the test video under consideration. If the initial result is a binary classification result, all samples are used for the subsequent training process. Otherwise, if probability or confidence scores are available, they can be used to select the top (or bottom) p percent of the best (or worst) samples as positive and negative training data. The automatically generated training data are used to select the best features for the classification task of this video via *Adaboost* [Freund and Schapire 1997]. Then, the set of selected features is split into two disjoint sets in order to enable the training of different classifier views. The feature sets are used to train new classifiers directly on the video. Finally, the newly trained classifiers and the initial classifier form an ensemble that uses majority voting to reclassify the video samples. The initial classifier is incorporated to prevent performance degradation for inappropriate cases.

## 2.7 Benchmarks

Progress in the field of visual concept detection clearly profited from publicly available benchmarks that allow the comparison of different approaches and promote the competition between researchers. The most important role in the field of video concept detection plays the *TRECVid Challenge*, which is described in Section 2.7.1. Two further benchmarks, which have been used within this thesis, are presented in Section 2.7.2 and Section 2.7.3: the *Mediamill Challenge* [Snoek et al. 2006b] and the *Pascal VOC* Challenge [Everingham et al. 2010].

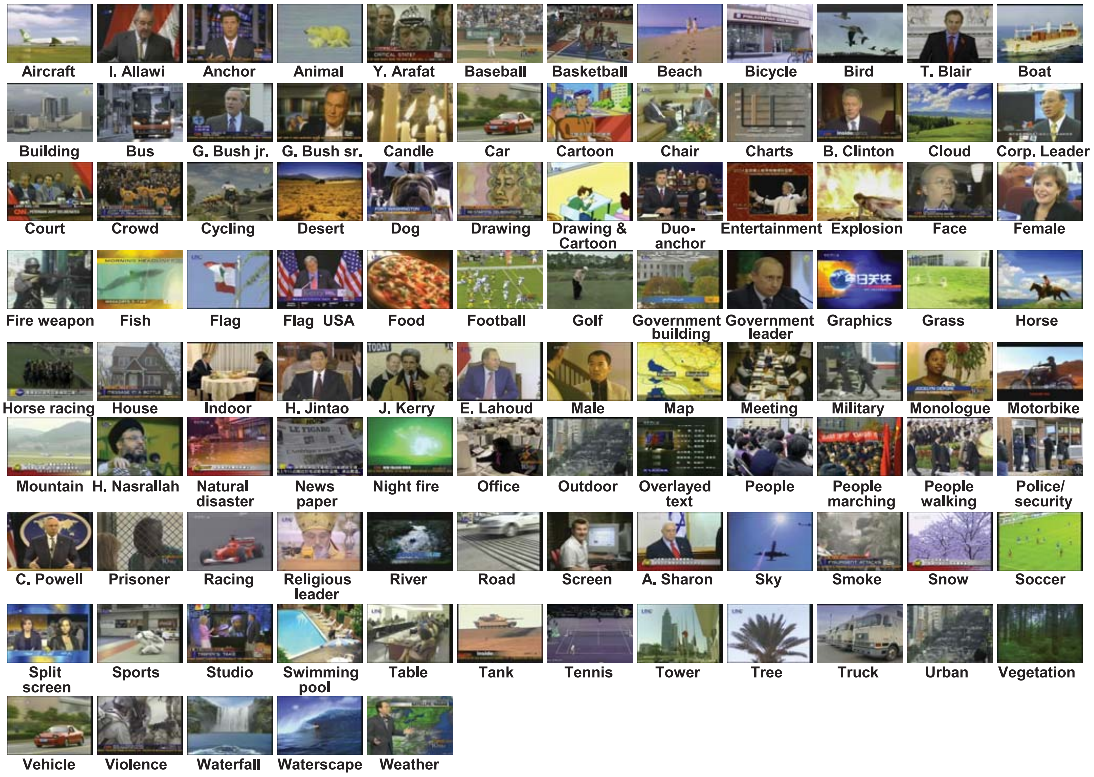


Figure 2.10: Visual impressions of the 101 semantic concepts from the *Mediamill Challenge*, source [Snoek et al. 2006b].

## 2.7.1 TRECVid Challenge

The aim of the *TREC Video Retrieval Evaluation* (TRECVid) conference series [Smeaton et al. 2006] sponsored by the *National Institute of Standards and Technology* (NIST) is to encourage research in digital video retrieval. *TRECVid* was born within the *Text REtrieval Conference* (TREC) series and became an independent workshop in 2003. The annually held international workshop provides a forum for researchers to compare their approaches on a common test set and to discuss their video retrieval results. The video analysis tasks range from shot segmentation over known item search, content-based copy detection and video surveillance to concept detection, which is also called high-level feature extraction or semantic indexing. A 5-year retrospective of achievements with respect to the semantic indexing task was given by Smeaton et al. [2009]. *TRECVid* provides large video data sets, concept vocabularies and standard evaluation procedures. The labeling of the training data has been accomplished by collaborative annotation efforts of the international participating teams. The domain of the video data set approximately changes every three years with an annually increasing amount of video data, so that the current test collection is new and unknown

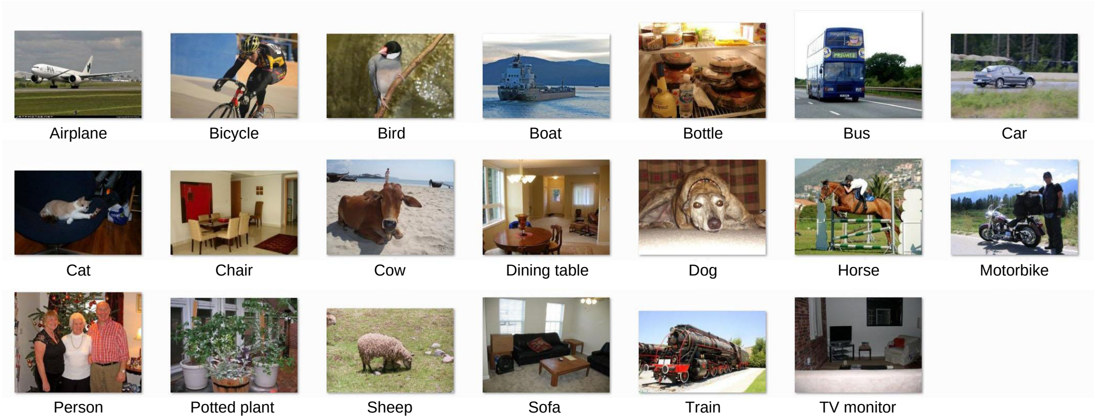


Figure 2.11: Visual impressions of the 20 object classes from the *VOC Challenge*.

to the participants. The semantic indexing task started with approximately 169 hours of news videos, continued with documentary films and recently employed internet archive videos. Not only the size of the data set grew, but also the number of concepts was annually increased to encourage researchers to build generic concept detectors. Started with 39 LSCOM-lite concepts in 2005, a large set of 346 concepts had to be detected in 2011 and the data set comprised 600 hours of internet archive videos with durations between 10 seconds and 3.5 minutes.

### 2.7.2 Mediamill Challenge

The *Mediamill Challenge* [Snoek et al. 2006b] offers a dataset which is based on the *TRECVid* 2005 [Smeaton et al. 2006] training set with an extensive set of 101 fully annotated concepts. Visual impressions of the 101 semantic concepts, including objects, scenes, events and personalities, are given in Figure 2.10. The data set consists of 86 hours of news videos containing 43,907 completely annotated video shots. These shots are divided into a training set of 30,993 shots and a test set of 12,914 shots. Per video shot a single representative keyframe as well as pre-computed low-level features are given. Furthermore, the challenge provides baseline performances for different experimental settings. It is distinguished between uni-modal and multi-modal experiments as well as between early and late fusion settings. The uni-modal experiments are based on either visual or textual features.

### 2.7.3 Pascal VOC Challenge

The annual *VOC Challenge* [Everingham et al. 2010] provides a dataset for object detection, image classification and object segmentation. For the *VOC Challenge* 2007, this dataset together with the ground truth information is completely online available. In total, the dataset consists of 9,963 fully annotated Flickr images. The images are approximately equally splitted into training and test set, 5,011 and 4,952 images, respectively. The image classification task comprises 20 object classes, like “bird”, “car”, “cow”, “aeroplane”, “bicycle”, “boat”, “bottle”, “chair”, “dining table” or “person”. Visual impressions of the 20 object classes are provided in Figure 2.11.

## 2.8 Performance Measures

Different performance metrics exist to evaluate retrieval results. Two of the most famous measures in the context of information retrieval are recall and precision. While recall describes the hit rate

$$recall = \frac{\#correctDetectedDocuments}{\#relevantDocuments}, \tag{2.41}$$

precision represents the probability that a detected document is relevant

$$precision = \frac{\#correctDetectedDocuments}{\#correctDetectedDocuments + \#falseAlarms}. \tag{2.42}$$

Recall and precision are negatively correlated to each other. This means that the recall value can be maximized for example by classifying all documents as positives, which in turn leads to a high false positive rate and thus to a low precision value. Since both measures interact they are mostly considered together. An optimized system should find as many relevant documents as possible with minimal false alarms.

Joint measures are the *equal error rate* (EER), where recall and precision are assumed to be equal, and the *f1-score*, which combines recall and precision in a single value by calculating the harmonic mean

$$f1\text{–}score = \frac{2 \cdot recall \cdot precision}{recall + precision}. \tag{2.43}$$

The most commonly used quality measure to evaluate concept retrieval results is the *average precision* (AP) score. This value is calculated from a list of ranked images or video shots as follows:

$$AP(\rho) = \frac{1}{|R|} \sum_{k=1}^{N} \frac{\left|R \cap \rho^k\right|}{k} \psi(i_k) \tag{2.44}$$

$$\text{with} \quad \psi(i_k) = \begin{cases} 1 & \text{if } i_k \in R \\ 0 & \text{otherwise} \end{cases}$$

where $N$ is the length of the ranked document list, $\rho^k = \{i_1, i_2, \ldots, i_k\}$ is the ranked document list up to rank $k$, $R$ is the set of relevant documents, $\left|R \cap \rho^k\right|$ is the number of relevant documents in the top-$k$ of $\rho$ and $\psi(i_k)$ is the relevance function. Generally speaking, AP is the average of the precisions at each relevant document. To evaluate the overall performance for multiple concept classes, the *mean AP* score is calculated by taking the mean value of the AP scores from the individual concepts.

Large concept sets and the always increasing amount of data at the *TRECVid Challenge* made the judgement of the complete test set prohibitively expensive. Therefore, a depth pooling strategy has been employed to evaluate the performance of the retrieval systems. To reduce the judgement effort, a depth $k$ pool is build per concept, that consists of the top-$k$ documents from each of the submitted retrieval results. The judged pool documents are then used to calculate the AP scores of the retrieval results, whereby unpooled documents are ignored. Cutoffs at about 100 documents have shown to be effective in evaluating retrieval systems [Smeaton et al. 2006] and led to pool sizes considerably smaller than the entire test set. Nevertheless, the annotation effort still remains enormous. If less shots have to be inspected, the same judgement effort could be used to evaluate more concept classes. Yilmaz and Aslam [2006] introduced an estimation of AP, called *inferred average precision* (infAP). The infAP measure is a statistical method for evaluating large-scale retrieval results using incomplete judgements. The judgement set is the result of a uniform random sampling of the documents in the pool. Their experimental results show that a sampling rate of about 50% yields very good estimates of the standard AP measure [Aslam et al. 2006; Yilmaz and Aslam 2006]. To obtain more accurate estimates with less judgement effort, the pooling strategy was adapted to take multiple pools into account with different sampling rates [Yilmaz et al. 2008]. Three pools were formed from the retrieval results. The first pool consists of the union of the top 10 documents from each retrieval result, the second pool summarizes the ranks 11 to 100 and the last pool contains the remaining documents. Corresponding to the relative importance of the pooled documents for the AP measure the first pool was sampled at the rate of 100%, the second at 20% and the last one at 5%. In the context of the *TRECVid* workshop this strategy is also called extended infAP. It allows the judgement of considerably less than 50% of the pooled documents while maintaining the estimation quality of infAP [Yilmaz et al. 2008]. Nevertheless, the fairness of this pooling strategy for evaluating retrieval results that did not contribute to the pool is still in question.

Since the AP score depends on the frequency of a concept class in the test set, it is difficult to compare AP scores from different test sets and of different concept classes. Yang and Hauptmann [2008] proposed *deltaAP*, which evaluates a concept classifier based on how much it performs better than a random classifier. It can be easily shown that the AP score of a randomly generated retrieval result with at least one positive document, is not zero. The randomAP score corresponds to the concept frequency in the given test set. Thus, deltaAP is given as the difference between the AP and the randomAP value:

$$deltaAP = AP - randomAP. \tag{2.45}$$

Furthermore, the official partial randomization test of the *TRECVid* evaluation [Smeaton et al. 2006] is used to determine whether a system is significantly better than a reference system or whether the difference is only due to chance.

## 2.9 Summary

In this chapter, the basic terms necessary to understand the remainder of the thesis have been explained and a survey of fundamental methods and research work related to generic concept detection in images and videos has been provided.

After describing the most popular concept lexicon LSCOM, the general pattern recognition process for concept detection has been explained, which is generally considered as a supervised learning task.

The focus of this chapter has been on state-of-the-art BoVW concept detection approaches. The main components of these approaches are the extraction of local keypoint features with SIFT as the most prominent descriptor, the coding step where the local descriptors are typically processed by using a kind of vector quantization and the pooling step where the codes are transformed into the final feature representation. A general conclusion of which BoVW strategy is superior and therefore preferable for visual concept detection is difficult because of the dependency on the visual concepts under consideration, the used training and testing set, the machine learning settings and further implementation details. However, the following tendencies can be observed. The K-means algorithm is the predominant method for the construction of visual vocabularies. The gains of special codeword selection methods are small compared to the use of larger codebooks. The most important factor in the design of visual vocabularies is the size of the codebook itself. The larger the codebook, the better the concept detection performance. 4,000 visual words have shown to be the best tradeoff between codebook size and computational costs. For more than 4,000 visual words, the increase in concept detection performance seems to slow down. While early BoVW approaches mainly focused on scale- and rotation-invariant interest point descriptors, today it seems that this representation is often insufficient to describe natural images. For scene classification, dense sampling strategies have outperformed the previously mentioned interest point detectors. In any case, an increased number of image keypoints has a positive effect on the concept retrieval performance. Concerning coding and pooling, soft assignment and sparse coding are superior to hard vector quantization. Furthermore, considering only the top k nearest neighbors at the coding step has a positive effect on the retrieval performance. For linear SVMs, the max-pooling strategy turned out to be preferable. The use of spatial as well as color information yields performance improvements for most visual concepts. While spatial pyramids tend to obtain better overall concept detection performance than the spatial coordinate coding approach, spatial coordinate coding is much more computational and memory efficient.

Furthermore, the predominant SVM algorithm used for building concept models has been described in this chapter. The ideas and fundamentals of SVMs are explained, important kernel functions are listed and MKL is introduced. MKL is a SVM-based method that considers a weighted combination of multiple kernel functions. This method can be used to combine different feature representations, where each kernel takes a different feature representation into account.

After explaining the differences between inductive and transductive learning, publicly available image and video benchmarks as well as performance metrics used within this thesis have been presented in Section 2.7 and Section 2.8, respectively.

The *TRECVid Challenge* [Smeaton et al. 2006] has been dominated by BoVW approaches over the last years. The best semantic indexing system at *TRECVid* 2010 used sparse and dense sampling, multiple color SIFT descriptors, spatial pyramids, multi-frame video processing per shot, and kernel-based machine learning [Snoek et al. 2011]. At the *TRECVid Challenge*s 2011 and 2012, a supervector coding approach yielded the best results [Inoue et al. 2012, 2013]. But the answer what may ultimately be the decisive factor in the design of BoVW approaches, is difficult. Man power and high computational capacities, for example, via GPU support, are also important factors that influence the retrieval results at the *TRECVid Challenge*. High computational capacities allow the integration of various extensive feature representations. Furthermore, the analysis of an increased number of frames per shot becomes possible, yielding a better chance to find the occurring concepts. Additionally, the collaborative annotations of the training set are partly erroneous. Given enough man power, the annotations of the training data can be revised and refined to improve the quality of the learned concept models. These differences in the training set impede the comparability of the used approaches.

Altogether, a continuous increase in the complexity of concept detection systems is observable. Approaches are evolving towards systems using, for example, many different combinations of keypoint sampling strategies, local keypoint descriptors, coding and pooling methods. “A lot helps a lot” seems to be the guiding principle in selecting the proper feature set, which is not surprising considering the wide range of diversified concepts.

*"Those who wish to succeed, must ask the right preliminary questions."*

Aristotle

# 3 Enhanced Local Descriptors for Mid-Level Feature Coding

## 3.1 Introduction

Visual concept detection plays a key role in multimedia retrieval, navigation and browsing. Current approaches mainly focus on local keypoint features, with SIFT [Lowe 2004] as the most successful descriptor. These local descriptors are usually clustered to build a visual vocabulary, where the cluster centers are regarded as "visual words". Similar to the representation of documents in the field of text retrieval, images are represented as histograms indicating the presence of each visual word. While features, like global color or Gabor histograms, are referred to as low-level features, BoVW approaches attempt to encode higher level information and are therefore also called mid-level features. Current concept detection systems mainly rely on this BoVW representation.

In this chapter, different aspects and extensions to the BoVW concept detection approach are investigated. Section 3.2 investigates the effect of different spatial extents in a state-of-the-art system for visual concept detection. Based on the observation that SIFT descriptors with different spatial extents yield large performance differences, a novel concept detection system is proposed that combines feature representations for different spatial extents using MKL. In Section 3.3 the BoVW approach is extended to a multi-modal video concept detection system. To enhance concept detection, the Bag-of-Words approach is leveraged for audio features and MKL is again proposed as the appropriate fusion scheme for these BoAW and BoVW features. Section 3.4 presents a method for the SIFT-

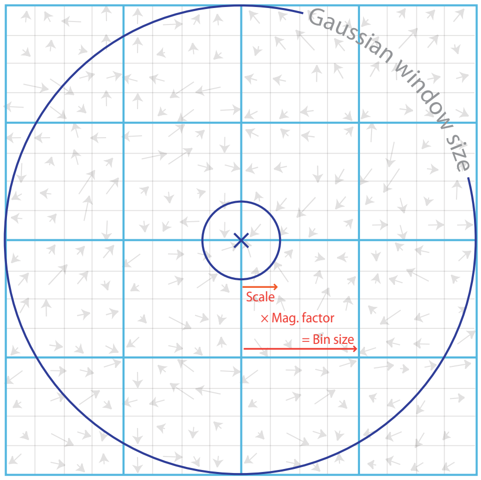


Figure 3.1: SIFT descriptor geometry.

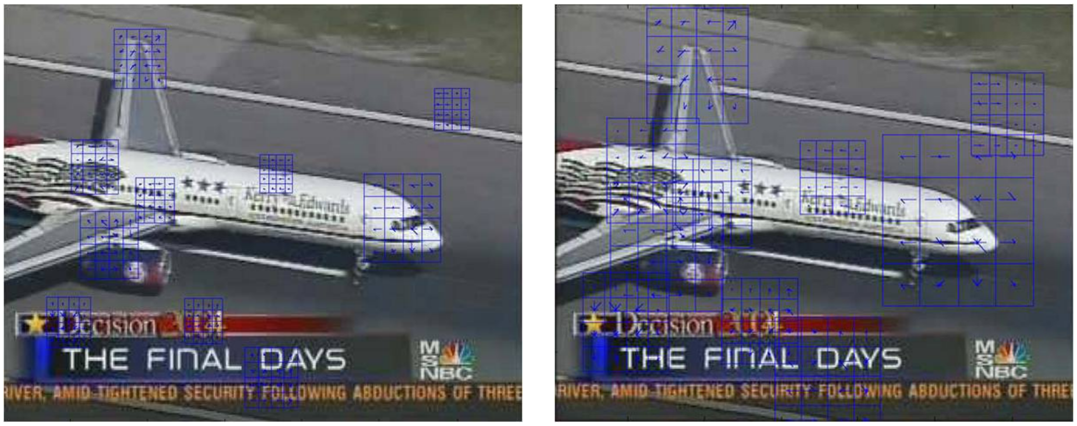
based integration of color information, called color moment SIFT (CMSIFT). CMSIFT achieves the same concept detection performance as the state-of-the-art transformed color SIFT descriptor, but is considerably faster. Finally, Section 3.5 summarizes the chapter.

Parts of this chapter have been published by Mühling et al. [2011b, 2012].

## 3.2 On the Spatial Extents of SIFT Descriptors

Although several variations of keypoint sampling strategies, vocabulary construction techniques, local descriptor projection methods, and machine learning algorithms have been evaluated, the impact of the spatial extents of local SIFT descriptors has not been studied for BoVW approaches in the field of visual concept detection. Previous work related to BoVW approaches has been extensively discussed in Chapter 2. In the current section, the impact of the spatial extents of local SIFT descriptors using a state-of-the-art visual concept detection system is investigated. It has been observed that for particular semantic concepts different spatial extents cause large performance differences. Based on this observation, a system is proposed that combines feature representations for different spatial extents using MKL.

This section is organized as follows: In Section 3.2.1 the terms "spatial bin size" and "magnification factor" are explained. Section 3.2.2 presents a state-of-the-art concept detection system as well as the fusion scheme for multiple spatial extents. A note on SVM parameter optimization is given in Section 3.2.3. Finally, experimental results are presented in Section 3.2.4, followed by a discussion in Section 3.2.5.


Decision
THE FINAL DAYS
MSNBC
RIVER, AMID TIGHTENED SECURITY FOLLOWING ABDUCTIONS OF THREE
Decision
THE FINAL DAYS
MSNBC
RIVER, AMID TIGHTENED SECURITY FOLLOWING ABDUCTIONS OF THREE

(a) Magnification factor 3.0

(b) Magnification factor 6.0

Figure 3.2: Visualization of SIFT descriptors from scale-invariant keypoints with different magnification factors.

### 3.2.1 Spatial Bin Size and Magnification Factor

As already described in Section 2.4.1, a SIFT descriptor encodes the appearance of a keypoint by using a spatial histogram of image gradients, where a Gaussian weighting function is applied to reduce the influence of gradients further away from the keypoint. The SIFT descriptor geometry is specified by the number and size of the spatial bins and the number of orientation bins. Using eight orientation bins and 4x4 spatial bins, the local descriptor results in a 128-dimensional vector. Based on this geometry the spatial bin size is a measure for the spatial extent of a SIFT descriptor. The spatial bin size of a keypoint descriptor is the product of the keypoint scale and the magnification factor (Figure 3.1). Thus, in the case of scale-invariant keypoints the magnification factor determines the spatial bin size and patch size. The default magnification factor of the SIFT implementation is 3 [Lowe 2004]. The effect of different magnification factors on the spatial extents of SIFT descriptors extracted at scale-invariant keypoints is visualized in Figure 3.2. Since in the case of dense sampling a keypoint does not provide scale information, the spatial bin size for dense sampled keypoints is specified directly.

### 3.2.2 Concept Detection System

In this section, the proposed BoVW system for visual concept detection is presented. The visual vocabulary is generated using the K-means algorithm, and the images are described as a histogram indicating the presence of each visual word. Further implementation choices of the used BoVW approach, such as the soft-weighting scheme, the integration of color and spatial information, the used classification scheme, and the proposed MKL framework to combine feature representations based on different spatial extents are described in the following subsections.

### Local Feature Extraction

Two sampling strategies are applied to extract local SIFT descriptors: the DoG salient point detector and the dense sampling strategy. To extract SIFT features, the implementation of the `Vision Lab Features Library` (`VLFeat`) [Vedaldi and Fulkerson 2010] is used. It also provides a fast algorithm for the calculation of densely sampled SIFT descriptors of the same scale and orientation. For dense sampled keypoints a step size of 5 pixels is applied.

Color information is integrated using RGB-SIFT. Therefore, the SIFT descriptors are computed independently for the three channels of the RGB color model. The final keypoint descriptor is the concatenation of the individual descriptors. Due to the normalizations during the SIFT feature extraction, RGB-SIFT is equal to the transformed color SIFT descriptor, and thus invariant against light intensity and color changes [Sande et al. 2008].

### Feature Coding and Pooling

A kernel weighted approach is used for coding and pooling. Instead of mapping a keypoint only to its nearest neighbor, like in the codeword plausability approach, or to all visual words, like in the kernel codebook approach, the codebook candidates are locally constrained to the $l$-nearest visual words. This locality constraint has shown to be superior for BoVW approaches [Liu et al. 2011]. Using a visual vocabulary of $N$ visual words, the importance of a visual word $v_t$ in the image is represented by the weights of the resulting histogram bins $w = [w_1, \dots, w_t, \dots, w_N]$ with

$$w_t = \sum_{i=1}^{K} \begin{cases} sim(d_i, v_t) & \text{if } v_t \in N_i^l \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where $K$ is the number of local descriptors, $N_i^l$ are the $l$-nearest neighbors of the local descriptor $d_i$ in the visual vocabulary and $sim(d_i, v_t)$ is a similarity function based on the Euclidean distance $d$:

$$sim(d_i, v_t) = exp(-\gamma \cdot d(d_i, v_t)). \tag{3.2}$$

The $\gamma$-value is chosen as the maximum Euclidean distance between two codebook candidates. Finally, the resulting histograms are normalized using the $L_1$-norm.

To capture the spatial image layout, a spatial pyramid of 1x1 and 2x2 equally sized subregions is used. The HoVW features are generated independently for each subregion and are concatenated into the final feature vector. The weighting of the final feature vector is realized as specified by Lazebnik et al. [2006].

#### Classification

Concept detection is considered as a supervised classification task. SVMs that have proven to be powerful for visual concept detection [Smeaton et al. 2006] are used for the classification of the concepts. The kernel choice is a critical decision for the performance of a SVM. Since histogram representations are used, the $\chi^2$-kernel, as already described in Section 2.5, is applied:

$$k_{\chi^2}(x, y) = e^{-\gamma\chi^2(x,y)} \tag{3.3}$$

Jiang et al. [2010a] used the $\chi^2$-kernel successfully for BoVW features in the context of visual concept detection. In their study, the $\chi^2$-kernel has outperformed the traditional linear and RBF kernels. A note on the optimization of the parameters $\gamma$ and $C$ is given in Section 3.2.3.

#### Fusion of Multiple Spatial Extents

In order to combine the capabilities of feature representations based on different spatial extents, MKL is applied to find an optimal kernel combination

$$k = \sum_{i=1}^{n} \beta_i k_i \quad \text{with} \quad \beta_i \geq 0 \tag{3.4}$$

where each kernel $k_i$ takes a different feature representation into account. The sparsity of the kernel weights can be controlled by the $L_p$-norm. The $L_2$-norm is applied which leads to a more uniform distribution of kernel weights than the $L_1$-norm. Throughout the experiments, the MKL framework provided by the `Shogun library` [Sonnenburg et al. 2010] is used in combination with the SVM implementation of Joachims [1998], called SVM$^{\text{light}}$.

### 3.2.3 A Note on SVM Parameter Search

SVM tutorials generally advise extensive parameter search to find the optimal SVM parameters [Ben-Hur and Weston 2011; Hsu et al. 2010], especially for the kernel parameter $\gamma$ and the SVM penalty factor $C$. Grid search and cross-validation are often applied to optimize these parameters. Generally, a coarse-to-fine grid search is used [Hsu et al. 2010]. A typical coarse search starts with exponentially growing sequences for $\gamma$ ($2^{-5}, 2^{-4}, \ldots, 2^{5}$) and $C$ ($2^{-15}, 2^{-14}, \ldots, 2^{3}$). A SVM has to be trained and validated for each parameter combination, resulting in 209 parameter settings for the previously given values. Afterwards, a finer grid search is performed in the neighborhood of the best parameter setting from the previous coarse search.

For validating a parameter setting a part of the training set is left out in the learning stage. This set of training samples, also called validation set, is used

| **AP** [%] | **Heuristic** | **Grid search** |
|---|---|---|
| Aeroplane | 67.80 | 72.17 |
| Bicycle | 52.77 | 53.75 |
| Bird | 40.54 | 42.19 |
| Boat | 60.22 | 59.00 |
| Bottle | 18.82 | 18.10 |
| Bus | 52.84 | 53.17 |
| Car | 70.50 | 73.06 |
| Cat | 48.87 | 48.38 |
| Chair | 46.41 | 46.38 |
| Cow | 28.97 | 28.94 |
| Dining table | 38.45 | 36.47 |
| Dog | 39.94 | 38.87 |
| Horse | 74.96 | 75.31 |
| Motorbike | 53.73 | 55.65 |
| Person | 81.28 | 82.41 |
| Potted plant | 19.96 | 18.99 |
| Sheep | 31.76 | 31.17 |
| Sofa | 44.86 | 44.30 |
| Train | 71.00 | 70.60 |
| TV monitor | 47.98 | 47.48 |
| Mean | 49.58 | 49.82 |

Table 3.1: Parameter selection for the SVM penalty factor $C$ using heuristic values and grid search.

to predict the performance of the learned classifier. To improve the performance prediction this process can be repeated several times on different training and validation subsets. Then, the overall evaluation score is yielded by averaging the separate results. A commonly used technique is cross-validation. Therefore, the training data is randomly divided into k folds. Stratified cross-validation additionally ensures that the proportion of positive and negative examples keeps approximately the same in each fold. To evaluate a classifier on the training set, stratified k-fold cross-validation performs k rounds of training and testing. In each round one of the folds is used as validation set and the remaining k-1 folds are used as training set.

In general, many concepts offer only a few positive and many negative training examples. This leads to the risk that the optimization process using the default accuracy criteria tends towards a classifier that always predicts a negative label. Therefore, the average precision measure is used as evaluation criteria due to the unbalanced training sets.

However, cross-validation as well as grid search are computationally very expensive or even impractical for training sets with several ten thousand positive and negative training samples and high feature dimensions in the range of 4,000 to 20,000. Zhang et al. [2007] showed that setting the $\gamma$ value of the $\chi^2$-kernel to the average distance between all training instances yields comparable results and saves much computational effort.

The second parameter, that has to be optimized, is the penalty factor $C$. By default, the SVM$^{\text{light}}$ implementation [Joachims 1998] sets the penalty factor $C$ to:

$$\begin{aligned} C_{heuristic} &= \frac{1}{x_{avg} \cdot x_{avg}} \\ \text{with} \quad x_{avg} &= \frac{\sum_{i=1}^{N} \sqrt{k(x_i, x_i) - 2 \cdot k(x_i, x_0) + k(x_0, x_0)}}{N}, \end{aligned} \tag{3.5}$$

where $k$ is the kernel function, $N$ is the total number of training examples and $x_0$ is the null vector. Experimental results on the *VOC Challenge* show that using the heuristic value for the SVM parameter $C$ leads to results comparable to an exhaustive parameter search as presented in Table 3.1. The experiments are based on densely sampled SIFT descriptors with a visual vocabulary of 4,000 visual words. The $\gamma$ value of the $\chi^2$-kernel is set to the average distance between all training instances. Three-fold stratified cross-validation and a coarse-to-fine search is used to optimize the parameter $C$. Within the coarse search the values $2^{-15}, 2^{-14}, \dots, 2^{3}$ are validated, followed by a fine search for the values $2^{b-\frac{3}{4}}, 2^{b-\frac{1}{2}}, \dots, 2^{b+\frac{3}{4}}$, where $2^b$ is the best performing parameter value from the coarse search.

Both the heuristic approach and the cross-validation approach lead to similar results. While the parameter optimization using cross-validation and grid search leads to 49.8% mean AP, the heuristic approach achieves 49.6%. Thus, the tremendous amount of additional computational efforts for parameter search is not worse the negligible performance improvement.

### 3.2.4 Experimental Results

In this section, the performance impact of the spatial extents of SIFT descriptors and the combination of different spatial extents using MKL in the field of visual concept detection is investigated. For this purpose, two benchmarks are used, the *Mediamill Challenge* [Snoek et al. 2006b] with 101 semantic concepts and the *VOC* Challenge [Everingham et al. 2010] with 20 object classes. Within the experiments based on the *Mediamill Challenge*, the positive and negative training instances are each restricted to 5,000 samples per concept to speed up the training process.

The experiments are based on a pure visual analysis of the data. Since color SIFT descriptors achieve superior performance for concept detection [Sande et al. 2008], the RGB-SIFT descriptor is used exclusively. Several experiments have been conducted on the two benchmarks to investigate the impact of spatial bin sizes in combination with different sampling strategies (sparse and dense sampling), different vocabulary sizes (1,000 and 4,000 visual words) and spatial pyramids.

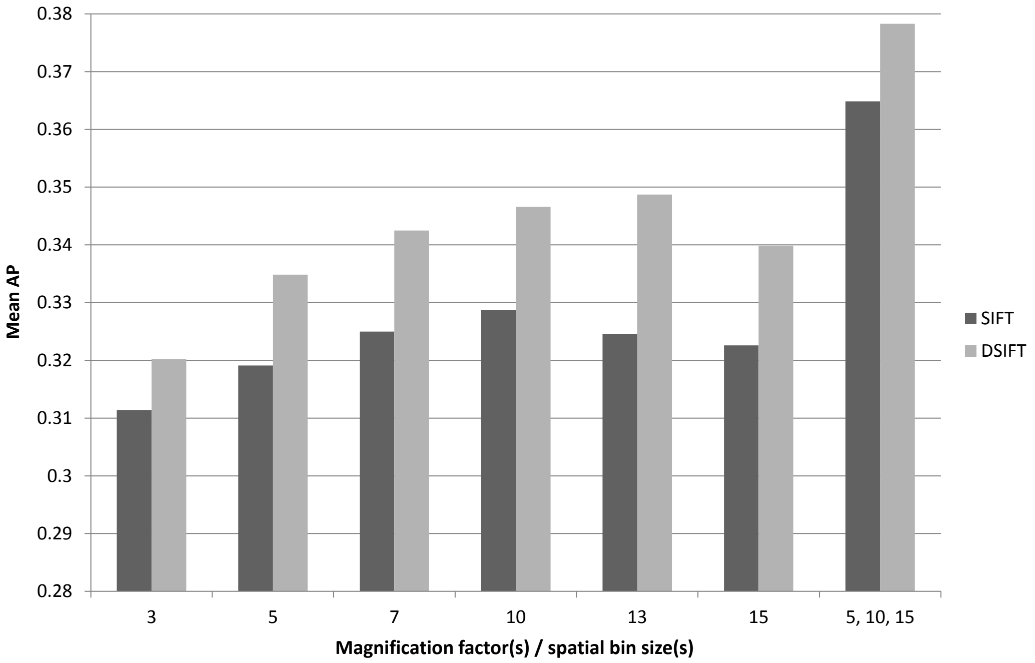
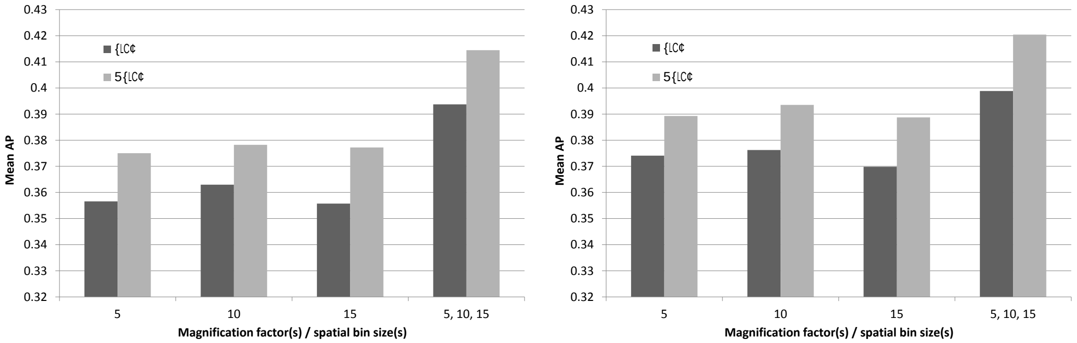


Figure 3.3: Evaluation of different spatial sizes on the *Mediamill Challenge* using a 1,000-dimensional vocabulary (averaged over 101 concepts).

The experimental results on the *Mediamill Challenge* are presented in Figures 3.3 and 3.4.

First, the impact of different spatial bin sizes in combination with a vocabulary size of 1,000 visual words and different sampling strategies has been investigated. Using a magnification factor of 10, an improvement of 5.5% was achieved compared to the default factor of 3 (Figure 3.3). In the case of dense sampling, the best performance was achieved using a spatial bin size of 13. To measure the influence of the non-deterministic K-means algorithm on the results several runs for the magnification factor and spatial bin size of 10 have been performed. Using 10 iterations, the mean AP and the standard deviation amounts to 32.91% ± 0.06 for scale-invariant keypoints and 34.6% ± 0.08 in the case of dense sampling.

Second, the experiment has been repeated with an increased vocabulary size of 4,000 visual words for the magnification factors and spatial bin sizes 5, 10 and 15 (Figure 3.4). Additionally, the experiment has been conducted in combination with a spatial pyramid representation. The spatial pyramids are constructed using a spatial grid of 1x1 and 2x2 regions. In both experiments based on salient points, the best performance was achieved using a magnification factor of 10, and the best spatial bin size for dense sampled SIFT descriptors was 10, too. Furthermore, it can be observed that some visual concepts yielded large performance differences for varying magnification factors and spatial bin sizes, respectively. Table 3.2 shows these differences for selected concepts.



(a) W/o spatial pyramids

(b) Spatial pyramids

Figure 3.4: Evaluation of different spatial sizes on the *Mediamill Challenge* using a vocabulary of 4,000 visual words (averaged over 101 concepts).

Third, the feature representations of different magnification factors and spatial bin sizes have been combined using MKL. Significant performance improvements were achieved in all experiments for both sparse and dense sampling. Using a vocabulary of 4,000 visual words, the relative performance improvement was up to 10.7% in the case of sparse sampling and with spatial pyramids up to 8.2% in the case of dense sampling as depicted in Figures 3.4(a) and 3.4(b), respectively. The combination of all spatial pyramid representations based on sparse and dense sampling using MKL achieved a mean AP of 43.2%.

Finally, different spatial bin sizes have been analyzed on the *VOC Challenge* in combination with sparse and dense sampling. In these experiments, spatial pyramids and a vocabulary size of 4,000 visual words were used (Figure 3.5). The best performance based on sparse sampling was achieved using a magnification factor of 10, like on the *Mediamill Challenge*. When dense sampling was used, the best performance was achieved for a spatial bin size of 5. In both cases, the combination of different spatial sizes yielded significant performance

| **AP** [%] | **Sparse sampling** | | | **Dense sampling** | | |
|---|---|---|---|---|---|---|
| | **5** | **10** | **15** | **5** | **10** | **15** |
| Bicycle | 0.5 | 1.5 | 5.5 | 2.4 | 17.6 | 65.4 |
| Beach | 7.5 | 15.6 | 14.6 | 7.7 | 17.1 | 16.0 |
| Boat | 10.4 | 12.6 | 13.1 | 16.4 | 18.4 | 20.0 |
| Marching | 46.4 | 42.9 | 39.9 | 46.6 | 38.2 | 33.9 |
| Tennis | 73.9 | 69.8 | 68.7 | 77.9 | 69.7 | 62.0 |
| Court | 38.0 | 35.0 | 36.9 | 34.1 | 38.6 | 41.6 |
| Desert | 18.8 | 17.3 | 15.7 | 22.1 | 20.5 | 18.0 |

Table 3.2: Average precision values for 7 selected concepts from the *Mediamill Challenge* for different spatial bin sizes using a vocabulary size of 4,000 visual words.

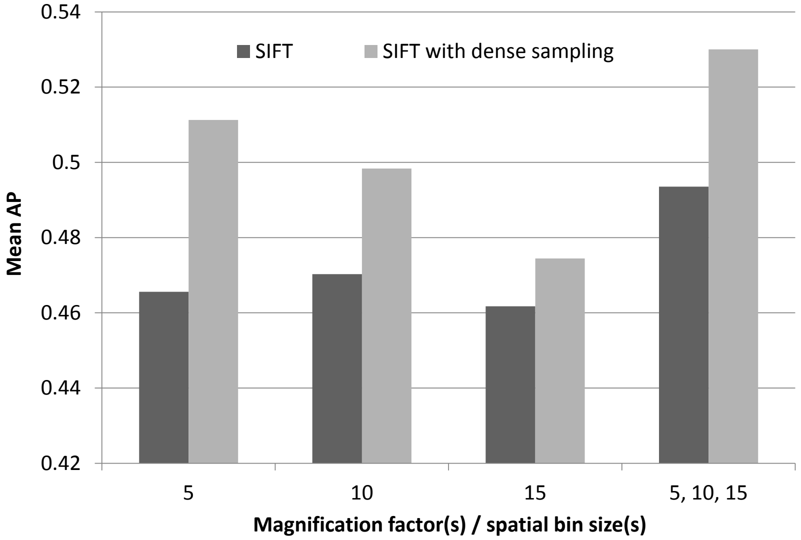


Figure 3.5: Evaluation of different spatial sizes on the *VOC Challenge* using spatial pyramids and a 4,000-dimensional vocabulary (averaged over 20 object classes).

improvements of up to 6.4% and 10.5%, respectively. The combination of all feature representations based on sparse and dense sampling achieved a mean AP of 54.1%.

### 3.2.5 Discussion

Experimental results on the *Mediamill* and on the *VOC Challenge* show that the concept detection performance can be significantly boosted by combining feature representations of different spatial sizes using MKL. Furthermore, the results indicate that the magnification factor which determines the spatial bin size depending on the keypoint scale should be chosen considerably larger than the default value of 3. In all experiments on the *Mediamill* as well as on the *VOC Challenge*, a magnification factor of 10 achieved the best detection performance. Due to the large visual variations within concept classes, larger patch sizes seem to result in a more generalizable representation. While small patch sizes only describe the near neighborhood of a keypoint, larger patch sizes describe rather coarse image structures. In the case of dense sampling, the impact of the bin size varies depending on the data set. While the concept detection results on the *Mediamill Challenge* also suggest larger bin sizes, the best performance on the object classification test set was already achieved using a bin size of 5. It seems that larger bin sizes are better suited for detecting scenes than for detecting objects. In general, it can be noticed that the best spatial bin size depends on the used dataset and most notably on the detected concept class.

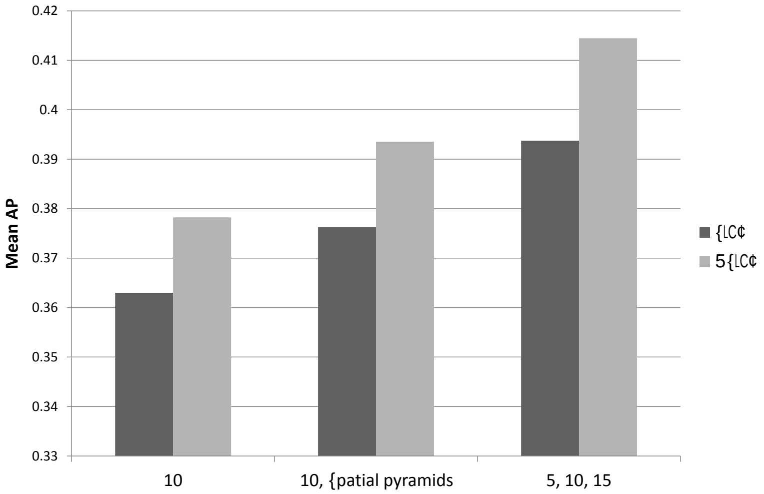


Figure 3.6: Comparison of spatial pyramids versus combining different magnification factors / spatial bin sizes on the *Mediamill Challenge* using a 4,000-dimensional vocabulary (averaged over 101 concepts).

In all experiments, the combination of different spatial bin sizes or magnification factors using MKL significantly improved the concept detection performance. These results show that the feature representations based on different spatial bin sizes are complementary.

Figure 3.6 depicts the performance improvement of the spatial pyramid representation versus the combination of different magnification factors / spatial bin sizes. The combination of different spatial bin sizes using MKL is much more effective than the use of spatial pyramids. Using a 4,000-dimensional vocabulary and a magnification factor, respectively, a spatial bin size of 10, the performance improvement of the spatial pyramid representation was only up to 3.8%. In contrast, the relative performance improvement of combining different spatial bin sizes was 7.8% in the case of sparse sampling and 8.7% in the case of dense sampling. While the storage complexity of the spatial pyramid representation adds up to 20,000-dimensional histograms, the combination of 3 different spatial bin sizes yields only 12,000-dimensional histograms per image or video shot.

The combination of different spatial bin sizes, different sampling strategies and spatial pyramids achieved state-of-the-art performances on the *Mediamill* as well as on the *VOC Challenge*, 43.2% and 54.1% mean AP, respectively. Considering further frames per shot on the *Mediamill Challenge* instead of only one keyframe, even 44.6% mean AP has been obtained. This is an improvement of over 100% compared to the baseline provided by the *Mediamill Challenge*. To the best of our knowledge, the best reported result for the same color features on this challenge is approximately 42% [Sande et al. 2008].

## 3.3 Multi-Modal Concept Detection

The detection of visual concepts in video shots is an essential prerequisite for semantic video retrieval, navigation and browsing. In addition to the visual modality, the audio signal of videos carries important information that can help to improve the retrieval performance of generic concept detection systems.

In previous approaches, audio features were used for visual concept detection, either by using low-level features such as MFCCs or by using detection results of specific audio events such as silence, speech, music and noise as mid-level features for subsequent training of video concept classifiers.

In the current section, the Bag-of-Words approach is leveraged for audio features to enhance video concept detection and MKL is proposed as the appropriate fusion scheme for these BoAW and state-of-the-art BoVW features.

The remainder of this section is organized as follows. Related work is reviewed in Section 3.3.1. In the following sections, the proposed system for multi-modal video concept detection is introduced. The application of the Bag-of-Words representation to audio features along with the classification scheme is presented in Section 3.3.2. Section 3.3.3 describes the state-of-the-art visual features and the MKL framework to combine the feature representations of both modalities. Experimental results are presented in Section 3.3.4. Finally, the results are discussed in Section 3.3.5.

### 3.3.1 Related Work

Most of the concept detection approaches that incorporate audio information directly use additional low-level features such as MFCCs, ΔMFCCs, pitch, zero-crossing rate, energy, or log-power to classify semantic concepts [Bredin et al. 2010; Gorisse et al. 2010; Li et al. 2010a]. For example, Bredin et al. [2010] extracted low-level features including MFCCs and their derivatives to build concept-specific GMMs.

In other approaches, the results of audio event detectors are additionally used as mid-level features. Besides acoustic events such as speech, non-speech, background and gender, Snoek et al. [2009] additionally detected the occurrence of 16 audio events such as “child laughter”, “baby crying”, “airplane propeller”, “sirens”, “traffic noise”, “car engine”, “dog barking”, or “applause” and used the results as additional input for concept classifiers like SVMs. Inspired by classical text document analysis, Lu and Hanjalic [2008] tried to automatically determine these audio elements by regarding them as natural clusters of the audio data. Between 2 and 20 elements are discovered per audio document using an iterative spectral clustering method.

The audio concept classification framework used by Feki et al. [2011] first removes segments of silence and then separates the audio signal into speech, music and environmental sound. The environmental sound segments are further classified using a time-frequency analysis based on MFCC features. For video concept detection, visual features and the previously described audio classification results are fed into a fuzzy reasoning system to fuse the different modalities [Elleuch et al. 2010].

Jiang et al. [2009a] introduced a novel representation called short-term audio-visual atoms. Audio features based on a matching pursuit representation [Mallat and Zhang 1993] of the audio signal and region-based color, texture, edge, and motion features are combined and a joint audio-visual codebook is built using multiple instance learning.

Inoue et al. [2010a] used a statistical framework to combine visual and audio features for video concept detection. The distribution of SIFT descriptors for each shot was described by GMMs, and a SVM with a GMM-kernel that compares GMMs was used for training and classification. In addition, hidden Markov models (HMM) were built for each concept based on audio features, including MFCCs, log-power and the corresponding derivatives. The final classification result is a weighted combination of log likelihood ratios from the audio models and from the SIFT GMMs. By additionally using audio features, the results for 20 semantic concepts on documentary films could be improved from 15% mean AP to 16.4%. At the *TRECVid* evaluation in 2010 [Smeaton et al. 2006], the GMM kernel was also applied for MFCC features resulting in a noticeable performance improvement for several concepts like “singing”, “dancing”, “cheering” or “animal” [Inoue et al. 2010b].

Peng et al. [2009] proposed a method that performs an audio-only analysis of the video data and investigated the use of an audio pLSA model for video concept detection. An audio vocabulary based on MFCC features from acoustically homogenous segments is built and the latent audio topics are discovered using pLSA. Each shot is then described by the probabilities of the discovered latent topics and classified by a SVM. Results are reported on 85 hours of news videos for 10 concepts from the *Mediamill Challenge*. Diou et al. [2010] combined BoW audio features based on MFCCs with visual features in an early fusion scheme. However, for the 30 evaluated concepts of the *TRECVid* 2010 semantic indexing task, the additional use of BoW audio features clearly decreased the performance from 4.5% to 3.5% mean infAP.

The BoAW approach has recently been successfully applied in the fields of music information retrieval and multimedia event detection. Riley et al. [2008] represented songs as a bag of auditory words showing robust results for a variety of signal distortions and Jiang et al. [2010b] combined Bag-of-Words representations for audio and visual features using a late fusion scheme to detect special events like “making a cake” or “assembling a shelter”.

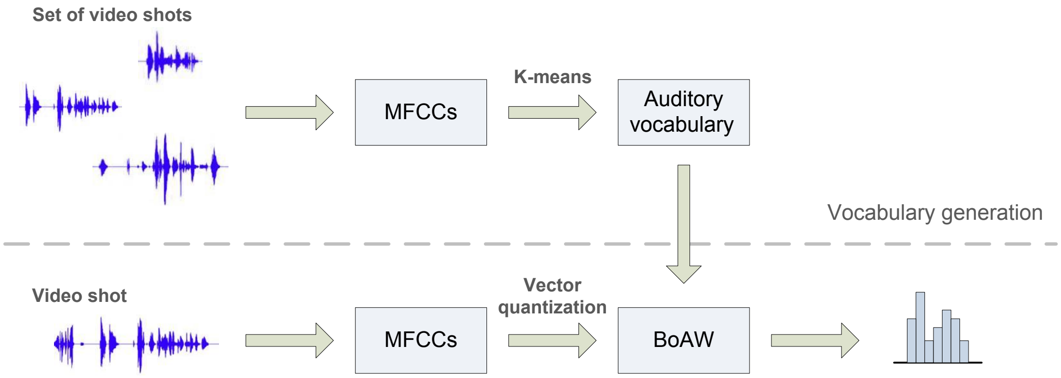


Figure 3.7: BoAW representation.

### 3.3.2 Bag-of-Auditory-Words

Since the BoVW representation based on local SIFT descriptors achieves superior performance in the field of visual concept detection [Smeaton et al. 2006], the Bag-of-Words paradigm is leveraged for audio features. Using a time-frequency analysis of the audio signal, 12-order MFCCs are extracted from audio frames of 20 ms length with an overlap of 50%. Thus, a video shot is represented as a set of 12-dimensional MFCC vectors. First, an auditory vocabulary, also called codebook, is generated based on these MFCC vectors using the K-means clustering algorithm, where the final cluster centers can be interpreted as "auditory words". Similar to the representation of documents in the field of text retrieval, a video shot can then be represented as a bag of auditory words which are the results of a vector quantization process using the generated vocabulary. Finally, a shot is described as a histogram, counting the occurences of auditory words. To diminish the quantization loss during histogram generation, the soft-weighting scheme, as described in Equation 3.1, is used. Instead of mapping a MFCC vector only to its nearest neighbor, the K nearest auditory words are selected.

Having proven to be powerful for visual concept detection [Smeaton et al. 2006], SVMs are used to build audio models and to classify video shots. Therefore, a kernel function needs to be specified, which intuitively measures the similarity between two data instances. Commonly used kernels are already described in Section 2.5.2. Besides the linear and the RBF-kernel (Equations 2.27 and 2.29), the $\chi^2$-kernel, as already described in Equation 2.30, is applied because of the used histogram representations.

### 3.3.3 Multi-Modal Fusion

In a multi-modal fusion setting, BoAW features are combined with state-of-the-art visual features. As visual features the BoVW representation is used. Densely

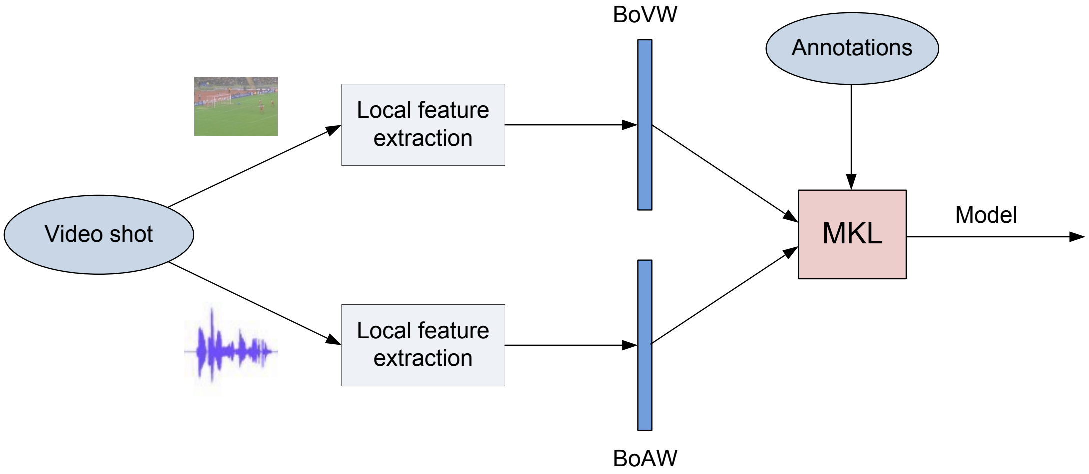


Figure 3.8: Multi-modal fusion of BoVW and BoAW features.

sampled local SIFT descriptors from the keyframes are extracted using the implementation of the `VLFeat library` [Vedaldi and Fulkerson 2010]. Color information is integrated using RGB-SIFT, where the SIFT descriptors are computed independently for the three channels of the RGB color model. Thus, the final local feature vector is the concatenation of the individual descriptors. Based on these local descriptors, a global visual vocabulary is generated using the K-means algorithm. Each keyframe or rather shot is described as a histogram indicating the presence of each visual word. Again, the previously described soft-weighting scheme is applied to consider the similarities of the local descriptors to the codebook entries.

The easiest way to combine BoAW and BoVW features is the early fusion scheme. Using this method, visual and audio features are simply concatenated and directly fed into a SVM. A more sophisticated method to combine the capabilities of different modalities is the MKL approach as described in Figure 3.8. Since audio information is more or less important depending on the semantic concept, e.g., there is no discriminative audio information for the concept house, MKL is applied to find an optimal kernel weighting

$$k_{multimodal} = \alpha \cdot k_{audio} + \beta \cdot k_{visual} \quad \text{with} \quad \alpha \geq 0,\ \beta \geq 0 \tag{3.6}$$

where the kernel functions $k_{audio}$ and $k_{visual}$ take both feature modalities into account. The $L_2$-norm is used to control the sparsity of the weights $\alpha$ and $\beta$ for audio and visual features, respectively. Throughout our experiments, the MKL framework provided by the `Shogun library` [Sonnenburg et al. 2010] is used in combination with the SVM implementation of Joachims [1998], called SVM$^{\text{light}}$.

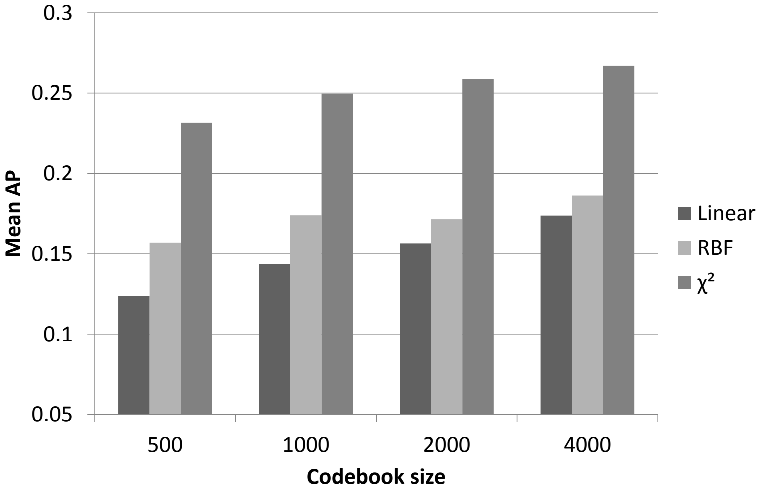


Figure 3.9: Performance evaluation of different kernel functions and codebook sizes using BoAW features.

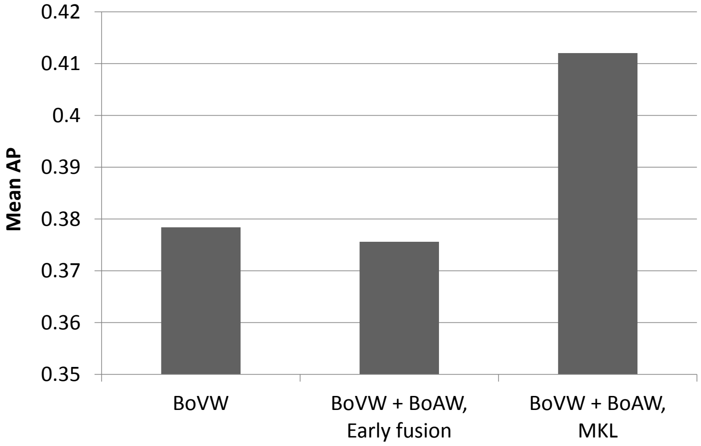


Figure 3.10: Performance evaluation of BoAW features in a multi-model setting using early fusion and MKL.

### 3.3.4 Experimental Results

In this section, the performance impact of BoAW features in the field of video concept detection is investigated. For this purpose, the Mediamill Challenge [Snoek et al. 2006b] is used. It offers an extensive dataset with 101 annotated concepts, including objects, scenes, events and personalities.

Several experiments have been performed to investigate the performance impact of BoAW features, both alone and in combination with visual features.

In a first experiment based on an audio-only analysis of the data, different auditory vocabulary sizes and kernel methods have been taken into account. The

| AP [%] | BoVW | BoVW+BoAW |
|---|---|---|
| Motorbike | 0.3 | 4.1 |
| Cycling | 13.7 | 91.7 |
| Racing | 11.4 | 52.3 |
| Bicycle | 17.6 | 80.0 |
| Baseball | 0.7 | 1.6 |
| Natural disaster | 8.7 | 18.0 |
| Boat | 18.3 | 34.1 |
| Golf | 36.4 | 51.3 |
| Waterbody | 36.9 | 49.4 |
| Aircraft | 16.6 | 21.8 |
| Football | 54.6 | 70.6 |
| River | 69.9 | 89.8 |
| Entertainment | 55.0 | 70.2 |
| Sports | 49.3 | 62.2 |
| Table | 10.9 | 13.7 |
| Food | 52.6 | 64.2 |
| Basketball | 54.6 | 65.7 |
| Soccer | 72.4 | 85.6 |

Table 3.3: Performance comparison between the visual-only baseline system and the multi-modal system using MKL, showing AP values of concepts with relative performance improvements of at least 18%.

linear, RBF and $\chi^2$-kernel have been compared in combination with codebook sizes between 500 and 4,000 auditory words. The experimental results are presented in Figure 3.9. The $\chi^2$-kernel significantly outperforms the linear as well as the RBF-kernel. Using 4,000 auditory words, the $\chi^2$-kernel yields a relative performance improvement of 43.3% compared to the RBF-kernel. Additionally, a larger vocabulary has a positive impact on the overall performance. In combination with the $\chi^2$-kernel, a vocabulary size of 4,000 auditory words achieves a mean AP of 26.7% compared to 23.2% for 500 words. Based on these results, the $\chi^2$-kernel and a vocabulary size of 4,000 auditory words are used exclusively in the experiments below.

In a second experiment, the impact of BoAW features in a multi-modal concept detection system is investigated. The state-of-the-art baseline system performs a visual-only analysis of the data using densely sampled RGB-SIFT descriptors with a vocabulary of 4,000 visual words. Both modalities, visual and audio features, are combined using MKL on the one hand and a simple early fusion scheme on the other hand. In order to save computation time, the models are trained using a reduced number of negative training samples per concept. The results of the two different fusion strategies are presented in Figure 3.10. While the early fusion strategy causes a slight performance decrease, the fusion of visual and audio features via MKL achieves a relative performance improvement of 8.9% compared to the baseline system. In total, 31 concepts yield a relative performance improvement of more than 10%. In particular, the concepts

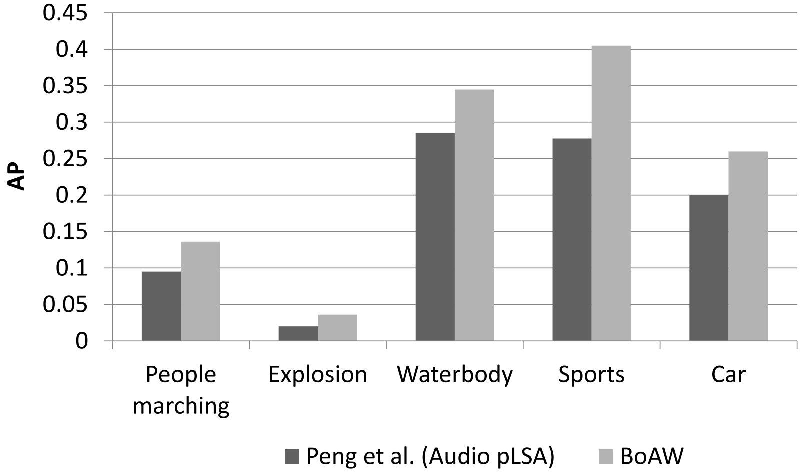


Figure 3.11: Performance comparison between the BoAW method and the audio pLSA approach of Peng et al. [2009].

representing personalities profit from the additional audio features, increasing the mean AP for this group of concepts from 9.2% to 11.1%. Further concepts with relative improvements of at least 18% are shown in Table 3.3.

### 3.3.5 Discussion

The experiments indicate that the kernel choice is a critical decision for the performance of the BoAW approach. While the RBF-kernel concentrates on the largest histogram differences due to the quadratic exponential decay, the $\chi^2$-kernel considers the bins more equally. This seems to be beneficial regarding the large intra-class variations of audio signals. Keeping in mind that the ground truth annotation of the 101 semantic concepts is based upon a visual inspection of the video shots, the BoAW approach achieves an impressive performance of 26.7% mean AP on the Mediamill Challenge data. The performance is even significantly better than the baseline system provided by the Mediamill Challenge with 21.6% mean AP, which uses local as well as global texture information. The state-of-the-art approach of Peng et al. [2009] relying on audio pLSA attained a mean AP of approximately 20.7% on a subset of 10 concepts from the MediaMill Challenge. On the same subset the proposed system achieves a superior performance of 26.8% mean AP using BoAW features, yielding a relative improvement of approximately 30%. Besides the mean AP, Peng et al. [2009] displayed AP scores for half of the ten concepts. For these concepts, performance comparisons between the BoAW method and the audio pLSA approach are shown in Figure 3.11.

Additionally using BoAW features via MKL clearly improves the performance of the state-of-the-art video concept detection system that relies on visual features only. The weak performance of the early fusion strategy confirms the results of Diou et al. [2010] at the *TRECVid Challenge* 2010, where the additional use of BoW audio features in an early fusion scheme clearly decreased the performance. This is not surprising since audio information is more or less important depending on the semantic concept. While the concepts “racing” or “motorbike”, for example, are characterized by engine noise, there is no discriminative audio information for concepts such as “house” or “grass”. In this case, audio features can be even misleading for the classification process. MKL instead of early fusion learns optimized kernel weights that provide information about the relevance of both modalities for the discrimination of semantic concept classes. Hence, audio features are more or less considered, depending on the corresponding concept.

## 3.4 Color Moment SIFT

Color information can be very helpful in classifying visual concepts, like for example “sunset”, “meadow” or “sky”. It contributes to improve the discriminative power of concept classifiers. State-of-the-art concept detection approaches embed color information at the local descriptor level.

Current approaches combine SIFT descriptors with local color histograms [Weijer and Schmid 2006] or concatenate SIFT descriptors extracted from different color channels, like Colored SIFT (CSIFT) [Abdel-Hakim and Farag 2006], RGB-SIFT or transformed color SIFT [Sande et al. 2010].

In this section, SIFT descriptors and local color moments are concatenated to embed color information at the local descriptor level. In the following, this local descriptor is called color moment SIFT (CMSIFT). CMSIFT descriptors offer several advantages: The extraction of CMSIFT descriptors is clearly faster than the extraction of SIFT features from multiple color channels, like RGB-SIFT or transformed color SIFT. The local feature dimension of CMSIFT descriptors is relatively small. Only six dimensions are added to the 128 dimensions of the SIFT descriptor, resulting in 134 dimensions. While Weijer and Schmid [2006] used 165 dimensions for the hueSIFT descriptor (37 extra dimensions for the color histogram), RGB-, HSV- and transformed color SIFT descriptors even result in 384 dimensions. The local descriptor dimension has a direct impact on the runtime of the histogram generation process which is the most time-consuming step of the BoVW approach. Doubling the feature dimension leads to a two times higher runtime.

This section is organized as follows. Related work is presented in Section 3.4.1. The extraction of CMSIFT features is described in Section 3.4.2. Section 3.4.3 shows the experimental results followed by the discussion in Section 3.4.4.

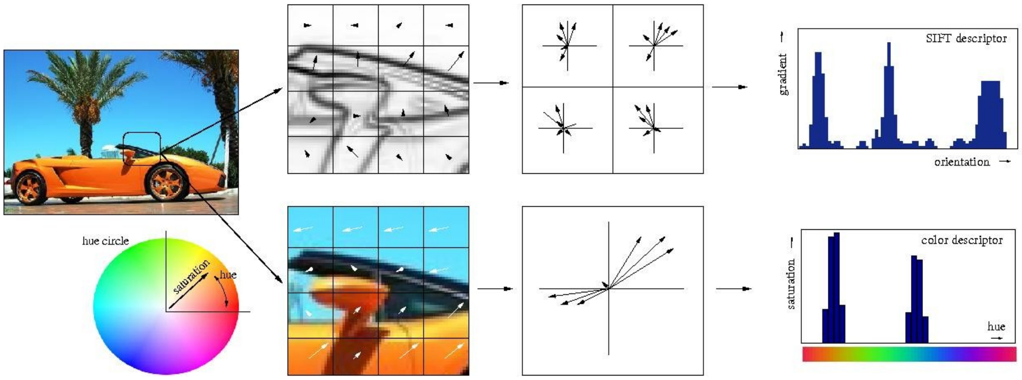


Figure 3.12: HueSIFT, source [Weijer and Schmid 2006].

### 3.4.1 Related Work

An important property of SIFT descriptors is the robustness against illumination changes. However, SIFT descriptors are extracted from gray-scale images, and thus ignore color information. Two different strategies have been pursued in the literature to extend SIFT descriptors: the attachment of local color histograms at the descriptor level and the extraction of SIFT descriptors from different color channels.

In an early fusion scheme, Weijer and Schmid [2006] added local color histograms at the descriptor level as described in Figure 3.12. This concatenation of a local SIFT descriptor and a local hue histogram is called hueSIFT. Photometric robustness is achieved by using the hue value. Therefore, colors of the RGB (red, green, blue) color model are transformed into the opponent color space:

$$\begin{aligned} O_1 &= \frac{R-G}{\sqrt{2}} \\ O_2 &= \frac{R+G-2B}{\sqrt{6}} \\ O_3 &= \frac{R+G+B}{\sqrt{3}} \end{aligned}$$

where $O_1$ and $O_2$ encode the chrominance information and $O_3$ represents the luminance component. While the hue value refers to the angle of the chrominance vector

$$\text{hue} = \arctan \frac{O_1}{O_2}, \tag{3.7}$$

the saturation value is determined by its magnitude

$$\text{saturation} = \sqrt{O_1^2 + O_2^2}. \tag{3.8}$$

But hue values are unstable at low saturations: the lower the saturation the higher the uncertainty of the hue value. Therefore, each hue value is weighted in the histogram generation process by its saturation to deal with this uncertainty [Weijer and Schmid 2006]. This method is similar to the generation of edge histograms, where the counted edge orientations are weighted by their magnitude.

The second strategy for color integration is the extraction and concatenation of SIFT descriptors from separate color channels. Thus, the dimension of a local descriptor is increased by the factor corresponding to the number of channels. This strategy was first used by Abdel-Hakim and Farag [2006] building SIFT descriptors in an opponent color invariant space, called CSIFT. Various color models lead to different local feature representations, like RGB-SIFT, Opponent-SIFT or HSV-SIFT. A comprehensive comparison of different color features for visual concept classification has been presented by Sande et al. [2008, 2010]. They showed that color SIFT variants perform substantially better than color features like local color histograms or color moments. Moreover, local descriptors like RGB-SIFT or transformed color SIFT clearly outperformed hueSIFT features. Furthermore, they argued that due to the normalizations during the SIFT feature extraction, RGB-SIFT is equal to the transformed color SIFT descriptor and thus invariant against light intensity and color changes [Sande et al. 2008]. In their experiments, Opponent-SIFT and RGB-SIFT achieved the best performance on the image classification task of the *VOC Challenge* [Everingham et al. 2010] as well as on the *Mediamill* video benchmark [Snoek et al. 2006b]. Moreover, Sande et al. [2010] revealed that a late fusion of different color descriptors leads to a further performance improvement.

### 3.4.2 CMSIFT Descriptor

For the CMSIFT descriptor, local color moments are combined with local SIFT descriptors. While the SIFT descriptor is computed from the gray-scale image, local color moments are extracted from the transformed color space. Compared to the RGB or HSV color space it has the advantage that it is invariant to photometric changes, including changes and shifts of light color and light intensity. Therefore, the RGB values of an image are converted to the transformed color space in the following way:

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \frac{R-mean_R}{stdev_R} \\ \frac{G-mean_G}{stdev_G} \\ \frac{B-mean_B}{stdev_B} \end{pmatrix} \tag{3.9}$$

where $mean_k$ and $stdev_k$ are the mean value and the standard deviation, respectively, of the image channel $k$.

The local color moment features are extracted from the same patch size as the corresponding SIFT descriptors. The color values are also weighted by a Gaussian window diminishing the influence of pixels further away from the keypoint center. Furthermore, the Gaussian window avoids that small changes of the keypoint position can trigger large changes of the resulting descriptor.

The first two local color moments of a keypoint at position (x,y) for the three color channels of the transformed color model are defined as:

$$\mu_k = \sum_{i=0}^{4 \cdot b} \sum_{j=0}^{4 \cdot b} w_{i,j} \cdot c^k_{x-2b+i,\, y-2b+j} \tag{3.10}$$

$$\sigma_k = \sqrt{\sum_{i=0}^{4 \cdot b} \sum_{j=0}^{4 \cdot b} w_{i,j} \cdot \left(c^k_{x-2b+i,y-2b+j} - \mu_k\right)^2} \tag{3.11}$$

where $c^k_{i,j}$ is the pixel value at position (i,j) of the k-th color channel, $w$ is a precomputed matrix of Gaussian weights and $b$ is the spatial bin size of the corresponding SIFT descriptor.

Finally, the SIFT feature vector $d$, the mean vector $\mu$ and the standard deviation vector $\sigma$ are concatenated to form the final CMSIFT descriptor $(d^T,\, \gamma \cdot \mu^T,\, \gamma \cdot \sigma^T)^T$ where $\gamma$ is a balancing factor to control the influence of the color components.

### 3.4.3 Experimental Results

The experiments are conducted on the *Mediamill* [Snoek et al. 2006b] and on the Pascal *VOC Challenge* [Everingham et al. 2010]. Within the concept detection setting the proposed CMSIFT features are compared to other state-of-the-art color SIFT variants.

The images of the *VOC Challenge* are scaled to a size of 90,000 pixels to speed up the subsequent feature extraction process. In a preprocessing step visual vocabularies have to be build for each local descriptor type as well as for different balancing factors. In order to create the visual vocabularies, local descriptors are extracted from a set of randomly selected training images using dense sampled keypoints with a step size of 10 pixels. On the *Mediamill Challenge* 30 keyframes are randomly chosen per concept class. After removing duplicates 2,213,060 local descriptors are extracted from the underlying 2,830 keyfames. At the *VOC Challenge* 100 images are randomly selected from each concept class, resulting in 1,557,016 descriptors from 1,852 images. 4,000-dimensional visual vocabularies

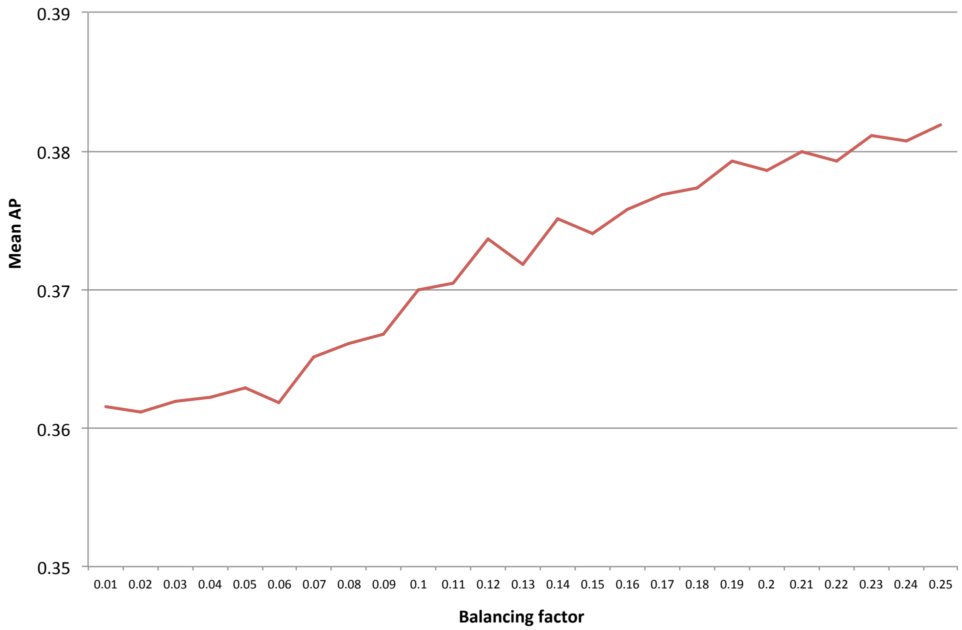


Figure 3.13: Performance evaluation of CMSIFT descriptors with different balancing factors on the *Mediamill Challenge* [Snoek et al. 2006b].

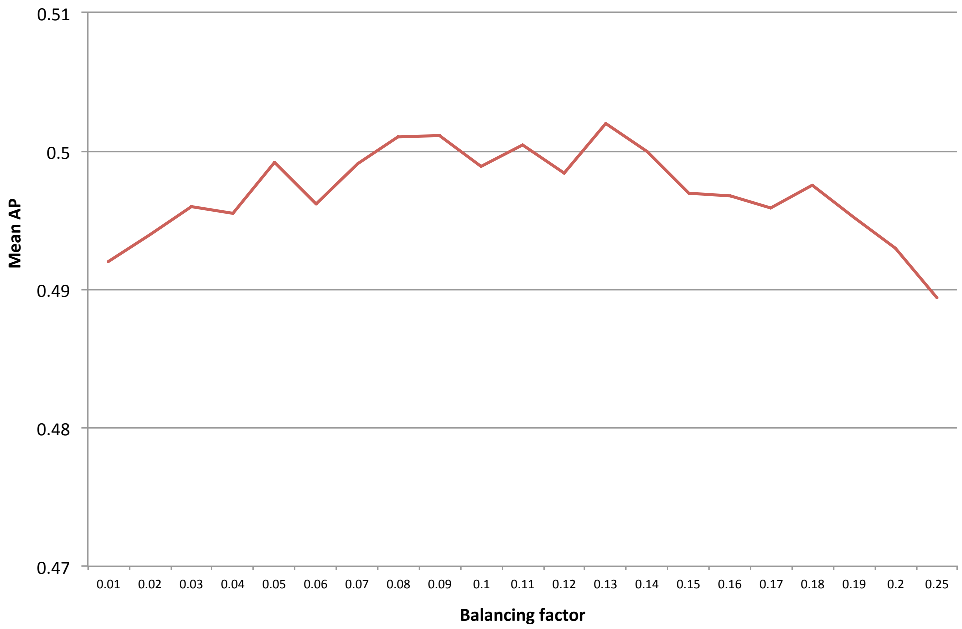


Figure 3.14: Performance evaluation of CMSIFT descriptors with different balancing factors on the *VOC Challenge* [Everingham et al. 2010].

are constructed using the K-means algorithm. For the initialization of the clustering process the approach of Arthur and Vassilvitskii [2007] is applied, which tries to find a good coverage of the data and improves upon random selection. Totally, a maximum number of 30 iterations is performed to refine the cluster centers.

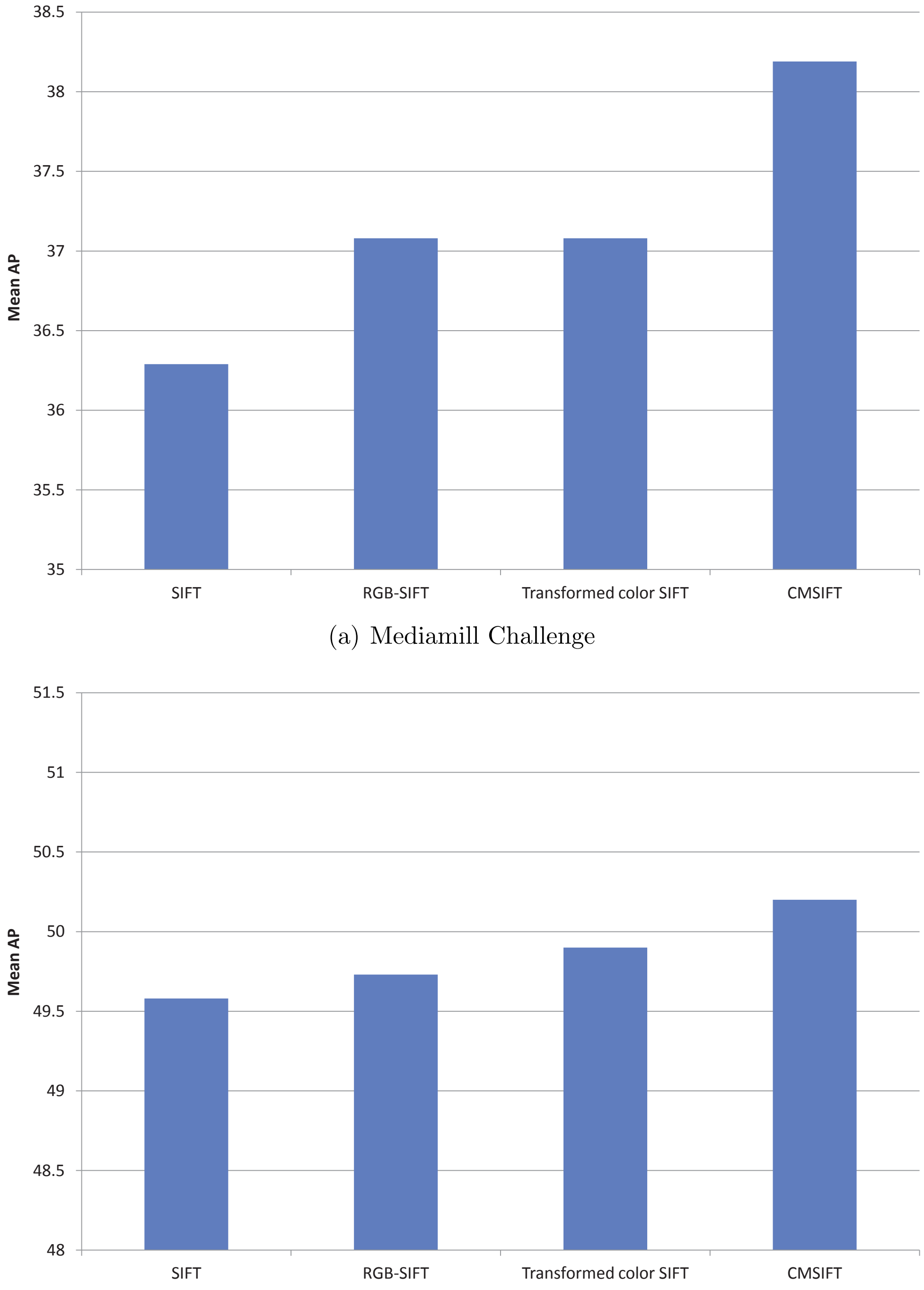
Throughout the experiments a dense sampling strategy with a step size of 5 pixels is performed and the HoVWs are generated using the soft weighting scheme, as described in Equation 3.1, followed by a $L_1$-normalization. The concept models are built using the SVM$^{\text{light}}$ implementation of Joachims [1998] with a $\chi^2$-kernel.

In a first experiment, different balancing factors are investigated as shown in Figures 3.13 and 3.14. The best performance is achieved using a balancing factor of 0.25 on the *Mediamill Challenge* and 0.13 on the *VOC Challenge*. These values are used as balancing factors in the following experiments.

In a second experiment, a comparison is made between CMSIFT, RGB-SIFT and transformed color SIFT descriptors. While CMSIFT achieves comparable results on the *VOC Challenge*, it even outperforms the other color SIFT variants on the *Mediamill Challenge* (Figure 3.15).

Furthermore, the runtimes for the feature extraction process are measured on a 2 GHz Dual-Core AMD Opteron$^{TM}$ CPU. Table 3.4 shows the runtimes for the local descriptor extraction as well as for the coding and pooling step. It can be observed that the runtime for the extraction of CMSIFT descriptors increases only slightly compared to SIFT descriptors. This shows that the extraction of SIFT descriptors is much more expensive than the calculation of local color moments. Since SIFT descriptors in the case of CMSIFT are only extracted from gray-scale images instead of multiple color channels, the local feature extraction process can be speeded up by approximately a factor of three. Additionally, the lower dimension of the CMSIFT descriptors has a positive impact on the runtime of the coding step. The nearest-neighbor search for mapping local descriptors to the visual vocabulary depends directly on the feature dimension of the local descriptors. Thus, the runtime for the coding and pooling step of CMSIFT descriptors is also approximately three times faster than for RGB-SIFT and transformed color SIFT descriptors. Altogether, the whole feature extraction process is around three times faster than using state-of-the-art RGB- or transformed color SIFT descriptors.

Furthermore, the length of the local descriptors clearly effects the runtimes for building the visual vocabularies, too. The vocabulary generation for CMSIFT descriptors including the extraction of the necessary local descriptors takes considerably less than half of the runtime compared to RGB- and transformed color SIFT vocabularies.



(a) Mediamill Challenge

(b) VOC Challenge

Figure 3.15: Comparison of CMSIFT with other color SIFT variants.

| | **Local descriptor extraction** | **Coding and pooling** |
|---|---|---|
| SIFT | 0.91 s | 3.47 s |
| CMSIFT | 0.93 s | 3.60 s |
| RGB-SIFT | 2.72 s | 10.78 s |
| Transformed color SIFT | 2.82 s | 14.50 s |

Table 3.4: Runtimes per image for the extraction of the local descriptors and for the coding and pooling step.

### 3.4.4 Discussion

Color information can be very helpful for the detection of visual concepts. Regarding individual concepts the integration of color information using CMSIFT improves the detection performance in terms of AP for most concepts. This is not surprising since color information is not discriminative or beneficial for every concept, in special cases it can be even misleading.

Important properties of the CMSIFT descriptor are its photometric robustness and the compact representation. Experimental results on the *VOC* and on the *Mediamill Challenge* show that CMSIFT achieves similar or even better concept detection performance compared to the state-of-the-art RGB and transformed color SIFT descriptors while being at the same time much more faster in the local descriptor extraction and coding process. Altogether, the proposed CMSIFT descriptor is the most compact color SIFT variant, resulting in a significantly faster feature extraction process.

## 3.5 Summary

In this chapter, enhanced local descriptors for BoVW concept detection systems have been proposed.

First, the impact of the spatial extents of SIFT descriptors for visual concept detection has been investigated. It turned out that the magnification factor that determines the spatial bin size depending on the keypoint scale should be much larger than the normally used default value. Based on the observation that SIFT descriptors with different spatial extents yield large performance differences, a system has been proposed that combines feature representations based on different magnification factors and different spatial bin sizes, respectively, using MKL. Experimental results on the *Mediamill* as well as on the *VOC Challenge* have demonstrated that these feature representations complement each other: The concept detection performance was significantly boosted by combining different spatial sizes of local descriptors using MKL.

Second, a multi-modal approach for video concept detection has been presented that models MFCC features in an auditory vocabulary. The auditory vocabulary is used to describe video shots via histograms of auditory words. SVMs are employed to build the audio models and to finally classify the video shots. Experimental results on a large set of 101 semantic concepts have shown the effectiveness of the proposed approach. The proposed system relying on BoAW features outperforms a state-of-the-art audio approach that uses pLSA [Peng et al. 2009] and is even significantly better than the baseline system provided by the *Mediamill Challenge*, which used local as well as global texture features. Furthermore, the resulting BoAW features are combined with BoVW features via MKL. Using MKL instead of an early fusion scheme significantly improves the results of a state-of-the-art video concept detection system that relies on visual features only.

Finally, CMSIFT descriptors have been introduced representing an effective method for the SIFT-based integration of color information. Color information can be very helpful to classify visual concepts, like for example “sunset”, “meadow” or “sky”. It contributes to improve the discriminative power of concept classifiers. CMSIFT descriptors combine SIFT descriptors with local color moments and are invariant to photometric changes. Experimental results on the *VOC* and on the *Mediamill Challenge* show that CMSIFT descriptors achieve similar or even better concept detection performance compared to state-of-the-art RGB- and transformed color SIFT descriptors while being at the same time much faster.

*"Time is free, but it's priceless.*
*You can't own it, but you can use it."*

Harvey MacKay

# 4

# Improving Concept Detection via Object-Based Features

## 4.1 Introduction

In this chapter, an approach for the task of visual concept detection is presented which systematically utilizes results of object detectors. Based on the observation that the use of face detection results improved the performance of several face related concepts [Mühling et al. 2007b, 2008], the hypothesis is that it is beneficial to utilize further reliable object detectors trained on separate data sets. These object detectors can be used for two purposes: (a) to directly provide retrieval results (such as using a face detector for the concept face), (b) to provide semantic features as additional input for the SVM-based concept classifiers. Thus, other related concepts can also profit from object-based features. For example, a car drives mostly on a road and a television screen normally does not appear in an outdoor setting. Using object detectors trained on separate data sets, semantic features are generated based on the object detection results. For videos detection results are assembled to object sequences and a shot-based confidence score as well as further features, such as position, frame coverage or movement, are computed for each object class.

Experimental results on the *Mediamill*, *VOC* and *TRECVid Challenge* show clear improvements in terms of retrieval performance not only for the object classes, but also and in particular for a large number of indirectly related concepts.

Furthermore, the generalization capabilities of concept models across different domains are addressed, constituting a severe problem in the field of concept detection. Object-based features are proposed to improve the cross-domain concept detection performance. In several cross-domain experiments using different news channels and genres the generalization capabilities of object-based features are investigated.

The remainder of the current chapter is organized as follows: Related work is presented in Section 4.2. Section 4.3 describes the used state-of-the-art object detection approaches. Class-specific *Hough forests* are another approach, that has been recently proposed for object detection. In Section 4.4 this method is extended to a multi-class approach to speedup the detection of multiple object classes. Section 4.5 explains the generation of object sequences in the case of videos and the derivation of feature representations from the object detection results. Extensive experimental results on the *Mediamill* and *VOC Challenge*, the participations at the annual *TRECVid Challenge* as well as channel and genre cross-domain experiments are presented in Section 4.6. Finally, Section 4.7 summarizes the chapter.

Parts of this chapter have been published by Mühling et al. [2007b, 2008, 2009a,b, 2010, 2011a,c].

## 4.2 Related Work

Current systems for generic concept detection mainly rely on BoVW features and in some cases additionally on features based on face detection, optical character recognition and/or speech recognition in the case of videos. For the *TRECVid* 2006 evaluations, Over et al. [2007] summarized that the 30 participants used the following feature types in the concept detection task: color, texture, shape, edges, acoustic, face and text.

Up to now, there have been only a few systems that integrate object-based features. For example, the *Semantic Pathfinder* [Snoek et al. 2006a] uses car, face and text detection results. The Semantic Pathfinder represents a generic multimedia indexing system for videos from the authors' perspective. Three consecutive analysis steps are derived from the authoring metaphor. Multi-modal features are analyzed on three different levels: namely content, style, and context. In the content analysis step segmentation-based features, called proto-concepts, and textual features from transcribed speech are extracted. The second step is the style analysis step. During this step, a video is viewed from the production perspective, considering features like "shot length", "shot size", "overlaid text", "camera motion" and "faces". In the last step, the previously detected concepts are analyzed in their semantic context. Finally, the Semantic Pathfinder explores the possible paths through the three consecutive analysis steps and chooses the path providing the highest performance based on a three-fold cross-validation.

Meanwhile, a similar idea to the proposed approach has been picked up by Li et al. [2010b]. In contrast to our approach, where fewer, but more reliable and sophisticated object detectors are used, the system of Li et al. [2010b], called

Object Bank, is aimed at using thousands or even millions of object detectors to describe the image content. Therefore, semantic feature sparsification methods are proposed to deal with high dimensional feature vectors. Furthermore, Li et al. [2010b] used spatial pyramid representations to assemble object detection results, while our approach considers spatial information via spatial coordinates.

Generally, concept detection is a multi-class multi-label problem which is typically broken down into binary classification problems. These binary classifiers are used to detect each individual concept, ignoring inherent correlations between the concepts. Semantic context can be modeled at different stages of the concept detection process. In the literature it can be distinguished between three types of context-aware approaches.

The easiest way is to exploit concept relations in a post-processing step like for example in the concept detection system of Elleuch et al. [2010]. After classifying each concept separately, a learned context model is used to refine the classification results. Chang et al. [2005] described another system using a parts-based statistical approach to represent an entire keyframe as an attributed relational graph. The parts-based concept classifiers trained on weakly labeled data were combined with conventional concept classifiers in a late-fusion scheme.

The second approach considers semantic context in a two-stage classification process. The context-based models are built on top of the independent binary detectors. Therefore, the results of the independent detectors are combined in a context vector. In the following stage, these context vectors are used to learn individual context models for each concept. In this setting, two models have to be learned per concept and detection errors of unreliable concept detectors of the first step are propagated to the second fusion step.

Qi et al. [2007a] have proposed a third approach that simultaneously classifies concepts. Class-specific information and correlations between the concepts are modeled in a single step by using a novel correlative multi-label framework. This approach is reminiscent of the multi-class instantiation of the structured output SVM from Tsochantaridis et al. [2004]. In the correlative multi-label framework the individual concepts as well as the concept correlations have to be modeled in the feature space, which results in an extraordinary high feature vector dimension. In their experiments they used 200 dimensional low-level features and 39 concepts resulting in 18,564 dimensions. Keeping in mind that state-of-the-art BoVW representations tend towards 4,000 dimensions, this already leads to 314,964-dimensional feature vectors. Using thousands of concepts further increases the feature dimension. The sparsity of this vector for an image or video shot depends on the number of occurring concept classes. Especially frequently occurring concepts additionally reduce the sparsity of the feature vector.

Continuous progress has been reported in the field of concept detection using BoVW approaches. These approaches have been described in detail in Chapter 2. Nevertheless, the success of concept detection systems is limited, especially across different domains. Yang and Hauptmann [2008] have shown that standard

SVM-based concept detection approaches learn little beyond memorizing most of the positive training data and thus generalize poorly to domains other than the training domain.

Yang et al. [2007b] investigated methods for adapting existing concept classifiers to domains other than the training domain. They proposed adaptive SVMs to adapt one or more existing classifiers to a new sparsely-labeled data set. An objective function similar to SVMs has been used to learn a “delta function” between the original and adapted classifier. To select the best existing classifier for adaptation, the performance is estimated by analyzing the score distribution on a few labeled instances of the new data.

Thus, it is not only difficult to apply concept models to video domains other than the training domain, but also to include training data from foreign domains into the learning process.

Altogether, the most important factor in the design of concept detection systems are robust feature representations which are necessary to provide the basis for successful within- and cross-domain concept detection.

## 4.3 Object Detection

The task of finding a given object category in an image or video sequence has received considerable attention in the literature. For videos, special object recognition approaches exist that rely on a motion-based segmentation of the video objects (e.g., [Kopf 2006; Kühne et al. 2001]). However, these methods do not detect non-moving objects. While early object detection approaches were sensitive to real world imaging conditions, such as pose and occlusion, significant progress has been made in recent years [Everingham et al. 2010]. In the following, state-of-the-art object detection approaches utilized within the experiments of this chapter are presented. A generic object detection approach and two special object detectors for face and text are applied to find object appearances in images and videos.

The object detection approach introduced by Felzenszwalb et al. [2008, 2010b] has been chosen, since it achieves superior performance at the object detection task of the annual PASCAL VOC Challenge. Key strategies of this approach are a discriminative training method with latent object locations and part placements, robust *Histogram of Oriented Gradients* (HOG) features and a fast matching procedure of pictorial structures. The approach is an extension of the person detector of Dalal and Triggs [2005], which is based on HOG features. Instead of using only one filter per object class, an object model consists of a global template that covers the whole object, several smaller part templates, and a model describing the spatial arrangement of the smaller parts. This kind of object model which uses parts and connections between parts is also called *pictorial structure* [Felzenszwalb and Huttenlocher 2005]. Felzenszwalb et al. [2010b] introduced a fast matching procedure of pictorial structures that uses dynamic programming combined with a generalized distance transform.

At the detection stage a sliding window is used to find object appearances. The confidence score at a particular image position and scale is the sum of the root filter score plus the scores of the matched part filters minus a deformation cost, whereby the part filters are applied at twice the spatial image resolution as the root filter.

The models are trained in a discriminative setting with positive and negative labeled training examples $D = (\langle x_1, y_1 \rangle, \ldots, \langle x_N, y_N \rangle)$ where $N$ is the number of training instances, $x_i$ are the HOG pyramids of the image regions, and $y_i$ are the corresponding class labels with $y_i \in \{-1, +1\}$. The vast amount of potentially negative examples is limited by data mining "hard negatives". Not only the object parts, but also the exact locations of the objects itself are treated as latent variables during the learning process. For each training instance $x$, the best scoring part placements $z$ are found using a function of the form:

$$f_\beta(x) = \max_z \beta \cdot \Phi(x, z) \tag{4.1}$$

where $\beta$ represents the model parameters, $z$ the latent variables and $\Phi(x, z)$ the feature vector. The model parameters consist of the filter values and the weights for the deformation costs. The feature vector itself is the concatenation of the HOG features from the root and part locations. Additionally, the vector includes the part deformation features. A generalization of the SVM, called latent variable SVM, is introduced to learn the object models by minimizing the following objective function:

$$L_D(\beta) = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} max(0, 1 - y_i f_\beta(x_i)). \tag{4.2}$$

where C is the penalty factor for misclassified examples. In practice the problem is solved by an iterative procedure, which alternates between finding the best fitting placement for each positive example and optimizing $\beta$ with a linear SVM.

Furthermore, mixtures of deformable part models are considered. The underlying idea is that each component of a mixture represents a different kind or view of the object class. Components are initialized by merging bounding boxes with similar aspect ratios into the same component. During the training process the membership information is also treated as a latent variable.

To further accelerate the detection process, a cascade algorithm is proposed [Felzenszwalb et al. 2010a]. A model with m parts leads to a cascade with m models where the root filter represents the weakest model. The complexity of the cascaded models is increased by iteratively adding the object part filters. Only locations with high-scoring weak models are further propagated through the object detection cascade, whereby the scores of weaker models are reused. Using object detection cascades the computation for most of the image locations is pruned after evaluating the first few models. This leads to an average speedup by a factor of more than 20 with negligible decrease in accuracy.

Within the experiments of this chapter, two further object detection approaches are applied.

First, frontal and profile face appearances are detected using the face detector provided by the `OpenCV library` [Bradski 2000]. This face detector is an implementation of the approach suggested by Viola and Jones [2004] with extensions as proposed by Lienhart et al. [2003]. The Adaboost-based approach of Viola and Jones [2004] uses Haar wavelet features, which are calculated based on integral images. It has been chosen since it is a very fast method that nearly operates in real-time on today's computers and thus can even be applied to every single frame of a video sequence. Since this approach usually reports many detections for a face of slightly different sizes and positions, an average rectangle is computed based on the reported detections, and the number of detections is used as confidence score.

Second, for overlaid text a special approach as introduced by Gllavata and Ewerth [2004] is used. It can automatically detect horizontally aligned text with different sizes, fonts, colors and languages. For this purpose, a wavelet transformation is applied to a video frame and the distribution of high-frequency wavelet coefficients is considered to statistically characterize text and non-text areas. Then, the K-means algorithm is used to classify text areas in the image. The detected text areas undergo a projection analysis in order to refine their localization.

## 4.4 Multi-Class Object Detection using Hough Forests

In general, the task of object detection is posed as a binary classification problem. Object models are learned to distinguish between specific object classes and background. To detect multiple object classes, the usual procedure is to apply a large number of independently trained single-class object detectors, like in the approach of Felzenszwalb et al. [2010b]. However, this approach is computationally expensive and does not scale to thousands of object classes. Later in this chapter it is shown that it is beneficial to integrate object detection results as additional input for concept classifiers. For this purpose, many object classes have to be detected in large image and video databases.

An appealing approach to reduce the computational overhead is the concurrent detection of several object classes by using a multi-class learning framework. Instead of learning class-specific object detectors, the aim is to learn a common classification model for multiple or all object classes.

Currently, class-specific Hough forests [Gall and Lempitsky 2009] have been successfully applied for object detection. In this section, an extension of Hough forests for multi-class object detection is presented. Hough forests are random forests that use local features to vote for object locations in order to realize a generalized Hough transform for a single-class object detection problem. It is shown that the features and the split function used in Gall and Lempitsky's approach are not appropriate for multi-class object detection. To resolve this issue, local HoVW features in conjunction with an appropriate node split function for multi-class random Hough forests are proposed. The presented approach classifies multiple classes without any significant computational overhead, in contrast to other multi-class approaches, such as multi-class SVMs that have to build one-against-one SVM classifiers for all class combinations. Experimental results for the *Caltech-101* test set demonstrate that the presented multi-class approach relying on local HoVW features achieves similar performance as single class-specific Hough forests, even when detecting as many as 24 object classes at a time.

This section is organized as follows: Previous work is discussed in Section 4.4.1. In Section 4.4.2, the construction of multi-class Hough forests is explained. Experimental results are presented in Section 4.4.3. Finally, the results are discussed in Section 4.4.4.

### 4.4.1 Related Work

Random forests introduced by Breiman [2001] consist of an ensemble of decision trees. They inherit the positive characteristics of decision trees, but they do not suffer from the problem of overfitting. Breiman [2001] showed empirically that random forests are more robust to noise in the training data, i.e., mislabeled training examples, than Adaboost. Moreover, the construction of trees has a rather low computational complexity compared to a SVM, the classification is very efficient at runtime, and the training as well as the classification can be easily parallelized at the level of decision trees.

Recently, random forests have been successfully applied to image classification. Bosch et al. [2007] investigated random forests for multi-class image classification using spatial shape and appearance descriptors. Simple linear classifiers on random feature subsets are used as decision functions within the trees. The authors showed that random forests are significantly faster with only a slight performance decrease in comparison to multi-class multiple kernel SVMs.

Recently, Hough-based approaches received considerable attention in the field of object detection. Exploiting the fact that object parts provide useful spatial information, local features are used to vote for object locations. Thus, these approaches are relatively robust to partial occlusions, shape and appearance variations.

Leibe et al. [2008] presented the implicit shape model as a probabilistic formulation of the Hough transform. It consists of a class-specific codebook and a spatial probability distribution. The codebook is learned from local feature descriptors using the K-means clustering algorithm, and the probability distribution specifies where each codebook entry can be found within the object area.

At the detection stage, local descriptors are matched to codebook entries, and probabilistic Hough votes are generated based on the corresponding spatial probability distribution. The Hough votes are aggregated in a voting space where object locations are determined by searching for local maxima. To optimize the detection performance, Maji and Malik [2009] extended the implicit shape model by placing the Hough transform into a discriminative framework. The authors used a max-margin formulation to learn weights on the entries of the appearance codebook. The weights for possible object location votes indicate whether a codebook entry is a good predictor for an object location. Another way of improving the object detection performance is to discriminatively learn the codebook. Gall and Lempitsky [2009] used a random forest framework, called *Hough forest* for realizing a generalized Hough transform to detect object appearances. Kumar and Patras [2010] used a different criterion based on intermediate Hough images for tree construction. They tried to explicitly maximize the response at the true object locations in the Hough images. Therefore, Hough spaces for all training images have to be calculated at all non-leaf nodes during training. Hough forests are also used by Fanelli et al. [2009] for mouth localization in facial images and by Yao et al. [2010] for action recognition in videos.

An approach for multi-class object detection has been presented by Torralba et al. [2007]. Instead of training object detectors individually, the authors used a joint-boosting algorithm to share features among object classes. Using 21 object classes from the *LabelMe dataset*, the authors have shown that jointly learning object classes needs less training data and yields a better object detection performance than using single-class object detectors.

### 4.4.2 Multi-Class Hough Forests

The proposed multi-class object detection approach is based on the class-specific Hough forest presented by Gall and Lempitsky [2009]. Besides the extension of this approach for multi-class object detection, different local feature representations are investigated. In the following, the construction of the underlying random forest, the Hough voting extension, the required leaf node information, as well as the detection of object centers in the Hough images are explained.

The random Hough forest approach as proposed by Gall and Lempitsky [2009] uses decision functions that directly compare pixel values. Therefore, each local patch consists of a number of image channels: three color channels, four edge channels with first- and second-order derivatives and nine HOG-like channels. Apart from these HOG-like features, two further feature representations are investigated. First, densely sampled RGB-SIFT descriptors are used. Second, based on RGB-SIFT descriptors, the usefulness of local HoVW features is analyzed. The assumption is that these descriptors are more suitable to describe local object parts, because they capture the local spatial arrangement of visual words. In a first step, a vocabulary of visual words is generated by clustering the

SIFT descriptors from the set of training images in their feature space. For this purpose, the K-means algorithm is used. Then, image regions are represented as local HoVW features by mapping the keypoint descriptors to the visual words of the vocabulary. The soft-weighting scheme as described in Equation 3.1 is used to diminish the quantization loss during histogram generation.

### Random Forest Construction

A random forest is an ensemble of decision trees. To realize an efficient multi-class object detection system, the decision trees are trained in a multi-class fashion. The training data consist of a set of instances $P_i = (I_i, c_i, d_i)$ with the image patch $I_i$, the class label $c_i$ and the relative position to the object center $d_i$. The training subsets for the different trees are generated using subbagging. The decision trees are built in a top-down manner by selecting at each node the best split function of a set of randomly instantiated split functions, so that the impurity of class labels and class specific offsets in the child nodes are minimized. Thus, to build the trees a binary split function for decision making and an uncertainty measure have to be defined that guarantee the purity of class labels and offsets in the leaf nodes.

Two different split functions have been investigated. The decision function of the original approach [Gall and Lempitsky 2009] directly compares values of a pair of pixels in an image patch $I$ within the same channel $a$:

$$t_{a,p,q,r,s,\tau}(I) = \begin{cases} 0 & \text{if } I^a(p,q) < I^a(r,s) + \tau \\ 1 & \text{otherwise} \end{cases} \tag{4.3}$$

with a decision threshold $\tau$ and two locations $(p, q)$ and $(r, s)$ within the image patch. Randomness is introduced by randomly choosing the channel and pixel positions. For local HoVW features as well as for the edge histograms of SIFT descriptors, the following simple linear classifier is applied:

$$t_{n,b}(x) = \begin{cases} 0 & \text{if } n^T x + b \leq 0 \\ 1 & \text{otherwise} \end{cases} \tag{4.4}$$

where $n$ is a vector of the same size as the feature vectors. Randomness for the linear classifiers is introduced by randomly choosing the components of the vector $n$ in the range of $[-1, 1]$.

Two types of uncertainty measures are used: the class-label uncertainty $U_{label}$ and the offset uncertainty $U_{offset}$. For each node, a set of decision functions $t^k$ with randomly chosen parameters is considered. The following optimization function with $U \in \{U_{label}, U_{offset}\}$ is solved to find the binary test that optimally splits the data:

$$argmin_k(U(\{P_i|t^k(I_i) = 0\}) + U(\{P_i|t^k(I_i) = 1\})). \tag{4.5}$$

In the case of multiple classes the class-label uncertainty is given by

$$U_{label}(A) = |A| \cdot \mathit{Entropy}(A)$$
$$\text{with}$$
$$\mathit{Entropy}(A) = -\sum_{c=0}^{C-1} \frac{|A^c|}{|A|} \log_2 \left( \frac{|A^c|}{|A|} \right) \tag{4.6}$$

and the offset uncertainty by

$$U_{\mathit{offset}}(A) = \sum_{c=1}^{C-1} \sum_{i:c_i=c} ||d_i - d_A^c||^2 \quad \text{with} \quad d_A^c = \frac{1}{|A^c|} \sum_{i:c_i=c} d_i \tag{4.7}$$

where $C$ is the number of classes, $A^c$ is the subset of $A$ that contains all instances of class $c$, and $d_i$ is the offset of the $i$-th local patch. For calculating the offset uncertainty, the background class is not considered. The type of uncertainty is randomly chosen for each node.

The final leaf node information represents the visual codebook and stores the class as well as the spatial information. Therefore, the training data are recursively split until a maximum depth is reached or the number of patches falls below a minimum. Each leaf node consists of a list of offset vectors and corresponding class labels for the containing instances. Furthermore, the class probabilities, i.e., the percentage of the corresponding object class patches, are stored. These probabilities determine the weight of the associated *Hough votes* in the object detection stage.

### Hough Voting

During object detection, the local feature descriptors are propagated through the trees of the random forest according to the split criteria in the nodes. At the leaf nodes, Hough votes for locations of possible object centers are triggered using the stored offset vectors. The votes are weighted by the corresponding class probabilities. Two voting strategies are investigated. The first strategy votes for all classes in the leaf node. Thus, weighted votes are generated for all offset vectors. The second strategy only considers offset vectors from the dominating object class. To detect objects at different sizes, the Hough forest algorithm is applied to a series of images at different scales resulting in several Hough images, one Hough image per object class and scale.

### Local Maxima Detection

Finally, the objects are detected as local maxima in the Hough images. A Hough image contains the accumulated votes. The idea of the Hough transformation

| | | | **Pre-scaled** | |
|---|---|---|---|---|
| [%] | **Single-class** | **Multi-class** | **Single-class** | **Multi-class** |
| Gall and Lempitsky | 56.7 | 14.4 | 70.6 | 39.2 |
| DSIFT | 42.4 | 36.5 | 55.6 | 49.1 |
| HoVW | 57.9 | 54.5 | 67.8 | 65.0 |

Table 4.1: Mean f1-scores for 24 object classes of the *Caltech-101 Challenge*.

is that the triggered votes of local patches yield peaks in the Hough image at the positions of the object centers. These local maxima in the Hough images are detected using the mean-shift algorithm, which is a local, iterative and non-parametric approach. The implementation of the `Intel OpenCV library` [Bradski 2000] is used. The detected local maxima have to exceed a predefined threshold to be accepted as an object center. The corresponding bounding boxes are determined based on the scale of the corresponding Hough image.

### 4.4.3 Experiments

In this section, experimental results are presented for a subset of the *Caltech-101* test set [Fei-Fei et al. 2007]. *Caltech-101* is a challenging dataset containing 101 object classes and a background class. The bounding boxes are provided as ground truth data for all object appearances. For the experiments, the following 24 object classes are randomly selected: "airplane", "bonsai", "brain", "buddha", "butterfly", "car", "chandelier", "ewer", "face", "grand piano", "hawksbill", "helicopter", "kangaroo", "ketch", "laptop", "leopard", "menorah", "motorbike", "revolver", "scorpion", "starfish", "sunflower", "trilobite", and "watch". For each object class and for the background class, 65 randomly chosen images were used for training, and from the remaining images, 15 images per object-class were randomly chosen for testing. The f1-score, as described in Section 2.8, is calculated for the point in the ROC-curve where the difference of recall and precision is minimal.

In a first experiment, the multi-class and single-class object detection performance on the 24-classes subset of the *Caltech-101* dataset is investigated. The experimental results are displayed in Table 4.1. The application of a large number of class-specific object detectors achieved better performances compared to the multi-class approaches. While the accuracy of the multi-class extension for the original approach declined from 56.7% to 14.4%, the approaches based on SIFT descriptors and local HoVW representations showed a significantly smaller performance decrease from 42.4% to 36.5% and from 57.9% to 54.5%, respectively. Overall, the best performance was achieved using HoVW representations.

In a second experiment, the procedure has been repeated with pre-scaled test images. The test images have been scaled such that the objects are of the same size as in the training set. The results for this experiment are presented in Table

| | **Training** | | **Testing (per image)** | |
|---|---|---|---|---|
| | **Single-class** | **Multi-class** | **Single-class** | **Multi-class** |
| DSIFT | 1,473 h | 75 h | 17.8 s | 0.9 s |
| HOW | 1,284 h | 67 h | 74.6 s | 4.1 s |

Table 4.2: Runtimes for 24 object classes of the *Caltech-101 Challenge* on a linux cluster with 2 GHz Dual-Core AMD Opteron$^{\text{TM}}$ CPU.

4.1 either. As expected, the performance increased for all runs. The implementation of Gall and Lempitsky's approach seems to be more sensitive to differing object scales than the proposed approach that relies on local HoVW features. While the performance loss of the original approach comparing multi-class and single-class detection was 42.3% in the preceding experiment, it also declined by 31.4% using equally scaled objects. The performance loss of the proposed approach using HoVW features amounts to only 2.8% when pre-scaled images are used. The experiments suggest that the combination of local HoVW features and linear classifiers as decision functions are more appropriate for multi-class object detection Hough forests. The multi-class object detectors are significantly faster than the single-class detectors (Table 4.2). Moreover, if object detection results are used for semantic concept detection, the overhead for computing HoVW features is negligible since related state-of-the-art systems rely on visual words and thus these features do not need to be calculated twice.

### 4.4.4 Discussion

To detect a large set of object classes in images, it is inefficient to run a large number of single-class object detectors. In this section, a multi-class approach for the task of object detection has been presented. This approach is capable of detecting 24 different object classes at a time, instead of applying one object detector for each object class separately. To achieve this, a random Hough forest approach with appropriate measures for class and offset uncertainty has been extended. The proposed approach relies on local HoVW features with an adequate split function. It turned out that the choice of features is crucial for obtaining a multi-class detection performance that is comparable to the single-class case. While the performance of the multi-class extension of the original approach using HoG-like features clearly dropped, the multi-class Hough forest based on local HoVW features almost retained the performance compared to the class-specific version. Overall, it is shown how to construct multi-class Hough forests in order to speed up the concurrent detection of many object classes in images.

## 4.5 Object-Based Feature Representations

In the following, the results of the previously described object detectors are pooled to a feature vector describing the underlying image or video shot. Even if the kind of object detection approach is exchangeable, the method of Felzenszwalb et al. [2010b] is mainly used within the experiments, because it won several prizes and achieved superior performance at the annual *VOC Challenge*.

The object detectors are applied using low thresholds to also obtain detection results with low confidence scores. The confidence scores of the object detectors based on deformable part models can be transformed to probabilities by normalizing the values to the interval $[0, 1]$ using the following logistic function:

$$prob(x) = \frac{1}{1 + \exp(-Ax)} \quad \text{with} \quad A \in \mathbb{R}^+. \tag{4.8}$$

In the case of "faces" the number of detection hits is normalized to the interval $[0, 1]$ by using a linear function of the form

$$prob(x) = \max(A \cdot x, 1) \quad \text{with} \quad A \in \mathbb{R}^+. \tag{4.9}$$

The following feature representations are inspired by BoVW approaches. Let $K$ be the number of object classes and $D_k$ the number of object detection results in an image $I$ for the object class $k$. The object detection results are given by $d_i^k = (x, y, w, h, s)$ with $k \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, D_k\}$. The parameters $x$, $y$, $w$ and $h$ are the spatial coordinates of the bounding box and $s$ is the confidence score or probability.

Different pooling strategies are applied to assemble object detection results to image representations. The max-pooling, average pooling and mix-order max-pooling strategies, that are known from BoVW approaches, are adapted to pool object detection results.

Using the average-pooling strategy an image is represented as a $K$-dimensional feature vector $[v_1, \ldots, v_K]$ with

$$v_k = \begin{cases} \text{avg}_{i \in 1,\ldots,D_k} s_i^k & \text{if } D_k > 0 \\ 0 & \text{otherwise .} \end{cases} \tag{4.10}$$

In contrast to the average pooling strategy, the feature vector components in the case of the max-pooling strategy are given by the following equation:

$$v_k = \begin{cases} \max_{i \in 1,\ldots,D_k} s_i^k & \text{if } D_k > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

where $w_k$ intuitively measures the probability of object class $k$ being present in the image. The application of the mix-order max-pooling strategy, which is an extension of the max-pooling approach, to object detection results allows us to measure the probability, that an object class is at least $n$-times present in an image. Therefore, the detection results are sorted per object class and the top $N$ values are used as features. The resulting feature vector has the dimension of $K \cdot N$. For $N = 1$ the resulting feature representation corresponds to the max-pooling strategy.

Based on the general object detection approach as described in Section 4.3 a further strategy is introduced that takes the component information of the object models into account. This strategy, called *component-based max-pooling*, relies on the max-pooling strategy, but treating the mixture components as separate object classes. If each mixture consists of $M$ components, the feature vector dimension results in $K \cdot N \cdot M$.

Besides these from BoVW approaches inspired feature representations, the sizes of the detected bounding boxes are used as additional features. The average as well as maximum size of the bounding boxes are computed as follows:

$$z_{avg,k} = \begin{cases} \text{avg}_{i \in 1,\ldots,D_k} w_i^k \cdot h_i^k & \text{if } D_k > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{4.12}$$

$$z_{max,k} = \begin{cases} \max_{i \in 1,\ldots,D_k} w_i^k \cdot h_i^k & \text{if } D_k > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4.13}$$

For videos, the annotation is generally based on a shot segmentation, which is used as a preprocessing step for semantic video indexing. In current video retrieval applications a shot is typically analyzed by processing its keyframe(s). This procedure allows the application of image-based approaches to videos.

To improve the reliability of the object detection results, object models are applied to subsequent video frames of a shot instead of to keyframes only. Two different strategies are pursued to assemble object detections within video shots to object sequences.

The first strategy uses an optical flow algorithm to track object regions between subsequent frames. A feature detector is applied to an object region to find points of interest that are suitable for tracking. For this purpose, the pixels with the highest eigenvalues are selected and tracked using the optical flow computation method according to Bouguet [2001]. This approach is an extension of the Lukas-Kanade approach [Lukas and Kanade 1981], that processes image pyramids to enable the estimation of fast movements as well. Two object regions belong to the same sequence if a predefined ratio of feature points from the preceding frame is tracked successfully and lies within the object region of a following frame. If a matched object region is found the object is added to the corresponding sequence and new interest points are selected from the object region to continue the tracking process.

The second strategy uses an agglomerative single linkage clustering to assemble object detections within a shot to object sequences. This strategy is computationally much faster than the previous approach. It is a hierarchical bottom-up approach where the distance between two clusters $X$ and $Y$ arises from the minimum distance between its elements:

$$d_{SLC}(X, Y) = \min_{x \,\in\, X, y \,\in\, Y} d(x, y). \tag{4.14}$$

For this purpose, the distance between two detected object regions considers position, size, frame number and detection score and is calculated as follows:

$$d(obj_i, obj_j) = overlap(obj_i, obj_j) + \alpha \cdot \|f_i - f_j\| + \beta \cdot \|s_i - s_j\| \tag{4.15}$$

where an object region is defined by $obj_i = (x_i, y_i, w_i, h_i, f_i, s_i)$ with the position $(x_i, y_i)$, the size $(w_i, h_i)$, the frame position $f_i$ and the detection score or probability $s_i$. The weighting factors $\alpha$ and $\beta$ determine the influence of different frame positions and different scores, respectively. The overlap function measures the relative intersection between two bounding boxes as described in Listing 4.1. The clustering process stops if no more clusters can be merged due to overlapping object regions or if a predefined threshold is exceeded. In a post-processing step, object sequences with an insufficient number of detected objects are discarded to minimize false alarms. The minimum number of detected object regions per sequence is set relative to the detection frame rate.

```
overlap($obj_i$,$obj_j$)
do
    $w_{intersect} = \min(x_i + w_i, x_j + w_j) - \max(x_i, x_j)$
    $h_{intersect} = \min(y_i + h_i, y_j + h_j) - \max(y_i, y_j)$

    if ($w_{intersect} > 0$  &&  $h_{intersect} > 0$)
    then
        $w_{union} = \max(x_i + w_i, x_j + w_j) - \min(x_i, x_j)$
        $h_{union} = \max(y_i + h_i, y_j + h_j) - \min(y_i, y_j)$
        return 1 - $w_{intersect} \cdot h_{intersect}$ / $w_{union} \cdot h_{union}$
    else
        return 1
    fi
done
```

Listing 4.1: Relative intersection between two bounding boxes.

Text detection results are treated differently. Under the assumption that overlaid text is constant in position and size, text detection results are assembled to sequences if the overlaid text is detected at approximately the same position and size for several subsequent I-frames, otherwise it is discarded.

After generating the object sequences the following features are extracted. First, a shot-based confidence score or probability for each object class $k$ is derived. If a shot contains $S_k$ object sequences of class $k$ and a sequence consists of $D_{i,k}$ detected object regions with $k = 1, ..., K$ and $i = 1, ..., S_k$, the vector of the confidence scores $[v_1, \ldots, v_K]$ is calculated as follows:

$$v_k = \begin{cases} \max_{i \in 1,...,S_k}(\text{avg}_{j \in 1,...,D_{i,k}} s^k_{i,j}) & \text{if } S_k > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

where K is the number of object classes and $s^k_{i,j}$ is the confidence score or probability of the $j$-th detected region of the $i$-th object sequence for object class $k$. Second, the number of sequences per object class $S_k$ is used as additional shot-based feature.

Furthermore, for each object class the sequence with the highest confidence score is selected and the following features are extracted: average position, average frame coverage and movement. The value for “movement” describes the maximum distance between two object positions of a sequence.

Additional information about the extracted features, the applied machine learning algorithms and the used fusion schemes are given in the description of the corresponding experiments in the following section.

## 4.6 Experiments

In this section, the performance impact of object-based features in generic image and video indexing systems is investigated. For this purpose extensive experiments are conducted on the *Mediamill* and on the *VOC* Challenge as well as at the semantic indexing task of the annual *TRECVid Challenge* [Everingham et al. 2010; Smeaton et al. 2006; Snoek et al. 2006b]. Furthermore, experimental results are presented that investigate the generalization capabilities of concept models in a cross-domain concept detection setting.

The remainder of this section is organized as follows: In Section 4.6.1 the *Mediamill Challenge* with 137 news videos is used for the automated detection of 101 semantic concepts. Object-based features are generated by assembling detection results to object sequences. A shot-based confidence score as well as further features, such as position, frame coverage and movement, are computed for each object class and combined with pre-computed low-level features from the *Mediamill Challenge* in an early fusion scheme. In Section 4.6.2 experiments are performed on the image classification test set of the *VOC Challenge* 2007.

Object-based features are combined with state-of-the-art BoVW features. Different object-based feature representations and fusion strategies including early fusion, MKL and late fusion are investigated. Section 4.6.3 presents the concept detection systems based on the *TRECVid* participations of the years 2009, 2010 and 2011, where object-based features have been incorporated. In the last section, the impact of object-based features in cross-domain settings is investigated. Therefore, concept models are applied to domains other than the training domain. The domains cover documentary films and news videos including English, Arabic, and Chinese channels.

### 4.6.1 Mediamill Challenge

The *Mediamill Challenge* consists of 86 hours of news videos from the *TRECVid* 2005 development set with 43,907 video shots and provides pre-computed low-level features per shot. This challenge has been chosen since ground truth data are available for an extensive set of 101 concepts. The aim is to demonstrate that a few object-based features are beneficial for a large number of concepts.

The focus of the following experiments is laid on the first experimental setting of the *Mediamill Challenge*, which is based on a pure visual analysis of the data. The optical flow algorithm as described in Section 4.5 is used to assemble detection results to object sequences and a shot-based confidence score as well as further features like position, frame coverage and movement are calculated for each object class. The following object classes are used to build object sequences: “airplane”, “car”, “chair”, “motorbike”, “person”, “television screen”, “face” and “text”. The publicly available object models except the special detectors for face and text are based on the development data of the *VOC Challenge* 2008. The training data for the object models consists of images obtained from the Flickr website. The number of training instances per object class varies: 316 airplanes, 840 cars, 63 chairs, 272 motorbikes, 4,168 persons, and 274 television screens. Examples of object detection results on the *TRECVid* data set are shown in Figure 4.2.

The generated object sequences are used to extract the following features. A shot-based confidence score for each object class is calculated based on Equation 4.16. Since the object detectors for face and text do not directly provide confidence scores, the number of detection hits for faces and the accumulated frame coverage of the text elements are used as appropriate score. Furthermore, the number of object sequences per class, the average object position, the average frame coverage and the movement are computed, as described in Section 4.5. Additionally, a feature indicating the presence of a female person is computed. Therefore, a SVM-based female classifier is built upon the detected face sequences of the training set. Since the used ground truth annotations for the concept “female” are coded in binary form, positive training shots are restricted to shots with exactly one face sequence. In the training as well as in the classification

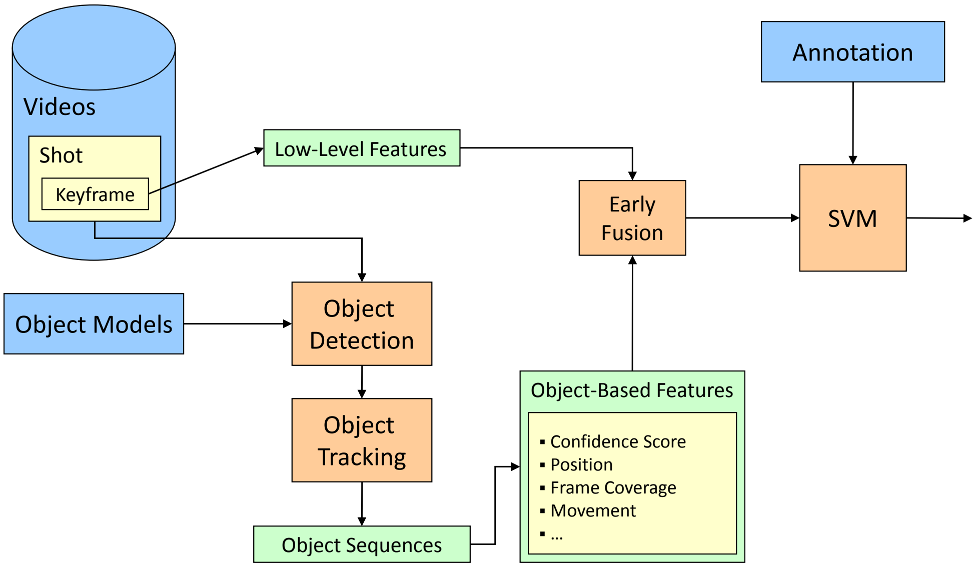


Figure 4.1: Video concept classification system.

stage the face with the highest number of detection hits per sequence is selected. This ensures that the used face images are as frontal and as large as possible. The face images are normalized to 24 x 24 pixels and a principal component analysis is applied to further reduce the dimensionality. The SVM model is learned using the RBF-kernel function. In the classification stage the female model is applied to the whole dataset providing confidence scores indicating the presence of a female person as additional object-based features.

In an early fusion scheme the pre-computed low-level features provided by the *Mediamill Challenge* are combined with the derived object-based features and are fed into a SVM as described in Figure 4.1. Since the focus is on the performance impact of object-based features, it is abstained from heavy parameter tuning and heuristic values are used as SVM parameters of the SVM$^{\text{light}}$ implementation [Joachims 1998]. More tweaking of parameters could achieve slightly higher mean AP values, but also carries the risk of overfitting.

The number of found shots, containing at least one corresponding object sequence, together with shot-based recall and precision values for each object class are presented in Table 4.3. The low detection performance of the airplane detector is understandable with respect to the *Mediamill* annotations that contain many positive examples showing close-ups of persons leaving an airplane. Thus, the airplane detector is incapable of recognizing an airplane only by means of such a small segment of the object.

To examine the impact of each object class, a series of eight experiments adding confidence scores for each separate object class to the low-level features was performed. Then, the number of concepts with clearly improved performance

| Object | #found shots | Recall | Precision | #improved concepts |
|---|---|---|---|---|
| Airplane | 6 | 3.3% | 66.7% | 5 |
| Car | 382 | 39.1% | 84.6% | 21 |
| Chair | 28 | 14.4% | 46.4% | 9 |
| Face | 6,190 | 73.9% | 96.2% | 34 |
| Motorbike | 7 | 27.3% | 85.7% | 3 |
| Overlaid text | 6,537 | 88.9% | 60.2% | 19 |
| Person | 9,727 | 87.4% | 88.1% | 18 |
| TV screen | 322 | 44.0% | 49.1% | 9 |

Table 4.3: Performance comparison of object detection results on the *Mediamill Challenge*.

| [%] | 101 concepts | 60 related concepts |
|---|---|---|
| Baseline | 29.3 | 32.0 |
| Experiment 1 | 31.3 | 35.4 |
| Experiment 2 | 31.1 | 36.6 |
| Experiment 3 | 31.8 | 37.2 |

Table 4.4: Overall concept detection performance in mean AP.

was counted as presented in Table 4.3. A concept is considered to be improved if the absolute performance improvement exceeds at least 1%. The experimental results show that the more frequently an object appears in the video data, the higher is the number of improved concepts.

Three experiments were performed to examine the overall performance. Experiment 1 extends the baseline system using an eight-dimensional context vector of shot-based confidence scores for each object class. Experiment 2 additionally includes all features that were derived from the object sequences. Due to the supposed importance of the derived features especially for the object classes "face" and "text", a third experiment was conducted where all object-based features for the object classes "face" and "text" are included as well as a six-dimensional context vector of confidence scores for the remaining objects. The last two experiments both include the feature obtained from the female classifier.

The experiments are evaluated on the entire set of concepts as well as on a subset of 60 concepts which are related to one of the detected object classes. It can be expected that especially these 60 concepts profit from the object-based features. Relations between concepts and objects were rated by three independent persons unaware of the experimental results. The final set of 60 related concepts was determined in a common discussion on the disagreements.

The overall results are presented in Table 4.4. The following observations can be made: All experiments relying on object-based features are significantly better than the baseline system at a significance level of 1%. Further tests at the same significance level on the set of non-related concepts (not shown in Table 4.4) have also shown to be significantly better.

Experiment 1 using shot-based confidence scores for the object classes obtains a relative performance improvement of 8.22% and 6.8% mean AP in comparison

| **AP [%]** | **Baseline** | **Exp. 1** | **Exp. 2** | **Exp. 3** | **Object detector** |
|---|---|---|---|---|---|
| Airplane | 14.9 | 21.5 | 15.7 | 15.8 | 77.1 |
| Car | 39.4 | 79.4 | 77.9 | 78.6 | 81.5 |
| Chair | 44.9 | 49.2 | 51.9 | 53.6 | 48.2 |
| Face | 96.0 | 99.6 | 98.3 | 99.0 | 95.8 |
| Motorbike | 1.3 | 12.0 | 4.6 | 9.7 | 91.5 |
| Overlaid text | 79.8 | 82.9 | 87.1 | 87.0 | 58.2 |
| Person | 97.6 | 99.9 | 98.6 | 98.9 | 88.0 |
| TV screen | 23.1 | 52.6 | 56.6 | 51.8 | 49.6 |

Table 4.5: Performance comparison of concept detection and object retrieval results for directly related concepts.

to the baseline system concerning related and all concepts, respectively. Not only concepts that directly correspond to the object classes, but also indirectly related concepts profited from the additional features. The concept “road”, for example, improved from 27.23% to 44.99% AP, “vehicle” from 36.08% to 62.7% AP, “walking or running” from 49.48% to 54.56% AP, “office” from 9.83% to 12.16% AP, or “indoor” from 80.4% to 85.31% AP. In total, 44 concepts were clearly improved. While Experiment 2 outperformed Experiment 1 on the related concept set, no performance gain could be achieved on the whole concept set. Nevertheless, the inclusion of features, like position, frame coverage and movement, leads to the best results for several concepts, like “anchor”, “crowd”, “people marching”, “cycling”, “graphics”, “golf”, “indoor”, “splitscreen”, “maps”, “soccer”, or “weather”. The best overall performance on both concept sets was obtained in Experiment 3 with 37.19% and 31.83% mean AP, respectively. Concerning the set of related concepts, Experiment 3 was even significantly better than Experiment 1 at a significance level of 5%. The concept “female” improved from about 26% AP in Experiment 1 to 55% in Experiment 2 and even 60% AP in Experiment 3. This improvement is probably due to the integration of confidence scores from the female classifier.

Furthermore, the direct use of shot-based object detection confidence scores for concepts directly corresponding to the object classes was evaluated (Table 4.5). When comparing direct object retrieval results and concept detection results in terms of AP, it should be kept in mind that the number of returned shots can be very different. While the number of found shots in the case of direct retrieval ranges from 6 shots for the object “airplane” to 9,727 shots for the object “person”, concept detection results were always limited to exactly 2,000 shots. For rarely occurring object classes it seems beneficial to immediately return retrieval results based on the confidence scores of the object detectors. In this case the related concept detector relies only on a small set of training samples and the higher amount of training samples for the object detectors becomes apparent.

Overall, it can be concluded that object-based features improve the performance not only for the concepts that directly correspond to the object detectors,

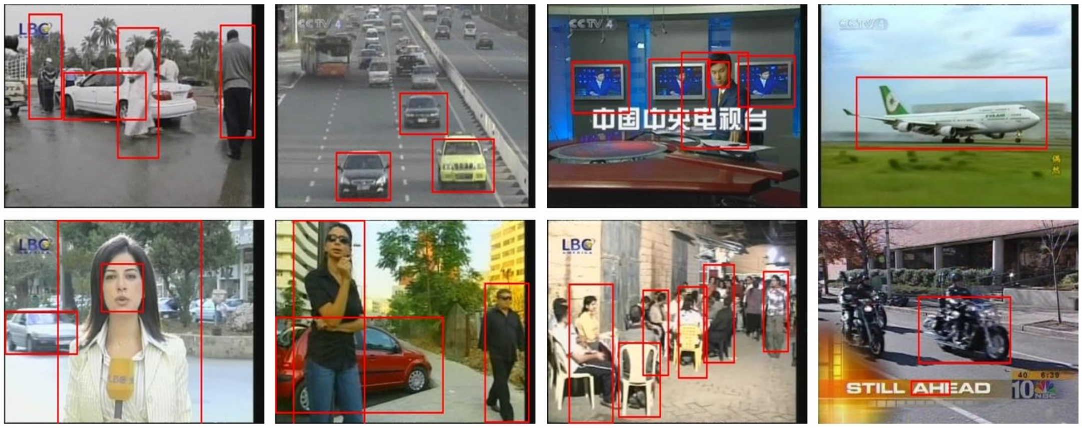


Figure 4.2: Visualization of object detection results from the *Mediamill Challenge*.

but also for a large number of indirectly related concepts. All experiments exploiting object-based features are significantly better than the baseline system on the subset of related concepts as well as on the whole concept set. The concept detection system relying on object-based features achieves a mean AP of 37.2% for the 60 related concepts, which corresponds to a relative performance improvement of 16.3% compared to the baseline system.

Furthermore, a severe problem in the field of concept detection is the inclusion of additional training data from foreign domains. Object detectors seem to be less domain-specific and thus overcome these limitations. The experiments show that the use of object detectors allows embedding object specific training data from foreign domains, like the Flickr website.

### 4.6.2 VOC Challenge

In this section, experiments are conducted on the image classification task of the *VOC Challenge* 2007 [Everingham et al. 2010]. Different object-based feature representations as described in Section 4.5 are evaluated on this test set. Furthermore, object-based features are combined with state-of-the-art BoVW features and different fusion strategies are investigated. Using the approach of Felzenszwalb et al. [2010b] object detectors are applied for the 20 object classes listed in Table 4.6. The experimental results are evaluated using the AP measure.

In a first experiment, different feature representation are evaluated. The used BoVW features are based on RGB-SIFT descriptors, a vocabulary of 5,000 visual words, a dense sampling strategy with a step size of 5 pixels and a spatial pyramid representation with 1x1 and 2x2 spatial subregions. The object-based feature representations are derived from the detection results using three different pooling strategies: max-pooling, mix-order max-pooling and component-based

| | | **Object-based features** | | |
|---|---|---|---|---|
| **AP** [%] | **Densly sampled RGB-SIFT with spatial pyramids** | **Max-pooling** | **Mix-order max-pooling** | **Component-based max-pooling** |
| Aeroplane | **70.2** | 59.5 | 60.4 | 57.7 |
| Bicycle | 55.7 | **71.1** | 70.4 | 68.7 |
| Bird | **40.8** | 20.1 | 17.9 | 15.0 |
| Boat | **63.6** | 46.4 | 41.3 | 29.4 |
| Bottle | 20.6 | **48.4** | 47.5 | 47.2 |
| Bus | 59.6 | 66.1 | 69.2 | **70.2** |
| Car | 72.7 | **86.6** | **86.6** | 85.9 |
| Cat | **53.5** | 45.6 | 46.7 | 38.1 |
| Chair | 50.2 | 51.7 | **52.5** | 51.0 |
| Cow | 32.7 | **57.4** | 48.1 | 34.0 |
| Dining table | **47.5** | 34.1 | 36.0 | 40.2 |
| Dog | **41.7** | 23.3 | 25.0 | 18.5 |
| Horse | **76.2** | 62.4 | 65.8 | 55.0 |
| Motorbike | 58.9 | **63.1** | 62.8 | 62.8 |
| Person | 82.6 | 88.8 | **89.1** | 88.5 |
| Potted plant | 23.0 | **31.8** | 29.9 | 15.9 |
| Sheep | 38.6 | 41.3 | **43.2** | 29.5 |
| Sofa | **48.1** | 44.6 | 39.7 | 30.6 |
| Train | **72.2** | 70.3 | 70.4 | 61.6 |
| TV monitor | 48.4 | **62.8** | 17.4 | 59.9 |
| Mean | 52.9 | **53.8** | 51.0 | 48.0 |

Table 4.6: Performance comparison of different object-based feature representations. The best performing feature representation per concept is indicated in bold.

max-pooling. The concept models are built using the SVM$^{\text{light}}$ implementation of Joachims [1998] with a $\chi^2$-kernel. The results for the different feature representations are presented in Table 4.6. Overall, the best result on the VOC Challenge was achieved by the max-pooling strategy. The max-pooling strategy clearly outperformed the other object-based feature representations, like the mix-order max-pooling strategy that encodes the probabilities of at least one, two or three objects of a class being present in the image. Interestingly, the mix-order max-pooling strategy achieved the highest performance in terms of AP for object classes frequently occurring several times in an image like “car”, “chair”, “person”, or “sheep”. The differences between the AP values of the results for object-based features and BoVW features are very large for several concepts, for example “bird”, “boat”, “bottle”, “cow”, “dining table”, or “dog”. This complementary information is used in a second experiment to build a system that combines both feature modalities.

In the second experiment, different fusion strategies are investigated to combine object-based features and BoVW features. The object-based features are built using the max-pooling strategy. The following fusion strategies are inves-

| **AP [%]** | **Early Fusion** | **Late Fusion Product** | **Late Fusion Average** | **Late Fusion Max** | **MKL** |
|---|---|---|---|---|---|
| Aeroplane | 76.7 | 73.0 | 72.7 | 72.9 | **79.0** |
| Bicycle | 75.7 | **76.5** | 75.7 | 74.1 | 76.2 |
| Bird | 37.4 | 30.9 | 34.0 | 23.2 | **41.1** |
| Boat | 60.5 | 58.4 | 59.9 | 59.6 | **64.5** |
| Bottle | **54.4** | 53.3 | 51.7 | 49.6 | 54.1 |
| Bus | 74.9 | 75.1 | 75.0 | 74.5 | **75.2** |
| Car | 88.4 | 88.2 | 87.7 | 87.5 | **88.7** |
| Cat | 55.9 | 54.3 | 54.3 | 53.9 | **57.9** |
| Chair | 57.8 | 58.1 | 58.1 | 56.2 | **58.7** |
| Cow | 53.2 | 53.8 | 55.5 | 53.1 | **56.1** |
| Dining table | 51.4 | 45.2 | 50.8 | 44.1 | **53.0** |
| Dog | 37.2 | 35.8 | 39.8 | 25.4 | **39.9** |
| Horse | 79.5 | 74.5 | 74.2 | 80.7 | **82.0** |
| Motorbike | 68.0 | 69.8 | **69.9** | 68.6 | 69.5 |
| Person | 90.3 | 90.5 | 90.1 | 90.0 | **90.8** |
| Potted plant | 33.3 | 33.4 | **33.5** | 30.3 | 33.0 |
| Sheep | 47.0 | 47.7 | 47.7 | 47.6 | **51.3** |
| Sofa | 52.3 | 53.2 | 55.2 | 43.2 | **55.9** |
| Train | 79.9 | 77.4 | 78.5 | 75.5 | **81.1** |
| TV monitor | 67.6 | 63.2 | 57.9 | 52.9 | **68.8** |
| Mean | 62.1 | 60.6 | 61.1 | 58.1 | **63.8** |

Table 4.7: Performance comparison of different fusion strategies. The best performing feature representation per concept is indicated in bold.

tigated: early fusion, late fusion and MKL. In contrast to early fusion, where the features are concatenated into a multi-modal representation, the late fusion approaches build appropriate models for each feature type. Three different rules are evaluated to combine the individual classification results: the product, the average and the maximum rule. In the MKL setting, as described in Section 2.5.4, the different feature types are fused using a weighted combination of kernel functions. The kernel weights are regularized during the training procedure using the $L_2$-norm.

The results for the different fusion strategies are presented in Table 4.7. The MKL approach performed significantly better than the other fusion strategies at a significance level of 1%. For almost all concepts the MKL fusion scheme achieved the best performance. The combination of BoVW and object-based features using the MKL fusion scheme yielded a superior performance of 63.8% mean AP. The best system at the *VOC Challenge* 2007 achieved 59.4% mean AP with an extensive set of feature types. Chatfield et al. [2011] yielded 61.69% mean AP using Fisher encoding with a very high dimensional feature vector of 327,680 dimensions.

Altogether, the experimental results on the *VOC Challenge* showed that the max-pooling approach is the most advantageous strategy for the image representation of object-based features. State-of-the-art BoVW features and object-based features seem to be very complementary. The combination of both representations led to a relative performance improvement of 18.9% in comparison to the BoVW approach. The MKL framework outperformed all other fusion strategies and led to a superior performace for the image classification task.

### 4.6.3 TRECVid Challenge

Five years of *TRECVid* experience have been gathered from participations at the semantic indexing task, also known as high-level feature extraction task [Mühling et al. 2007b, 2008, 2009b, 2010, 2011a]. The current section presents the experimental results of participations at the annual *TRECVid Challenge* where object-based features have been incorporated, i.e., results of the years 2009 to 2011. The corresponding concept detection systems are described in the following sections. Per concept class ranked lists of at most 2,000 shots have been submitted for evaluation. The concept detection results are evaluated by the *TRECVid* team [Smeaton et al. 2006] based on the infAP measure, as suggested by Aslam et al. [2006]. Additionally, the official partial randomization test [Smeaton et al. 2006] is used to determine whether a competitive system is significantly better than the reference system. The annotation of the development data has been accomplished in a collaborative effort of the participating teams.

#### TRECVid Challenge 2009

The video data of the *TRECVid Challenge* 2009 consists of news magazines, science news, news reports, documentaries, educational programmings, and archival videos. About 100 hours of video data are used as development data and about 280 hours as test data. This corresponds to 36,262 training shots and 95,110 test shots. Altogether, 20 concepts have been evaluated as listed in Table 4.8.

Our baseline video retrieval system relied on low-level features as well as on actual approaches for camera motion estimation, audio analysis, face and text detection. The low-level features are extracted from the keyframe, which is the frame in the middle of a shot. Black bars at the top and bottom of the frames are automatically detected and removed in a preprocessing step. The features of the baseline system, including visual and audio features, are briefly described in the following paragraphs:

*Color moments*: Color moments are extracted at two different granularities. The first three global color moments are computed for the whole image. Corresponding values are extracted for each region of a 3x3 grid in the HSV (hue, saturation, value) color space. The $i$-th pixel of the $j$-th color channel of an image region is represented by $c_{ij}$. The first three color moments are defined as:

$$mean_j = \frac{1}{N} \cdot \sum_{i=0}^{N-1} c_{ij} \tag{4.17}$$

$$stdev_j = \sqrt{\frac{1}{N} \cdot \sum_{i=0}^{N-1} (c_{ij} - mean_j)^2} \tag{4.18}$$

$$skew_j = \sqrt[3]{\frac{1}{N} \cdot \sum_{i=0}^{N-1} (c_{ij} - mean_j)^3} \tag{4.19}$$

*Color autocorrelograms*: Color correlograms describe the spatial relationship between colors, whereas autocorrelograms are limited to identical colors. An autocorrelogram expresses the probabilities of colors to reoccur in a certain distance. Small distances (1, 4, 7, 10, 13, 16 and 19 pixels) are used, so that local spatial correlations of identical colors are represented by the correlogram. Colors are described in the HSV color space. By choosing a smaller number of bins for the brightness component the features become more independent of illumination changes. In total, each color correlogram results in a 350-dimensional feature vector.

*Texture features*: The gray-scale image co-occurrence matrices $m_k$ are constructed at eight orientations. These matrices are used to extract the following values representing the global texture:

$$energy_k = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (m_{kij})^2 \tag{4.20}$$

$$contrast_k = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i-j)^2 \cdot m_{kij} \tag{4.21}$$

$$entropy_k = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} m_{kij} \cdot \log m_{kij} \tag{4.22}$$

$$homogeneity_k = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{m_{kij}}{1 + |i-j|}, \tag{4.23}$$

where $N$ is the number of gray values and $m_{kij}$ is the value of the co-occurrence matrix $m_k$ at position $(i, j)$.

*Gabor wavelet features*: The wavelet coefficients are computed by the following Gabor wavelet functions [Kruizinga and Petkov 1999]:

$$g_{\theta,\lambda,\varphi,\sigma,\gamma}(x,y) = e^{\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}} \cos(2\pi\frac{x'}{\lambda} + \varphi) \tag{4.24}$$

$$\text{with} \quad x' = x\cos\theta + y\cos\theta, \quad y' = -x\sin\theta + y\cos\theta.$$

A Gabor wavelet is controlled by five parameters: the orientation $\theta$, the wave length $\lambda$, the phase $\varphi$, the radius $\sigma$ of the Gaussian function, and the aspect ratio $\gamma$. The radius of the Gaussian function is chosen proportionally to the wave length, and the aspect ratio is fixed to 1. Gabor energies of a pixel for the different orientation and spatial-frequency combinations are obtained by a superposition of the phases 0 and $\pi/2$ using the $L_2$-norm. Gabor wavelet features are extracted for eight orientations and five frequencies. The resulting 40 Gabor energies per pixel are summarized in a Gabor histogram describing the whole image. By distinguishing ten energy classes, the size of the histogram results in 400 dimensions. Furthermore, the average result of each Gabor energy filter is computed for each region of a 4x4 grid. Thus, the total number of Gabor wavelet features amounts to 1,040 values.

*Camera motion features*: Motion vectors embedded in MPEG videos are employed to estimate camera motion at the granularity of P-frames, according to the approach presented by Ewerth and Freisleben [2004]. The following camera motion types are distinguished: translation along the x-axis, respectively y-axis, rotation around the x-axis, respectively y-axis and z-axis, and zoom. A shot is described by the distribution of the values concerning the different camera motion types using the following statistical values: mean, median, minimum, maximum, standard deviation, and skewness. In addition, the percentages of a shot concerning the different camera motion types (pan, tilt and zoom) are considered, so that the final camera motion vector results in 39 dimensions.

*Audio features*: Two sets of audio features are extracted: low-level and mid-level features. The low-level features are extracted from non-overlapping 25 ms frames. These features comprise 20 MFCCs with their first order derivatives, 10 line spectral pairs and a measure of pitch. These values are summarized per shot in a 510-dimensional histogram. The mid-level features are based on a set of low-level features, which is specifically tailored to facilitate audio type classification [Lu et al. 2003]: 8th-order MFCCs, zero-crossing rate, short time energy, sub-band energy distribution, brightness, bandwidth, spectrum flux, band periodicity, a measure of frame noisiness, and the position of the cepstral peak. These extracted low-level features are fed into a content-based audio classification and segmentation system [Stadelmann 2010]. The audio type classification system produces mid-level features on a per second basis in the form of acoustic class labels and related probabilities for “silence”, “speech”, “pure speech”, “non-pure speech”, “music”, “background”, and “action” sounds. The low-level features are therefore aggregated per second, normalized and then concatenated. The resulting feature vector is processed by a hierarchical tree of SVMs, if it was not previously classified as silence by a threshold based classifier. The classification tree is trained on more than 32 hours of audio samples including, among others, the *TIMIT* dataset for clean speech [Garofolo et al. 1993] and the noisy speech corpus *NOIZEUS* [Hu and Loizou 2006]. Finalizing the classifier’s decision, short silence periods within speech are labeled as “pause” by a heuristic decision function. A second algorithm based on the work presented by Ahmadi and Spanias [1999] processes the low-level features “energy”, “zero-crossing rate”, and “cepstral peak” to add “voiced” and “unvoiced” speech labels. Altogether, the eleven acoustic class labels and their probabilities are statistically summarized per video shot using mean, median, minimum, maximum, standard deviation, and skewness values. Furthermore, the percentage of each audio type label with respect to the shot length is calculated. Finally, these percentages and the distribution properties form a 77-dimensional audio mid-level feature vector additionally to the audio low-level features.

*Face and text features*: Face and text sequences are extracted as described in Section 4.5. For each shot the number of face sequences, the number of detected faces, the average shot size and the average number of detection hits are considered as features. Additionally, the percentage of detected profile faces and the ratio of sequence length versus shot length are calculated. For text sequences the following features are derived: number of appearing text elements, average text position and average, maximum, and accumulated text frame coverage.

Based on the observation that the use of face detection results in the baseline system improved the performance of several face related concepts, further object detectors are used. Additionally to the face and text detection approaches, object detectors for the following object classes are incorporated: “airplane”, “bicycle”, “boat”, “bus”, “car”, “chair”, and “person”. The object models were built based on the development data of the *VOC Challenge* 2008. Due to the lack of time and the huge amount of video data, agglomerative single linkage clustering was used to assemble object detections to object sequences as described in Section 4.5. This strategy is computationally much faster than the optical flow-based tracking approach. The generated object sequences are used to extract object-based features as additional input for the SVM-based concept classifiers. The multi-modal low-level, mid-level, and object-based feature representations are combined in an early fusion scheme and are fed directly into a SVM with a RBF-kernel as depicted in Figure 4.1. The implementation of the SVM algorithm provided by the `LIBSVM` library [Chang and Lin 2012] is used.

| **infAP** [%] | **Baseline** | **Run 1** | **Run 2** |
|---|---|---|---|
| Classroom | 2.8 | 1.9 | 3.6 |
| Chair | 4.1 | 5.8 | 7.0 |
| Infant | 0.2 | 0.1 | 0.1 |
| Traffic intersection | 9.1 | 10.3 | 10.7 |
| Doorway | 8.3 | 8.9 | 9.0 |
| Airplane flying | 4.6 | 7.0 | 9.1 |
| Person playing a musical instrument | 8.4 | 9.5 | 9.4 |
| Bus | 0.5 | 2.3 | 3.8 |
| Person playing soccer | 35.8 | 36.4 | 32.0 |
| Cityscape | 10.7 | 12.4 | 12.4 |
| Person riding a bicycle | 2.6 | 5.1 | 6.7 |
| Telephone | 2.2 | 2.3 | 3.1 |
| Person eating | 21.3 | 21.3 | 21.2 |
| Demonstration or protest | 0.9 | 0.9 | 1.3 |
| Hand | 6.1 | 6.2 | 7.0 |
| People dancing | 1.9 | 4.2 | 3.6 |
| Nighttime | 10.5 | 11.1 | 11.5 |
| Boat or ship | 7.8 | 13.5 | 17.2 |
| Female human face closeup | 10.9 | 11.1 | 13.0 |
| Singing | 7.4 | 7.3 | 8.9 |
| Mean | 7.8 | 8.9 | 9.5 |

Table 4.8: Performance evaluation of the *TRECVid* 2009 runs.

In a first run, shot-based confidence scores are calculated per object class as described in Equation 4.16. These confidence scores are joined with the features of the baseline system. For the second run, further object class-specific features, as described in Section 4.5, are derived from the sequences: the number of sequences, the position, the frame coverage and the movement. The results of the three submitted runs are presented in Table 4.8 in terms of infAP. By adding shot-based confidence scores (Run 1) a relative performance improvement of 12.1% mean infAP compared to the baseline system is obtained. Further improvements are achieved by the supplementation of the feature set with additional features like position, frame coverage and movement. This experiment (Run 2) attained a mean infAP of 9.53% for the 20 concept classes, which is a relative performance improvement of 22.2% compared to the baseline system.

Altogether, the experiments revealed that the approaches exploiting object-based features improved the concept detection results significantly. Almost all concepts and not only concepts that directly correspond to the object classes profited from the additional use of object-based features.

### TRECVid Challenge 2010

The *TRECVid Challenge* 2010 used a video database with approximately 11,200 videos from the *Internet Archive* with Creative Commons license. This video database is divided into 200 hours of development data and 200 hours of test data. While the development data consists of approximately 3,200 videos with durations between 3.6 and 4.1 minutes and 118,581 shots, the test data comprises 8,000 Internet Archive videos with durations between 10 seconds and 3.5 minutes and 144,988 shots. Altogether, 130 concepts have been selected for the semantic indexing task and 30 of them have been evaluated by NIST using the infAP measure. The set of 30 concepts that have been evaluated is listed in Table 4.9.

Our baseline system at the *TRECVid Challenge* 2010 relies on state-of-the-art BoVW features using RGB-SIFT descriptors. A dense sampling strategy with a step size of 5 pixels is used. The geometry of the RGB-SIFT descriptors is characterized by eight orientation bins and 4x4 spatial bins. For the spatial bin size 4, 6, and 8 pixels are used. Ten positively labeled training shots per concept are used to construct a 1,000-dimensional vocabulary using the K-means algorithm. Furthermore, the soft-weighting scheme as described in Section 3.2.2 and a spatial pyramid representation with a spatial image partitioning of 1x1 and 2x2 subregions is applied. This parameter setting results in 5,000-dimensional feature vectors. At the training and classification stage, the SVM$^{\text{light}}$ implementation of Joachims [1998] is used.

Based on the success of object-based features in the last year's concept detection system, further object detectors have been incorporated. State-of-the-art object detection approaches [Felzenszwalb et al. 2010b; Viola and Jones 2004] are utilized to find object appearances for the 20 object classes of the *VOC Challenge*, as depicted in Figure 2.11, plus the object class "face". Due to the large amount of video data (263,569 shots) it is abstained from building object sequences and the object-based features are directly extracted from the keyframe images. The object detector of Viola and Jones [2004] is used for faces and the approach based on deformable part models [Felzenszwalb et al. 2010b] is used for the remaining object classes. Using these object detectors trained on separate data sets, shot-based confidence scores as well as further face-related features are computed. Each object detector delivers a number of bounding boxes and associated confidence scores per shot. Based on these object detection results, shot-based average and maximum confidence scores as described in Section 4.5 are calculated for each object class. For the object class "face" the number of faces as well as the average and maximum size of the detected bounding boxes are additionally extracted. The different feature representations are combined using MKL. Besides $L_1$-norm MKL, non-sparse MKL using the $L_2$-norm is also investigated, which leads to a more uniform distribution of kernel weights.

In addition to the baseline system, three runs have been conducted that additionally use object-based features. Table 4.9 shows the submitted results of the baseline system and the three runs in terms of infAP.

In the first experiment (Run 1), object-based features are combined with state-of-the-art BoVW features using the MKL framework. The object-based feature representations include the average confidence scores, maximum confidence scores as well as the face related features. While the BoVW features are compared using

| **infAP [%]** | **Baseline** | **Run 1** | **Run 2** | **Run 3** |
|---|---|---|---|---|
| Airplane flying | 2.9 | 3.0 | 2.1 | 3.0 |
| Animal | 3.0 | 4.7 | 6.7 | 3.6 |
| Asian people | 0.2 | 0.3 | 0.3 | 0.7 |
| Bicycling | 2.5 | 3.3 | 7.7 | 4.6 |
| Boat or ship | 2.9 | 3.1 | 4.3 | 4.7 |
| Bus | 0.1 | 1.2 | 1.3 | 1.4 |
| Car racing | 1.3 | 1.4 | 0.6 | 1.1 |
| Cheering | 2.4 | 2.4 | 2.6 | 3.6 |
| Cityscape | 14.5 | 14.9 | 15.4 | 13.5 |
| Classroom | 0.6 | 0.6 | 0.5 | 0.6 |
| Dancing | 4.7 | 5.0 | 4.9 | 3.4 |
| Dark skinned people | 5.2 | 4.9 | 5.6 | 4.8 |
| Demonstration or protest | 6.1 | 6.2 | 7.4 | 8.4 |
| Doorway | 7.8 | 7.6 | 8.0 | 9.0 |
| Explosion or fire | 2.0 | 2.0 | 1.9 | 1.8 |
| Female human face closeup | 13.2 | 11.8 | 12.4 | 11.7 |
| Flowers | 2.3 | 2.3 | 2.8 | 4.8 |
| Ground vehicle | 9.8 | 18.2 | 20.2 | 20.2 |
| Hand | 3.4 | 3.8 | 5.7 | 6.1 |
| Mountain | 18.5 | 18.5 | 18.6 | 19.4 |
| Nighttime | 5.7 | 5.8 | 6.7 | 8.3 |
| Old people | 5.4 | 5.5 | 4.7 | 3.6 |
| Running | 2.1 | 1.8 | 1.7 | 2.0 |
| Singing | 3.6 | 4.2 | 5.9 | 5.3 |
| Sitting down | 0.1 | 0.2 | 0.2 | 0.4 |
| Swimming | 30.9 | 31.3 | 31.6 | 29.7 |
| Telephones | 0.9 | 0.8 | 0.4 | 1.2 |
| Throwing | 0.2 | 0.2 | 0.9 | 0.3 |
| Vehicle | 9.2 | 17.0 | 20.1 | 19.8 |
| Walking | 5.9 | 6.8 | 7.6 | 7.9 |
| Mean | 5.6 | 6.3 | 7.0 | 6.8 |

Table 4.9: Performance evaluation of the *TRECVid* 2010 runs.

the $\chi^2$-kernel, the object-based feature representations are taken into account by using RBF-kernels. The kernel weights are constrained using the $L_1$-norm. The additional consideration of object-based features significantly improved the baseline system from 5.58% to 6.29% mean infAP.

In a second experiment (Run 2), the impact of $L_2$-norm MKL was investigated, which results in a more uniform distribution of kernel weights. This run further improved the performance and achieved 6.96% mean infAP. In particular, the concepts “animal”, “bicycling”, “bus”, “vehicle” and “ground vehicle” profited from the additional use of object-based features and were partly increased by more than 100% (Table 4.9). In comparison to other participating teams the object-based system achieved the best result for the concepts “vehicle” with 20.1% infAP and “ground vehicle” with 20.2% infAP. Only one team submitted better results for the concept “cheering” and only two teams for the concepts “bicycling” and “animal”.

In the last experiment (Run 3), the feature set was supplemented with global features. Color and Gabor histogram representation were additionally taken into account. The feature vectors of these histograms were compared using the $\chi^2$-kernel. The kernel weights were again learned using $L_2$-norm MKL. This combination of local, global and object-based features achieved no performance gain compared to the previous system. While the concepts "flowers", "cheering", "nighttime", "demonstration or protest" and "doorway" were improved by additionally using global features, several other concepts like "animal" or "bicycling" dropped. It seems that especially concepts describing scenes profited from global color and texture information.

Overall, the experiments revealed that the approaches additionally employing object-based features significantly improved the overall performance. Some concepts like "animal", "bicycling" or "vehicle" were improved by more than 100% in terms of infAP. The concepts "vehicle" and "ground vehicle" yielded with 20.1% and 20.2% infAP respectively, the best results of all participating teams. Furthermore, it has been shown that a more uniform distribution of kernel weights achieved better results than using $L_1$-norm MKL. Finally, the combination of BoVW and object-based features using $L_2$-norm MKL achieved the best overall result and obtained a mean infAP of 6.96%.

### TRECVid Challenge 2011

At the *TRECVid Challenge* 2011 the experiments of the semantic indexing task were conducted on a large set of 346 concept classes. 50 out of the 346 semantic concepts have been evaluated by the *TRECVid* team [Smeaton et al. 2006] based on the infAP measure suggested by Aslam et al. [2006]. Table 4.10 shows the results of the four submitted runs in terms of mean infAP. The development set together with the test set of the last year's challenge form the new development set at the *TRECVid Challenge* 2011. Thus, the new development set contains 11,200 Internet Archive videos (400 hours of video data) with a duration between 10 seconds and 4.1 minutes and 263,569 shots. The new test data comprises again approximately 8,000 videos from the Internet Archive (200 hours of video data) with a duration between 10 seconds and 3.5 minutes and 129,111 shots.

The baseline run relies on BoVW feature representations using densely sampled RGB-SIFT descriptors with a step size of five pixels, a 2,000-dimensional visual vocabulary and a spatial pyramid representation of up to 2x2 regions. The BoVW feature representations are generated using the soft-weighting scheme as described in Section 3.2.2. These parameter settings result in 10,000-dimensional feature vectors. The concept models are learned using the SVM$^{\text{light}}$ implementation of Joachims [1998].

In addition to the baseline system three runs have been submitted that additionally use object-based features. Therefore, object detectors were utilized to find appearances for the same 21 object classes as at the *TRECVid Challenge*

| **infAP [%]** | **Baseline** | **Run 1** | **Run 2** | **Run 3** |
|---|---:|---:|---:|---:|
| Adult | 7.3 | 7.5 | 7.4 | 7.5 |
| Anchorperson | 38.0 | 43.9 | 43.6 | 44.5 |
| Beach | 13.8 | 15.0 | 15.0 | 15.2 |
| Car | 13.5 | 21.5 | 21.5 | 21.3 |
| Charts | 2.9 | 3.4 | 3.4 | 3.4 |
| Cheering | 8.1 | 8.3 | 8.3 | 9.3 |
| Dancing | 3.2 | 2.7 | 2.7 | 2.3 |
| Demonstration or protest | 7.0 | 9.3 | 9.3 | 8.0 |
| Doorway | 5.2 | 5.8 | 5.8 | 5.8 |
| Explosion or fire | 1.7 | 2.9 | 2.9 | 2.9 |
| Face | 11.5 | 11.0 | 11.1 | 12.0 |
| Female person | 7.3 | 9.2 | 9.8 | 10.4 |
| Female human face closeup | 20.5 | 20.2 | 20.0 | 20.1 |
| Flowers | 1.8 | 1.9 | 1.9 | 2.1 |
| Hand | 3.2 | 2.4 | 2.4 | 2.4 |
| Indoor | 3.8 | 4.9 | 4.9 | 4.6 |
| Male person | 6.8 | 6.6 | 7.3 | 8.2 |
| Mountain | 26.3 | 28.1 | 28.1 | 28.3 |
| News studio | 35.7 | 35.5 | 35.5 | 35.7 |
| Nighttime | 5.7 | 6.1 | 6.1 | 7.2 |
| Old people | 6.1 | 5.3 | 8.0 | 7.9 |
| Overlaid text | 14.2 | 12.8 | 12.8 | 12.8 |
| People marching | 1.8 | 1.8 | 1.8 | 2.0 |
| Reporters | 35.7 | 37.2 | 38.0 | 37.1 |
| Running | 5.1 | 5.2 | 5.2 | 5.3 |
| Scene text | 3.3 | 3.7 | 3.7 | 3.7 |
| Singing | 2.3 | 3.4 | 3.4 | 3.6 |
| Sitting down | 0.2 | 0.2 | 0.2 | 0.1 |
| Sky | 18.5 | 17.9 | 17.9 | 17.9 |
| Sports | 12.9 | 13.9 | 13.9 | 14.1 |
| Streets | 17.4 | 19.6 | 19.6 | 19.7 |
| Two people | 4.2 | 6.7 | 6.1 | 6.0 |
| Walking | 5.9 | 8.6 | 8.6 | 8.6 |
| Walking or running | 5.3 | 7.0 | 7.0 | 7.5 |
| Door opening | 3.0 | 2.0 | 2.0 | 2.0 |
| Event | 1.9 | 3.0 | 3.0 | 3.0 |
| Female human face | 9.8 | 11.6 | 13.0 | 12.9 |
| Flags | 0.5 | 0.9 | 0.9 | 0.9 |
| Head and shoulder | 6.6 | 7.7 | 7.7 | 7.7 |
| Male human face | 10.7 | 9.6 | 10.1 | 10.3 |
| News | 23.2 | 12.2 | 12.2 | 26.3 |
| Quadruped | 4.9 | 8.9 | 8.9 | 9.5 |
| Skating | 17.9 | 17.4 | 17.4 | 15.0 |
| Speaking | 9.6 | 10.6 | 11.2 | 11.4 |
| Speaking to camera | 9.2 | 13.7 | 15.6 | 15.8 |
| Studio with anchorperson | 42.3 | 46.1 | 46.1 | 51.7 |
| Table | 4.6 | 7.5 | 7.5 | 7.5 |
| Text | 10.1 | 11.0 | 11.0 | 10.4 |
| Traffic | 16.3 | 22.3 | 22.3 | 22.5 |
| Urban scenes | 10.6 | 12.2 | 12.2 | 12.4 |
| Mean | 10.7 | 11.7 | 11.9 | 12.3 |

Table 4.10: Performance evaluation of the *TRECVid* 2011 runs.

2010. Due to the large amount of video data the object models were only applied to the keyframe images as in the previous year. Each deformable part-based model consists of six components which intuitively corresponds to the different views of an object. The component-based max-pooling strategy was applied to the detection results of these models resulting in a 120-dimensional feature vector. For the object class "face", the number of detected faces, the average and maximum probability, as well as the average and maximum size of the detected bounding boxes are determined. The final object-based feature vector is the concatenation of the component-based probabilities and the face-related features.

In the following experiments, the BoVW features are combined with object-based feature representations using $L_2$-norm MKL. For all feature representations the $\chi^2$-kernel is used to measure the similarities between the data instances.

In the first experiment (Run 1), the BoVW features were combined with the object-based features using the MKL fusion scheme. This approach additionally considering object-based features was significantly better than the baseline system at a significance level of 1%. Many concepts clearly profited from object-based features, for example, "anchor person", "car", "quadruped", "streets", "speaking to camera", "table", "traffic" or "two people".

In a further experiment (Run 2), BoVW features were extracted from detected object regions and were used as an additional feature representation. Due to the large number of face-related concepts, like "adult", "female", "male", "dark skinned person", "first lady", "glasses", or "Arafat", BoVW feature representations were extracted from face regions. Therefore, an additional face-related codebook of 1,000 visual words was constructed. For the codebook as well as for the histogram generation, the face regions were scaled to 50x50 pixels and densely sampled RGB-SIFT descriptors were extracted with a sampling step size of two pixels, eight orientation bins, and a spatial bin size of four. Again, the soft-weighting scheme is applied to build the BoVW feature representations. If several face regions are detected in an image they were summarized in a single histogram. These region-based BoVW features were again combined with the previous representations using the MKL approach. Although most face-related concepts could be slightly improved, like "old people" or "speaking to camera", this run achieved only slight performance improvements compared to Run 1.

In the last experiment (Run 3), a new post-processing scheme is introduced that leads to a rescoring of shots based on the concept relations. Two types of relations between the semantic concepts are provided by the organizers of the semantic indexing task: implications and exclusions. The relation "A $\Rightarrow$ B" for example is valid if concept A is a subclass or specification of concept B. This relation is also true if concept B is a part of concept A. Besides implications, two concepts can exclude each other such as "Indoor" and "Outdoor". The given set of concept relations contains 427 implications and 559 exclusions. The relations are implemented in a simple post-processing framework by adding and subtracting scores from the individual concept detection results. The two relation types are processed as follows:

The relation "A $\Rightarrow$ B" is realized by taking the positive shots of concept A into account and increasing the scores of the corresponding shots in the ranked list of concept B (Lines 1-7 of Listing 4.2). Additionally, the logically implicated relation "$\neg$B $\Rightarrow$ $\neg$A" is considered and the confidence scores of concept A are reduced for shots with a negative score for concept B (Lines 11-17 of Listing 4.2). But this relation is only applied if the detector of concept B predicted at least one positive shot on the test set (Line 9 of Listing 4.2). Otherwise, the corresponding results of concept B are not to be trusted.

```
   foreach $shot$ in $shots$
 2 do
       if ($score_A(shot)$ > 0)
 4     then
           $score_B(shot) \mathrel{+}= score_A(shot)$
 6     fi
   done
 8
   if ($\max_{shot \in shots}(score_B(shot)) > 0$)
10 then
       foreach $shot$ in $shots$
12     do
           if ($score_B(shot)$ < 0)
14         then
               $score_A(shot) \mathrel{+}= score_B(shot)$
16         fi
       done
18 fi
```

Listing 4.2: Concept relation "A implies B".

```
if ($\max_{shot \in shots}(score_A(shot)) > 0$)
then
    foreach $shot$ in $shots$
    do
        $score_B(shot) \mathrel{-}= score_A(shot)$
    done
fi

if ($\max_{shot \in shots}(score_B(shot)) > 0$)
then
    foreach $shot$ in $shots$
    do
        $score_A(shot) \mathrel{-}= score_B(shot)$
    done
fi
```

Listing 4.3: Concept relation "A excludes B".

Exclusions are bidirectional which means that “A excludes B” as well as “B excludes A” are valid. For both directions a simple kind of confidence prediction is performed in order to decide whether the relation is useful (Lines 1 and 9 of Listing 4.3). The relation “A excludes B” is realized by subtracting the confidence scores of concept A from the corresponding scores of concept B and vice versa (Listing 4.3).

Compared to the reference system (Run 2), the post-processing step achieves a relative performance improvement of 3.8% in terms of mean infAP. In particular, the concepts “news” and “studio with anchorperson” took advantage of the relation-based rescoring.

Using all extensions, the last run achieved the best overall performance improving the baseline system from 10.7% to 12.3% mean infAP. In comparison to other participating teams, the best result was achieved for the concepts “overlaid text” with 14.2% infAP and “two people” with 6.7% infAP. Overall, the results were among the five best teams for the concepts “car”, “female human face”, “overlaid text”,“two people”, and “text”. Altogether, the experiments showed that the additional use of object-based features significantly improved the concept detection performance – even in combination with state-of-the-art BoVW mid-level features. Further improvements were attained using BoVW feature representations from face regions and a relation-based rescoring scheme.

### 4.6.4 Cross-Domain Concept Detection

This section deals with the improvement of cross-domain concept detection via object-based features. The generalization capabilities of visual concept models across different user domains are a severe problem. The visual appearance of semantic concepts strongly depends on the domain of the particular image or video source. This can be easily observed, for example, in the difference between television news and user-generated YouTube videos. Even within the news domain the appearance of several concepts depends on the broadcast channel due to different editing styles or studio layouts. The appearance of the concepts “anchor”, “maps”, or “weather forecast” in news videos, for example, are typically related to a TV broadcaster or program. The spatial composition of these shots is specific for a TV cast, such as the moderator’s position, the camera distance in an anchor shot, or the used colors for displaying a map. Example images for the concept “anchor”, whose appearance depends on the program channel, are presented in Figure 4.3.

In the following, the generalization capabilities of concept models are investigated in a channel cross-domain and in a genre cross-domain setting. The experimental setting involves two different data sets: news videos from the *TRECVid Challenge* 2005 and documentary films from the *TRECVid Challenge* 2007. The common concept set comprises the LSCOM-lite concepts except “entertainment”, “government leader”, and “corporate leader” which have been dropped at the *TRECVid Challenge* 2007.

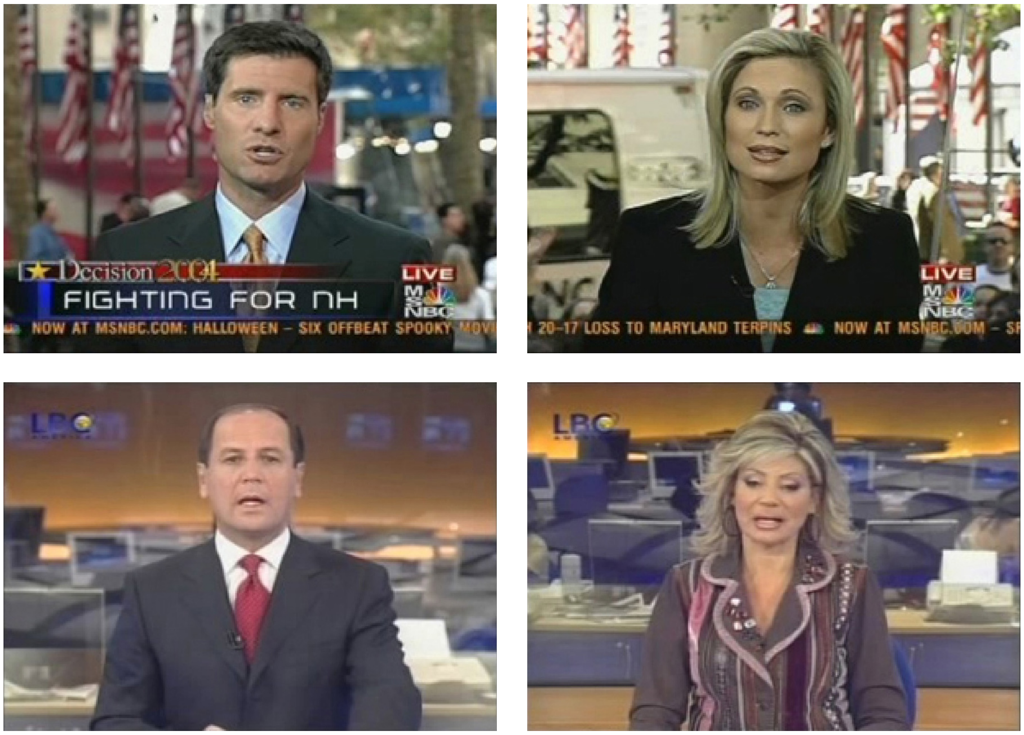

Decision 2004
FIGHTING FOR NH
LIVE
MSNBC
NOW AT MSNBC.COM: HALLOWEEN – SIX OFFBEAT SPOOKY MOVI
LIVE
MSNBC
20-17 LOSS TO MARYLAND TERPINS
NOW AT MSNBC.COM – SF
LBC
LBC

Figure 4.3: Keyframes from different news channels showing the concept "anchor".

The baseline system uses state-of-the-art BoVW features with densely sampled SIFT descriptors, a step size of five pixels, a 4,000-dimensional visual codebook and the soft-weighting scheme, as described in Section 3.2.2. Concept models are again built using the SVM$^{\text{light}}$ implementation of Joachims [1998] with the $\chi^2$-kernel. Object-based features are extracted based on the 20 object classes of the *VOC Challenge* as depicted in Figure 2.11. The component-based max-pooling strategy is used to derive a 120-dimensional feature vector, whose feature values indicate the probability of an object class being present in an image. Additionally, a face detector is used to derive the number of faces, the maximum and average probability as well as the maximum and average size of the facial areas. Altogether, the concatenation of the object-based feature representations yields a 125-dimensional feature vector per shot. The object-based feature representations are combined with BoVW features in a MKL setting. The kernel weights are constraint using the $L_1$-norm.

The channel cross-domain experiments were conducted using the *TRECVid* 2005 development set which has been divided by the *Mediamill Challenge* into a training and a test set. It consists of 86 hours of news videos from different news channels including English (CNN, NBC and MSNBC), Chinese (CCTV and NTDTV), and Arabic (LBC) channels. While the *Mediamill Challenge* merges

| Channel | Train | Test |
|---|---:|---:|
| CCTV4 | 8,147 | 2,749 |
| CNN | 8,036 | 2,992 |
| LBC | 11,875 | 3,398 |
| MSNBC | 6,407 | 2,498 |
| NBC | 4,784 | 4,538 |
| NTDTV | 4,766 | 1,715 |
| **Total** | **44,015** | **17,890** |

Table 4.11: Number of training and test images for the different news channels.

subsequent (sub)shots which are shorter than two seconds to “master shots”, the LSCOM-lite annotations are based on the subshots. Thus, the dataset comprises 44,015 training shots and 17,890 test shots. The number of training and test shots of the channel subsets are presented in Table 4.11.

While the average within channel performance using BoVW representations achieves 37.77% mean AP, the average channel cross-domain performance achieves only 23.51% mean AP. This result shows the poor generalization capabilities of concept models across different news channels. Additionally using object-based features achieves an average within channel performance of 37.82% mean AP and an average channel cross-domain performance of 24.35% mean AP. The individual results are presented in the confusion matrices of Table 4.12 and Table 4.13. In most cross-channel combinations the additional use of object-based features yields a significantly better result than using BoVW features alone. Table 4.14 shows the confusion matrix of the significance test results. The significance tests were conducted using the partial randomization test with a significance level of 5%.

Further experiments are conducted in a genre cross-domain setting between news videos and documentary films. The documentary films from the *TRECVid Challenge* 2007 comprises 50 hours of development data and 50 hours of test data. The shot segmentation yielded 21,532 training and 22,084 test shots. The ground truth labels for the *TRECVid* 2007 development set are the result of a collaborative annotation effort of the participating teams. While the development set is completely annotated with the 36 LSCOM-lite concepts, ground truth data for the test videos exists only for the subset of 20 concept classes listed in Table 4.17. The runs on the *TRECVid* 2007 test set are evaluated by the infAP measure, which is based on a total of 66,293 shots judged by NIST in 2007. For each of the 20 concept classes NIST pooled shots from the top of the submitted retrieval results. A 50% random sample of each pool was finally judged. Thus, for each of the 20 concept classes about 3,315 shots are annotated on average. The infAP measure is calculated using the `trec_eval` tool, which is publicly available at the *TRECVid* website.

In a first experiment, the generalization capabilities of “news models” to documentary films are investigated. Models for the 36 LSCOM-lite concepts are

| **AP [%]** | **CNN** | **CCTV4** | **LBC** | **MSNBC** | **NBC** | **NTDTV** |
|---|---|---|---|---|---|---|
| CNN | 38.8 | 27.7 | 23.1 | 28.0 | 24.1 | 22.7 |
| CCTV4 | 24.0 | 38.1 | 26.4 | 21.1 | 22.7 | 22.6 |
| LBC | 23.7 | 28.3 | 46.0 | 23.1 | 22.8 | 25.2 |
| MSNBC | 28.0 | 24.6 | 23.0 | 32.2 | 25.1 | 20.8 |
| NBC | 23.8 | 23.7 | 24.0 | 24.0 | 37.0 | 21.4 |
| NTDTV | 19.7 | 21.5 | 22.3 | 18.8 | 19.3 | 34.6 |

Table 4.12: Confusion matrix of the channel cross-domain experiment using BoVW features.

| **AP [%]** | **CNN** | **CCTV4** | **LBC** | **MSNBC** | **NBC** | **NTDTV** |
|---|---|---|---|---|---|---|
| CNN | 39.1 | 28.6 | 23.4 | 29.2 | 25.5 | 22.5 |
| CCTV4 | 25.4 | 38.1 | 22.8 | 23.0 | 24.7 | 23.1 |
| LBC | 24.5 | 28.9 | 44.6 | 24.1 | 24.8 | 26.2 |
| MSNBC | 28.6 | 25.6 | 23.0 | 32.6 | 27.0 | 21.9 |
| NBC | 24.4 | 24.6 | 22.4 | 24.3 | 36.6 | 22.7 |
| NTDTV | 21.2 | 22.7 | 23.5 | 20.4 | 21.2 | 35.9 |

Table 4.13: Confusion matrix of the channel cross-domain experiment using BoVW plus object-based features.

| | **CNN** | **CCTV4** | **LBC** | **MSNBC** | **NBC** | **NTDTV** |
|---|---|---|---|---|---|---|
| CNN | | > | | > | > | |
| CCTV4 | > | | | > | > | > |
| LBC | > | > | | > | > | > |
| MSNBC | > | > | | > | > | > |
| NBC | | | | | | > |
| NTDTV | > | > | > | > | > | > |

Table 4.14: Confusion matrix of significance test results, showing the runs additionally relying on object-based features, that performed significantly better than the reference system at a significance level of 5%.

built based on the training set of the news videos. These models are applied on the one hand to the news videos of the test set and on the other hand to the documentary films of the *TRECVid* 2007 development set. Both runs have been evaluated in terms of AP as well as deltaAP. DeltaAP has been calculated to consider the different occurrence frequencies of concept classes in the different domains. The results of the experiments from news videos to documentary films for the 36 LSCOM-lite concepts are presented in Table 4.15. For BoVW features the application of concept models built on news videos and applied to documentary films leads to a relative performance loss of 45.1% in terms of mean AP and of even 54.8% in terms of mean deltaAP in comparison to the performance within the same domain. These results show the poor generalization capabilites of the learned news models. By additionally using object-based features, the cross-domain concept detection performance has been significantly

| Source Domain | Target Domain | DSIFT | DSIFT+OBJ | Performance Measure | Rel. Performance Improvement |
|---|---|---|---|---|---|
| News | News | 41.7% | 42.1% | mean AP | 0.9% |
| News | Documentaries | 22.9% | 24.4% | mean AP | 6.62% |
| News | News | 35.6% | 36.0% | mean deltaAP | 1.1% |
| News | Documentaries | 16.1% | 18.0% | mean deltaAP | 9.44% |

Table 4.15: Genre cross-domain experiments from news videos to documentary films using the 36 LSCOM-lite concepts from *TRECVid* 2007.

| Source Domain | Target Domain | DSIFT | DSIFT+OBJ | Performance Measure |
|---|---|---|---|---|
| Documentaries | Documentaries | 9.74% | 9.84% | mean infAP |
| News | Documentaries | 5.96% | 6.97% | mean infAP |
| Documentaries | News | 8.94% | 9.83% | mean AP |
| News | News | 35.66% | 35.89% | mean AP |

Table 4.16: Genre cross-domain experiments based on the subset of 20 concept classes of the *TRECVid Challenge* 2007.

improved at a significance level of 1%. While the relative performance improvement amounts 0.9% in terms of mean AP and 1.1% in terms of mean deltaAP for BoVW features, the relative performance improvement averages out 6.62% and 9.44%, respectively, using object-based features.

In a second experiment, the results are evaluated based on the subset of 20 concepts, which have been judged by the *TRECVid* team on the documentary films of the *TRECVid* 2007 test set. The results on the *TRECVid* 2007 test set are evaluated using the infAP measure. First, the news as well as the documentary models are applied to the documentary films of the *TRECVid* 2007 test set. On the domain of the documentary films the news models achieve a performance of 5.96% mean infAP with BoVW features and 6.97% mean infAP additionally using object-based features. This is a relative cross-domain performance improvement of 16.9% which is significantly better at a significance level of 1%. While the relative performance loss compared to documentary models amounts to 38.8% using BoVW features, it averages out only 29.2% additionally using object-based features. Table 4.17 shows the cross-domain results of the 20 concept classes for the BoVW representation as well as for the additional object-based features. Most of the concepts profited from the additional object-based features. Second, news and documentary models for the 20 concept classes are applied to the test set of the news videos. On the news domain the documentary models achieved only 8.94% mean AP without object-based features and 9.83% mean AP with object-based features compared to the news models with 35.66% and 35.89% mean AP, respectively. The domain of the documentary films seems to be more difficult than the news domain. Nevertheless, this is a relative cross-domain concept detection performance improvement of 9.96%, which is significantly better at a significance level of 1%.

| **infAP [%]** | **DSIFT** | **DSIFT+OBJ** |
|---|---|---|
| Sports | 8.25 | 8.62 |
| Weather | 0.02 | 0.00 |
| Office | 7.04 | 7.05 |
| Meeting | 5.81 | 9.60 |
| Desert | 4.07 | 6.65 |
| Mountain | 8.35 | 7.67 |
| Waterscape or waterfront | 17.97 | 16.52 |
| Police security | 0.12 | 1.41 |
| Military | 0.78 | 0.67 |
| Animal | 4.30 | 4.69 |
| Computer or TV screen | 4.49 | 6.02 |
| Flag US | 0.32 | 0.08 |
| Airplane | 5.16 | 3.76 |
| Car | 17.52 | 23.75 |
| Truck | 3.41 | 6.07 |
| Boat or ship | 8.85 | 10.24 |
| People marching | 4.67 | 5.36 |
| Explosion or fire | 3.50 | 3.43 |
| Maps | 8.28 | 10.90 |
| Charts | 6.20 | 6.91 |
| Mean | 5.96 | 6.97 |

Table 4.17: Genre cross-domain results for the 20 evaluated concepts of the *TRECVid Challenge* 2007 using models trained on news videos and applied to the *TRECVid* 2007 test set.

The models built on news videos have been also applied to the development set of the documentary films. These models yielded 11.35% mean AP using BoVW features and 13.22% mean AP additionally using object-based features in comparison to 5.96% and 6.97% mean infAP on the test set. This shows that although infAP is an estimation of the AP score, the measurements does not seem to be directly comparable. Nevertheless, the relative performance improvement in both cases is almost the same.

Altogether, different source and target domains lead to a severe loss in detection performance, both for different channels and genres. But the cross-domain concept detection performance can be clearly improved via object-based features.

## 4.7 Summary

In this chapter, an approach for the task of visual concept detection has been presented that systematically utilizes results of object detectors. After the discussion of related work, current object detection approaches that are used throughout the experiments have been presented. Since it is inefficient to run a large number of single-class object detectors, it has been demonstrated how a concurrent multi-class object detection system can be constructed to speed up the detection of many object classes in images. This multi-class approach, which is an extension of class-specific Hough forests, is capable of detecting 24 different object classes at a time, instead of applying one object detector for each object class separately.

Using object detectors, semantic features were generated based on object detection results using different pooling strategies. For videos, detection results were assembled to object sequences and a shot-based confidence score as well as further features, such as position, frame coverage or movement, were computed for each object class. These object-based feature representations were used as additional input for the SVM-based concept classifiers. Thus, other related concepts can also profit from object-based features.

Experimental results were conducted on the *Mediamill*, *VOC* and *TRECVid Challenge*. Systems additionally using object-based features performed significantly better than the baseline system, even in combination with BoVW mid-level features. Additionally, different pooling and fusion strategies have been investigated. The experimental results on the *VOC Challenge* showed that the max-pooling approach is the most advantageous strategy for the representation of object-based features. BoVW features and object-based features seem to be very complementary. The combination of both representations led to a relative performance improvement of 18.9% mean AP compared to the BoVW approach. Furthermore, the MKL framework outperformed the early and late fusion strategies and led to a superior performace for the image classification task of 63.8% mean AP. Top results were also achieved at the annual *TRECVid Challenge* by using object-based features. The experiments revealed that the approaches additionally employing object-based features significantly improved the overall performance. Some concepts like "animal", "bicycling" or "vehicle" were improved by more than 100%. For "vehicle", "ground-vehicle", "overlaid text" and "two people" the best results in terms of infAP of all participating teams were achieved.

Finally, the generalization capabilities of concept models have been investigated. Different source and target domains led to a severe loss in concept detection performance, both for different channels and genres. It has been shown that object-based features clearly improve the cross-domain concept detection performance and thus the generalization capabilities of concept models.

Overall, it can be concluded from the experiments that object-based feature representations improve the performance not only for the concepts that directly correspond to the object classes, but also for a large number of indirectly related concepts. It has been demonstrated that a few object-based features are beneficial for a large number of concepts.

*"Man muß das Unmögliche versuchen, um das Mögliche zu erreichen."*

Hermann Hesse

# 5 Long-Term Incremental Web-Supervised Learning of Visual Concepts

## 5.1 Introduction

In recent years, many people have started to maintain private archives of images and expose them on Web 2.0 platforms. Tags and annotations of multimedia data in the internet are rich sources of information that can be employed for learning visual concept classes. However, such annotations have several drawbacks with respect to completeness and correctness due to limited time, objectivity, and expertise of the human annotators. Several proposals have been made to use the large amount of images in the WWW as training data. Training data sources are, for example, Flickr [Kennedy et al. 2006; Xu et al. 2009] or Google's image search [Fergus et al. 2005]. Different methods are applied to learn visual concepts: e.g., translation and scale invariant pLSA (TSI-pLSA) [Fergus et al. 2005], k-nearest neighbor (kNN) classifiers [Liu et al. 2009], or conditional random fields [Xu et al. 2009], whereas only a few proposals deal with incremental learning of classification models based on training data from the internet like the approaches of Datta et al. [2007], Li and Fei-Fei [2009], and Zhang et al. [2009]. Most approaches use images from the WWW in a batch learning mode, assume a homogeneous appearance of visual concepts, or focus on removing inappropriate training data (e.g., [Fan et al. 2009; Fergus et al. 2005; Gong et al. 2009; Kennedy et al. 2006; Liu et al. 2009; Xu et al. 2009]).

In this chapter, a novel purely web-supervised learning approach for modeling heterogeneous concept classes in images is proposed. The approach is aimed at continuous long-term learning of appearance models and improving these models periodically. It is called web-supervised since only images from the WWW are used for learning, and the learning process is initiated by simply providing the concept name. For this purpose, a novel learning component, called *random savanna*, is proposed that deals with the heterogeneous appearance of visual concepts. The learning framework consists of several components: a WWW crawling component, a multi-modal clustering component for spam detection as well as for subclass identification, a sub-space ensemble learning component, a self-validation component, an updating component, and a scalability manager.

Multi-modal clustering results are evaluated and non-spam clusters are considered as relevant concept subclasses. For each subclass a random forest is built. The updating component includes new training data while at the same time it ensures that the performance is improved by utilizing a validation set. The scalability manager guarantees that a limited number of training images is used in each training round. Experimental results demonstrate the capabilities of the proposed web-supervised learning framework. The framework outperforms a SVM on a test set that consists of representative web images exhibiting a number of heterogeneous image classes.

This chapter is organized as follows: Related work is discussed in Section 5.2. In Section 5.3, the web-supervised learning system is presented. Experimental results are shown in Section 5.4. Section 5.5 summarizes the chapter.

Parts of this chapter have been published by Ewerth et al. [2012].

## 5.2 Related Work

First, approaches using web images to train image classification models in a batch learning mode are presented. Then, methods based on incremental learning are discussed.

Kennedy et al. [2006] addressed the question when to use manually annotated or search-based data for learning a classification model. The authors investigated 19 features to predict the performance of search-based and manual annotations. They found out that cross-domain image similarity and model generalization are strong predictors, in contrast to concept frequency and in-domain model quality.

Fergus et al. [2005] presented an approach in which object classes are learned using results of Google's image search. The authors proposed TSI-pLSA as an extension of pLSA to gain scale and rotation invariance.

Schroff et al. [2011] presented a system that automatically gathers a large number of web images for a given concept class. Noisy images are ranked using a *Bayes posterior estimator* that has been trained on the surrounding text and metadata of the downloaded images. The visual features of the top-ranked images are used as positive training data to learn a SVM-based model. Finally, the resulting classifier is used to re-rank the downloaded images. Experiments on 18 concept classes showed that this system outperforms Google's image search.

Xu et al. [2009] used Flickr tags to automatically annotate web images. Correlations between Flickr keywords are exploited for analyzing semantic inter-relationships. Conditional random fields are used to model and integrate the information from different sources: visual features according to the MPEG-7 standard and textual keyword information. Experimental results were presented for a dataset of 5,000 images from the WWW where 1,000 images were used as training set.

Gong et al. [2009] presented a system that employs the text associated with images in websites. For a given term, the related web images are clustered using GMMs based on visual features. A proposed word promotion method is used to finally re-weight the term vector based on co-occurrences of semantic terms. The authors reported results for a training set of 12,000 web images and 200 test images.

Fan et al. [2009] proposed a framework for multi-label learning that includes a tag-cleansing algorithm to deal with the issue of synonymous, spam as well as loose and ambiguous tags. A visual concept network is generated that consists of concept classes and characterizes the inter-concept visual similarity contexts.

Web data from photosig.com has been used by Liu et al. [2009] to find training images for a textual query. Using the relevant web images for a given query, a kNN-classifier is used to retrieve the consumer photos. Cross-domain learning using regularized regression is realized via relevance feedback of users that label data from the target domain. Experimental results are presented for Kodak's consumer video benchmark and Corel's stock photo dataset.

Torralba et al. [2008] collected approximately 80 million images from the web covering a wide range of concepts. These images were resized to a very low resolution of 32x32 pixels. The authors showed that for certain concepts a simple nearest-neighbor approach in conjunction with semantic information from the WordNet database yields very good detection performance.

Jiang et al. [2009b] proposed an approach to leverage knowledge data from heterogeneous sources in order to achieve domain adaptation for video search. They explored context information associated with Flickr images and suggested Flickr context similarity to measure the similarity of query detectors. Offline and online semantic context transfer across different domains were distinguished. Experimental results were presented for *TRECVid* video data sets from the domains of broadcast news and documentary films.

Another approach relies on the assumption that Flickr users group their images into batches that are related to an event [Ulges et al. 2011]. The authors proposed to build group-specific annotation models that consider a specific context setting. In case of the Flickr domain, it is assumed that a Flickr group represents a specific context. The authors presented group-specific extensions for two image annotation models: texton and pLSA. Finally, a group-specific model was applied to annotate images.

Wang et al. [2006] presented an automatic annotation system to facilitate the search process for personal images. Content-based image retrieval is used to collect a set of similar web images. The annotations of these web images are merged to build a ranked list of possible keywords. The top-ranked keywords are used for the final annotation of a personal image. Experimental results on a test set of the University of Washington for content-based image retrieval demonstrated the effectiveness of the proposed system.

A similar approach has been presented by Wang et al. [2008c]. In this approach, a keyword is associated with the query image. Then, semantically similar web images are searched, and their annotations are used to tag the query image. A rejection step is used to filter out noisy annotations. Real-time capability was achieved by mapping images to hash codes and using a distributed service-oriented architecture. Experimental results were reported for Google search results and a test set of the University of Washington.

Tang et al. [2009] inferred semantic concepts for images from web images provided in community forums. A semi-supervised sparse graph-based approach is proposed that exploits labeled and unlabeled data simultaneously. To remove noisy tags, a label refinement strategy is proposed. Experimental results were presented for a test set of 55,000 Flickr images.

Zhao et al. [2010] presented a classifier-free system for annotating web videos that is based on using tags of near-duplicate web videos. Experimental results are presented for Google, Yahoo, and YouTube videos.

Apart from the approaches for learning visual concept models in a batch mode, there are only a few proposals that deal explicitly with incremental learning of classification models based on training data from the WWW. Li and Fei-Fei [2009] presented the system OPTIMOL that uses a number of manually given seed images to find more images of an object class in the internet. Web images are used to improve the object model, relying on Bayesian incremental learning of a latent topic model. Results for collecting images for 23 classes are reported.

Datta et al. [2007] presented an annotation system that exploits collaborative image tagging. Three sources of information are used for annotating images automatically: visual features, knowledge from the WordNet database, and an initial “black-box” classifier. The models are learned incrementally and are refined when new human tag information is available or has changed.

Zhang et al. [2009] proposed an incremental SVM with a fixed number of support vectors (n-ISVM) for web video categorization. Experimental results were presented for a test set of YouTube videos. It has been shown that n-ISVM achieves a comparable performance to a standard SVM.

## 5.3 Web-Supervised Random Savannas

In this section, a novel approach for continuous web-supervised learning is presented as depicted in Figure 5.1. The proposed system differs from related work in several ways. First, the learning process is initiated by simply providing the

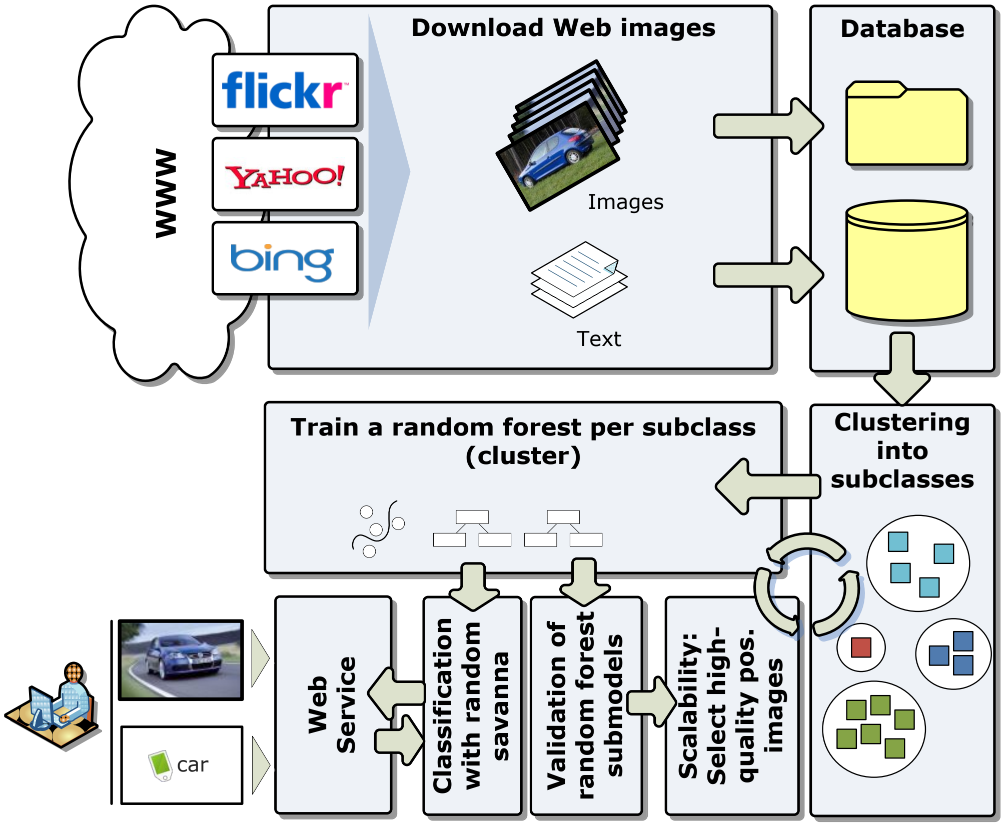

WWW
flickr
YAHOO!
bing
Download Web images
Images
Text
Database
Clustering into subclasses
Train a random forest per subclass (cluster)
Web Service
Classification with random savanna
Validation of random forest submodels
Scalability: Select high-quality pos. images
car

Figure 5.1: Overview of the web-supervised system.

name of the concept to the system, and no seed images are required (e.g., in contrast to Li and Fei-Fei [2009]). Second, it explicitly considers the learning of visual concepts as a continuous long-term incremental learning task that is aimed at improving its detection ability periodically. Third, it considers the task of concept learning as a task of subclass learning in order to model heterogeneous image classes. For example, images showing the concept "car" consist of frontal views, side views, and rear views, there are many types of different cars from different periods of time, they have different colors, and cars are shown in very different settings. Hence, the proposed system tries to find reasonable subclasses in the WWW training data that are used to learn specialized subclass detectors. The learning of subclasses is realized via a subspace ensemble learning approach. All subsequently described algorithmic steps are conducted separately for each specified target concept.

The system consists of the following components that are described in more detail in the subsequent sections:

- **Crawling**: A web crawler that automatically acquires training data from the WWW for a given concept.

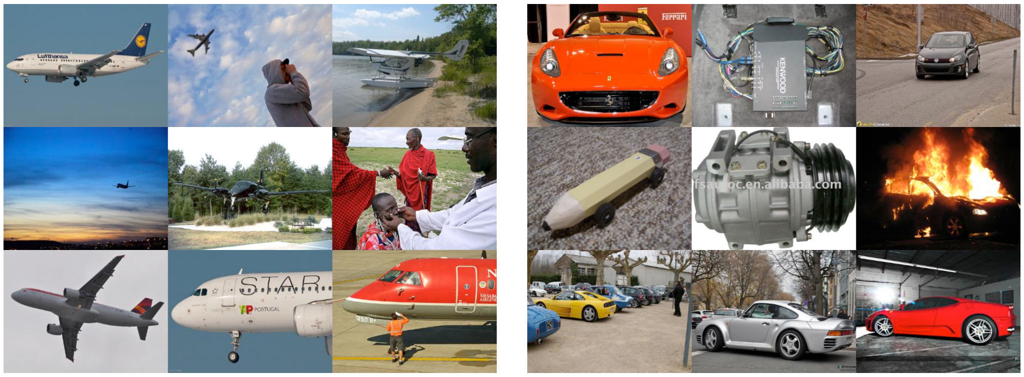


Figure 5.2: Example images downloaded from the WWW for the concepts "airplane" and "car".

- **Subclass modeling and spam detection**: A multi-modal and multi-feature clustering component that removes inappropriate training data and uses adequate remaining clusters for subclass modeling.

- **Subclass learning**: Random forests are used to train a classifier for each sub-model. These classifiers are bundled for ensemble classification.

- **Validation**: The validation component is employed to evaluate sub-models and their related classifiers on the validation set. The validation set can be either selected automatically from the training data, or it can be a predefined data set to address a target domain.

- **Ensemble classification via random savanna**: The ensemble classification component assembles the votes of all classifiers and generates a decision for a given test image.

- **Updating**: The updating component adds new training data, continues and updates the clustering process and proceeds with the subclass learning procedure. Good subclass models and related training examples are kept, while bad training data (i.e., spam) are discarded. The trees of a random forest are evaluated on the held-out validation set which can be generated automatically.

- **Scalability management**: This component is tightly coupled with the updating component and is responsible for the scalability of the approach. It retains a small number of images in the system of good quality for the training process.

## 5.3.1 Acquiring Training Data from the WWW

The system is initialized by simply specifying the target concept, e.g., “car”. Then, the system starts crawling the WWW and searching for appropriate training images of the target concept. The current version of the system focuses on querying selected websites. Given the name of a concept, the system searches for images tagged with that concept on the following websites: Yahoo image search (`http://images.search.yahoo.com`), Flickr (`http://flickr.com`), and Bing image search (`http://bing.com`). To enhance the retrieval results, the English concept term is translated to several languages: French, German, Spanish, and Italian. The corresponding five concept terms are used to query Flickr and the search engines of Yahoo and Bing. All three websites are accessed using the corresponding application programmering interface (API). Then, the query results are used to access and retrieve image data and related text information from the concerning website. Examples of downloaded web images are presented in Figure 5.2 for the concepts “airplane” and “car”.

## 5.3.2 Discovering Spam and Visual Subclasses

Once the training images are downloaded for the target concept, it is important 1.) to filter out spam images, and 2.) to arrange the remaining positive training images of the target concept in reasonable clusters that reflect their visual similarity according to one or more feature sets. For this purpose, each image is described via a feature set that consists of textual or visual information. Textual features are derived from related tags and surrounding textual information, as described in the following subsection. HoVW representations are used as visual features. The HoVW features are based on a 1,000-dimensional codebook that is generated using Lowe’s SIFT features [Lowe 2004].

### Spam Detection Using Textual Features

For the images taken from Flickr, the tags and the image-related comments are extracted. For search results using image search engines, surrounding text of the related webpage is used. The information from the HTML “img-tag” and the “alt-attribute” are used. Furthermore, the words before (e.g., 10 words) and after the “img-tag” (e.g., 50 words) are extracted for subsequent processing. Stop word removal and stemming is applied to all extracted textual information using the *Porter stemmer* [Porter 1980]. The remaining $N$ words form a lexicon for the given target concept. Each image is described via a feature vector whose components represent one of the lexical terms. The feature vector is normalized using term frequency and inverse document frequency (tf-idf). Then, the feature vectors are clustered using the K-means algorithm. This algorithm groups the data in $k$ clusters by minimizing the variance of the cluster centers; $k$ is iterated from 2 to $k_{max}$ (e.g., 20) to select the best $K$ using the *Silhouette coefficient* (SC) [Ester and Sander 2000]. Since the positive training images to be clustered are always related to only one specified target concept, a search range from 2 to 20 is sufficient in most cases. In this way, the best $K$ for clustering the web images based on textual information is estimated. A cluster is discarded as spam if the number of elements or its related SC is below a threshold.

```
TextualSpamDetection
do
    Build lexicon L of words from surrounding textual data;
    Represent images as histograms of words using tf-idf;
    Cluster images according to the histograms;

    foreach cluster C[i]
    do
        if (|C[i]| < minNr || silhouetteCoeff(C[i]) < minSC)
        then
            Discard C[i] as spam;
        fi
    done

    Estimate a set of hyponyms H for the concept using WordNet;
    Create ‘‘optimal’’ feature vector Q;
    foreach word L[i]
    do
        if (L[i] in H)
        then
            Q[i] = 1;
        else
            Q[i] = 0;
        fi
    done


    foreach cluster C[i]
    do
        d[i] = distance(C[i],Q);
    done

    Sort C according to d in descending order;

    s = argmax(d[i] - d[i+1]);
    Discard (C[s+1],...,C[|C|]) as spam;
    Choose (C[1],...,C[s]) as relevant non-spam clusters;
    return;
done
```

Listing 5.1: Pseudocode for spam filtering of a given concept based on textual information.

Then, the remaining clusters are analyzed with respect to their semantic similarity to the given target concept. For this purpose, a set of hyponyms $H$ is generated using the WordNet database. An artificial "optimal" feature vector $Q$ (with $N$ dimensions) is generated whose components are set to 1, if they represent a concept from H (and otherwise 0). In other words, all components of this feature vector are set to 1 that are semantically related to the given target concept (i.e., corresponding either to hyponyms of the target concept or to the target concept itself). Then, the distances of all $K$ cluster centroids with respect to Q are computed $(D_1, \ldots, D_K)$, and sorted in descending order $(D_1{}', \ldots, D_K{}')$. The idea of this step is that the first cluster of this sorted list is semantically most similar to the target concept. The maximum of two successive distances $(D_s{}', D_{s+1}{}')$ is calculated and used to split the sorted list of clusters into relevant (non-spam) clusters $(D_1{}', \ldots, D_s{}')$ and spam clusters $(D_{s+1}{}', \ldots, D_K{}')$. The related images of spam clusters are discarded from the training set. The pseudocode for the algorithm is presented in Listing 5.1.

### Unsupervised Visual Subclass Discovery

The non-spam images for the given target concept are now passed to a second clustering process that is aimed at grouping the positive training images according to object pose, setting and context and at searching for re-occurring patterns in the images. This task is somewhat related to unsupervised visual object discovery [Tuytelaars et al. 2009]. In the presented approach it is used to discover subclasses of the specified target concept. Each image is represented by a HoVW using salient point SIFT descriptors [Lowe 2004]. The K-means algorithm is used again to cluster the feature vectors, and $k$ is iterated from 2 to $k_{max}$ (e.g., again 20) to select the best $K$ for clustering of the target concept using the SC [Ester and Sander 2000]. A cluster is considered for subsequent classifier training if the following two conditions are met: 1.) the SC of the cluster exceeds a minimum value, and 2.) there is a minimum number of objects (images) in the cluster (e.g., 20). Then, for each cluster $C$, the average distance of its members is computed $(A_1, \ldots, A_K)$, and they are sorted in descending order $(A_1{}', \ldots, A_K{}')$. The maximum of two successive distances $(A_m{}', A_{m+1}{}')$ is calculated and used to split the sorted list of clusters into relevant (non-spam) clusters $(D_1{}', \ldots, D_m{}')$ and spam clusters $(D_{m+1}{}', \ldots, D_K{}')$. Only the remaining non-spam clusters are used for the subsequent training of random forests. In the following, these remaining clusters are called *subclass models*; they are expected to represent a reasonable subspace of the visual concept. Examples for some clustering results using visual features are presented for the concept "car" in Figure 5.3 and 5.4. Figure 5.3 shows two different clusters (subclasses) for "car", while Figure 5.4 shows two clusters for "car" that are discarded as spam. As an alternative for the K-means algorithm, another clustering method using an EM-based approach has been tested.

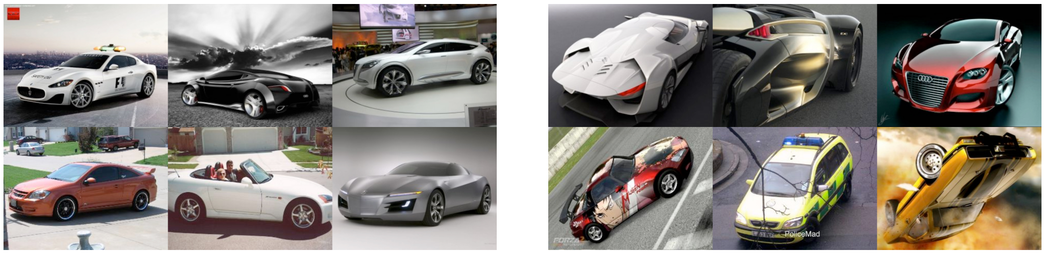


Figure 5.3: Two different subclasses (clusters) for the concept “car”.

### Subclass Learning via Random Forests

Each cluster of images that remains in the system after visual spam detection represents a subclass model that is used to train a classifier. All subclass classifiers are finally assembled in an ensemble. The images in each cluster are used as positive training samples for the related subclass of the target concept, while the downloaded images for all other concepts in the system are used as negative training examples. The goal of this component is to learn an adequate model for a given cluster of training images. Although the clustering process should yield clusters of similar images, in practice there will still be noisy images in the clusters. The question is which kind of supervised learning method is suitable for this kind of learning task. Random forests [Breiman 2001] have been chosen based on the following considerations. First, Breiman [2001] has shown empirically that random forests are more robust than Adaboost with respect to noise in the training data, i.e., mislabeled training examples. This is an important property for classifier training regarding the addressed scenario of web-supervised learning, where the probability of falsely labeled images is rather high. Second, another important advantage is that the construction of trees has a rather low computational complexity compared to other methods, e.g., SVMs. Third, it is easily possible to exchange trees in a random forest that do not perform well, or to insert trees that perform well. For example, the simplicity of realizing incremental learning by modifying existing random forests is an important advantage compared to SVMs. In addition, in Section 5.3.6 is demonstrated that random forests support the scalability of the system in a natural way. Moreover, training and classification are simply parallelizable at the level of decision trees. Last but not least, random forests have been applied successfully for image classification [Bosch et al. 2007].

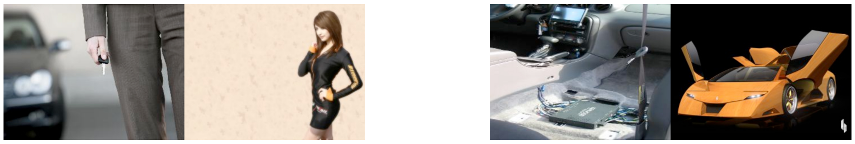


Figure 5.4: Two subclasses (clusters) for the concept “car” discarded as spam.

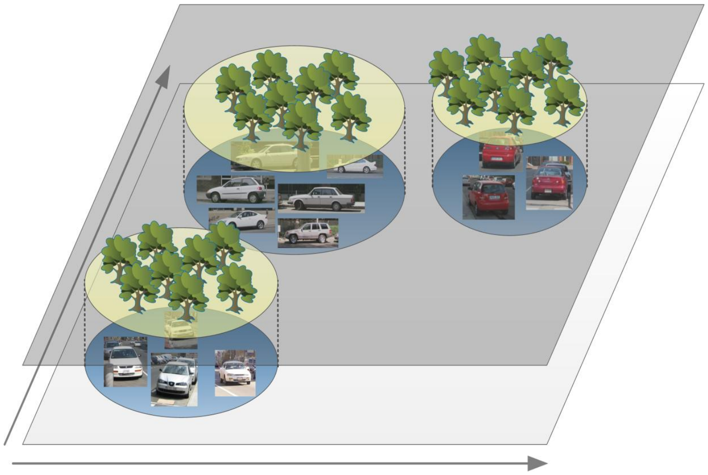


Figure 5.5: Random savanna abstraction.

Random forests generate an ensemble of random tree classifiers based on the given training data. Several possibilities exist to inject randomness in trees and forests. In the presented system, a random forest is generated by bagging, i.e., by randomly sampling the training examples for generating the random trees of the ensemble. A subset of $k$ features is randomly selected for a node in a random tree and the feature(s) that yield the best split for this node are selected. This best split (feature) for the data is chosen based on the information gain criterion. Given an unknown test image, the image is passed to all trees in the forest and evaluated by them. The total score of a random forest for a test image is the averaged sum of all leaves' decision scores.

### 5.3.3 Classification via Random Savannas

The proposed classification component, called *random savanna*, consists of an ensemble of random forests, where each of the random forests represents a subclass model that belongs to the best classifiers on the validation set. The name "random savanna" is motivated by the following analogy. Using visual codebooks of size 1,000 or higher yields a large feature space. With respect to the positive training data used during the learning process most of this feature space is empty and is considered as a "desert". The positive training samples of each cluster cover a small region of the feature space representing a reasonable subclass, e.g., the rear view of cars. A random forest is planted for each cluster (subclass). Hence, the ensemble of these random forests is considered as a random savanna (Figure 5.5).

Given a new test image, all random forests are applied to the test image. Each random forest returns a score that reflects its confidence that the image shows the related concept. This raises the question of how the votes of the classifiers should be evaluated to obtain the final classification result. This result describes whether a visual concept is present in an image or not. Basically, the classification result of the majority of the random forests could be used as the final result. Otherwise, it is possible that an image belongs to one subclass and is classified correctly by the corresponding classifier, but is misclassified by all non-related classifiers. In the following, several ensemble strategies are discussed:

- **Idealistic strategy**: First, the somewhat theoretical best case scenario that rarely appears in practice is discussed. In this case, the assumption is that all subclass classifiers can be totally trusted and that a subclass classifier outputs "true" (i.e., the image contains the concept) only if the image falls into the corresponding subclass and "false" otherwise. The following rule could be applied: the image contains the concept if exactly one random forest classifies the image "positive", and the image does not contain the concept if all of the random forests classify the image "negative". In the sequel, this strategy is called "ideal". If two or more classifiers rate an image as "positive", several strategies are possible: it could either be believed (case 1), the outcome can be considered as "undefined" (case 2, "strict"), or any other strategy as described below can be used, for example, the majority-driven strategy.

- **Validation-driven strategy**: Another possibility is to exploit the classifier behavior on the held-out validation set. For each classifier, its accuracy, precision and recall on this set are known. Thus, another strategy is to weight the classifiers' decision scores according to their accuracy on the validation set.

- **Most-confidence strategy**: This strategy means that the decision of the classifier that is most confident is chosen. It is calculated as follows: If $vote_{min}$ and $vote_{max}$ are the lowest and highest score, respectively, of all classifiers, the final classification is $vote_{max}$ (or: 1), if $vote_{max} > (1 - vote_{min})$, and $vote_{min}$ (or: 0) otherwise.

- **Majority-driven strategy**: In this strategy, first the majority vote of all ensemble classifiers is considered. Then, the average of (only) the majority votes is used as the final result.

- **Average strategy**: This strategy simply uses the average of all ensemble classifiers' decision scores as the final result.

### 5.3.4 Validation Set for Subclass Models

The generated random forests are validated on a held-out set of the training data. Of course, it is reasonable to use a validation set that comes from the target domain. Two possibilities for creating a validation set are considered:

automatic and manual creation. In the proposed system the automatic creation of a validation set works as follows. From the multi-lingual queries that are used to find training images, the top-20 results are exploited to generate the data of the validation set. These top results are normally of high quality compared to subsequent retrieval results at lower ranks. Each random tree of each random forest is evaluated on this validation set using the f1-score, which is based on recall and precision [Datta et al. 2007].

### 5.3.5 Updating the Subclass Models

The updating component is essential for the quality of the system. It is responsible for utilizing new data in a way that improves the system's overall classification performance while keeping the number of models in a moderate range. The following state is assumed before the updating component comes into play: several clusters representing subclasses and a number of related random forest classifiers exist. Training images that have contributed to the clusters and subclasses have remained in the system, whereas all other (positive) training images have been discarded. Now, new training data are acquired according to the description in Section 5.3.1 and have to be included in the learning system. The impact of duplicate images is discussed in Section 5.3.7. Images are passed to the clustering process that currently consists of the clustering based on the old training data. Three cases with respect to the old training data can be distinguished:

1. *Stability of object assignments and clustering*: The new images are simply assigned to clusters without changing the number of clusters and the assignments of the old data.

2. *Changes in the assignment of objects*: It is possible that the assignment of old data changes, i.e., they are assigned to another cluster due to the newly added images.

3. *Changes in the clustering structure*: It is possible that new clusters emerge or existing clusters are merged.

In the first two cases, updating works as follows. The clusters of the previous clustering are mapped to clusters of the new clustering. If a cluster of the current generation corresponds to a preceding generation's cluster, the corresponding random forest is extended with random trees using all positive training images of this cluster. Finally, each tree of the forest is evaluated on the validation set. Only those trees are "planted" in the forest that achieve an f1-score on the validation set that is above the average of the f1-scores for all trees of this forest. In the third case, the number of clusters has changed after adding new training images. First, for all current clusters the best match (i.e., the corresponding cluster) in the preceding clustering result is estimated by counting the number of shared positive training images. If there is no best match for a cluster, a new random forest representing this cluster is generated.

### 5.3.6 Managing Scalability

One of the advantages of the random savanna approach is that it can deal with scalability in a natural way. Since the proposed system is expected to run endlessly in its final envisaged development stage, a component is needed that leverages only a subset of all positive training images. Otherwise, the computational complexity of the incremental training phase would increase from iteration to iteration and the training time could become unacceptable at some point. To retain only a fixed number of positive training images in the system, it is made use of the fact that the random trees in the forests are trained via bootstrapping. In the updating phase, each tree is evaluated on the validation set, and all trees can be sorted according to the obtained f1-score. Due to bootstrapping, each image has contributed to none, one or several trees in a random forest. Furthermore, many random trees have been discarded in the updating stage, because their performance on the validation set was not good enough, hereinafter called "bad" trees. For each training image, the number of bad trees for which it has been used as a positive training image is counted, obtaining a "badness score" B. Then, all positive training images are sorted in ascending order according to this score; only the top-k images (e.g., k=400) remain in the system for subsequent training iterations.

It should be pointed out why the contribution to "bad" trees is counted and not the contribution to "good" trees. Assume that a score G is used that counts the number of good trees for which an image has served as a training image. The fact that an image has a low score in this respect does not necessarily imply that it is a bad training image. It is possible that it has simply been chosen less often or even never for training due to the bootstrapping procedure. Another possibility is to normalize the score for an image with respect to the frequency of how often it has been selected to train a tree. But even in this case it is possible that an actually very good training image has been chosen only once for one "bad" tree and its quality would be severely underestimated. In contrast, if an image has been chosen often and it often contributed to "bad" trees, the conclusion that this image is not well suited for training is more reliable.

### 5.3.7 System Impact of Duplicate Images

When training images are automatically acquired from the WWW, it is possible to download the same image several times from multiple sites. Such duplicates are considered implicitly by the system in the different processing steps. First, the duplicates that have different textual information might be partially filtered out as spam. Since the system relies on random forests and hence on bootstrapping of

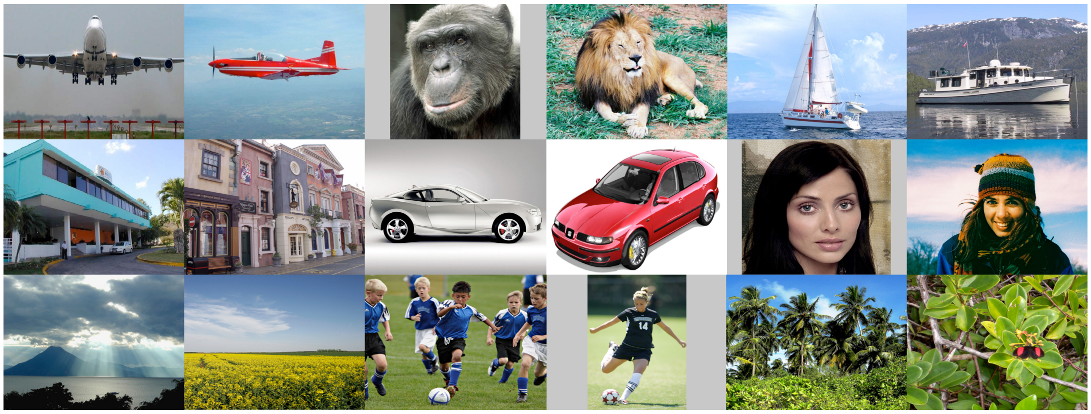


Figure 5.6: Example images of the test sets TS1 and TS2.

training samples, the probability that two or more images that are equal are used for learning is reduced. Finally, the scalability manager rates the usefulness of training images. If one or more of the duplicate images were identified to be less useful for learning, they will be discarded from the system. Overall, the fraction of duplicates is rather low compared to the total number of downloaded images. Our experience is that the impact of duplicates on the learning system is rather limited. In the test sets, duplicates have been filtered out manually in order to obtain unbiased results.

## 5.4 Experimental Results

The impact of all components of the proposed system for web-supervised learning has been evaluated on three different test sets. The first test set (TS1) consists of 400 web images, while the second test set consists of 1,000 web images (TS2). 9 concepts have been chosen to test the learning capabilities of the system: “airplane”, “animal”, “boat”, “building”, “car”, “face”, “sky”, “soccer”, and “vegetation”. Examples for both test sets are presented in Figure 5.6. It can be noted that the number of test images per concept in TS2 is considerably larger than in the often used experimental *Caltech* setting [Fei-Fei et al. 2007; Griffin et al. 2007], where only 20 or 30 test images are suggested and images within a class have a homogeneous appearance. Furthermore, results are reported for five concepts (“airplane”, “boat”, “car”, “chair”, “motorbike”) from the *VOC* 2007 test set that consists of 4,952 images [Everingham et al. 2010].

Training images were collected from the WWW as described in Section 5.3.1. The codebook generation is based on salient point SIFT descriptors, and a codebook with 1,000 visual words was generated. The images belong to the nine concept classes, and between 20 and 69 examples per class are used for building the codebook. In the *VOC* experiments, densely sampled SIFT features were

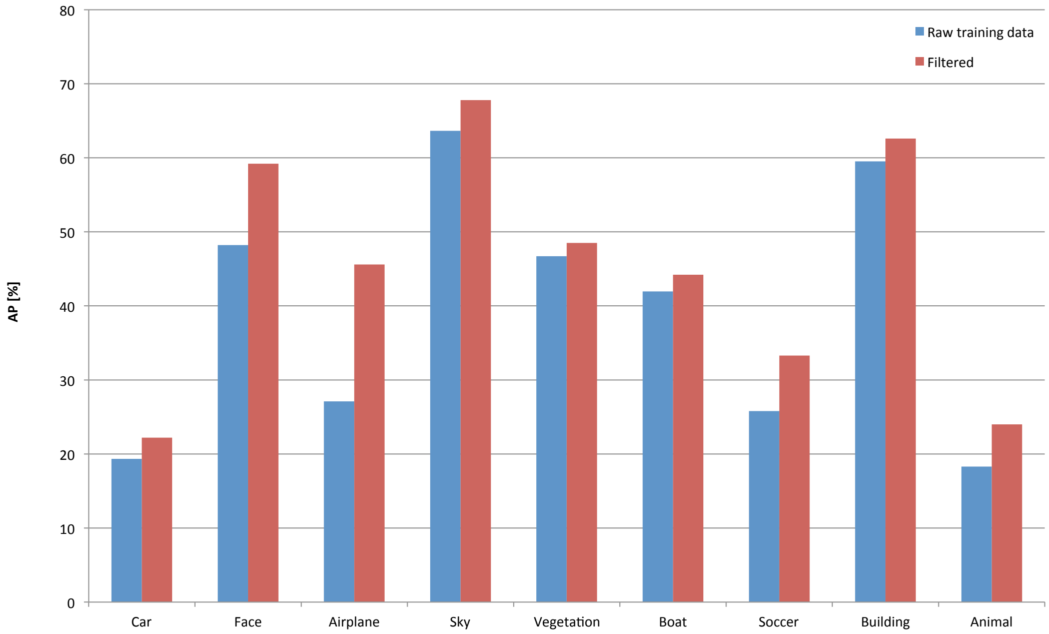


Figure 5.7: Web-supervised learning results for nine concepts with and without spam filtering based on visual cues.

used in conjunction with a codebook size of 4,000 visual words. The HoVW representations were constructed using the soft-weighting scheme as described in Section 3.2.2. The implementations of the K-means, the SVM, and the random forest algorithm are used from the WEKA collection of machine learning algorithms [Witten and Frank 2005]. The SVM algorithm uses the RBF-kernel. Furthermore, a parameter search has been conducted for the SVM parameters $C$ and $\gamma$.

For each concept query, the system returns a list of ranked images. To evaluate a retrieval result, the quality measure of AP is used. The default combination strategy for the classifiers in the ensemble is the majority-driven strategy. The individual experiments are described in more detail below.

**Experiment A (Spam detection capability)**: In the first two experiments, the impact of the spam detection component based on textual or visual features has been tested. Regarding textual features, the effect of spam detection for the concept "face" has been investigated exemplarily and 1,000 training images have been downloaded for it. Two different SVMs were build: a baseline SVM trained on all positive training images and a "spamless" SVM trained without images from the spam clusters. All test images were ranked according to the confidence scores of the SVM concept classifiers yielding two different rankings

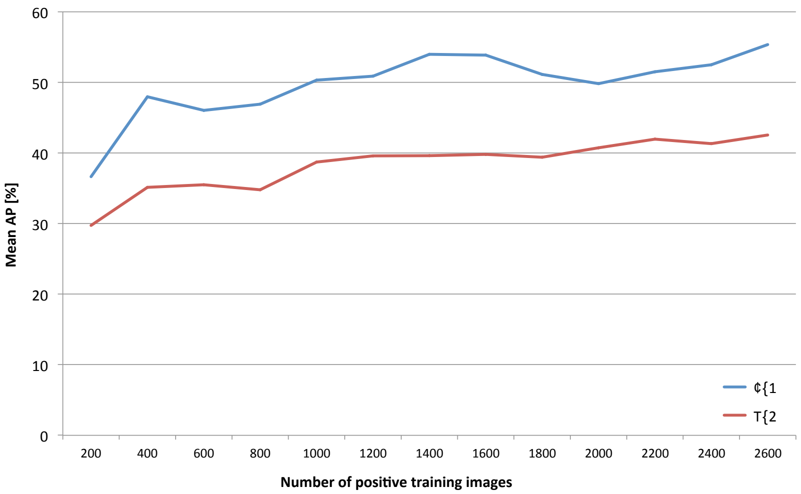


Figure 5.8: Web-supervised learning results for nine concepts on the test sets TS1 and TS2.

for the baseline and the "spamless" SVM. If all positive training images were used, a performance of 64.5% AP was achieved. Using the spam filter, the result was improved by 10% to 74.5% AP.

To test spam detection based on visual cues, 600 training images for each concept have been downloaded. The clustering of positive training samples was conducted using the HoVW features. Again, the positive images of all remaining relevant clusters were used to train a "spamless" SVM and compared with the baseline SVM that was trained with all downloaded positive images. The results, presented in Figure 5.7, clearly show an improvement for all concepts when using spam detection.

**Experiment B (Incremental learning capability)**: This experiment investigates the incremental learning capability of the web-supervised random savanna system for the nine mentioned concepts. The system performance in terms of AP has been evaluated on the test sets TS1 and TS2 at different (simulated) time points using an increasing number of positive training samples. The system has been initialized with 200 positive training examples. Each time when 200 new images had been added, the updating process took place. The results are presented in Figure 5.8 and demonstrate that incremental learning improves the performance on the average. For the more difficult test set TS2, the mean AP of the initial system could be improved significantly from 29.7% to 42.5% (2,600 training images).

**Experiment C (Comparison with random forest and SVM)**: In this experiment, the proposed random savanna system has been compared with the

| **AP [%]** | **RF** | | **SVM** | | **RS** | |
|---|---|---|---|---|---|---|
| **Pos. images** | **200** | **1400** | **200** | **1400** | **200** | **1400** |
| Airplane | 32.4 | 34.8 | 29.9 | 30.8 | 20.7 | 36.4 |
| Animal | 12.5 | 13.9 | 16.4 | 17.7 | 12.9 | 23.3 |
| Boat | 21.8 | 21.4 | 22.0 | 28.7 | 22.3 | 32.2 |
| Building | 48.2 | 42.0 | 48.3 | 52.5 | 43.4 | 53.8 |
| Car | 14.9 | 32.1 | 36.3 | 36.0 | 28.2 | 38.6 |
| Face | 30.9 | 32.1 | 26.4 | 36.4 | 32.2 | 43.8 |
| Soccer | 8.2 | 12.8 | 21.2 | 21.2 | 8.2 | 19.6 |
| Sky | 57.0 | 45.4 | 75.5 | 63.2 | 53.6 | 50.7 |
| Vegetation | 38.2 | 26.9 | 32.9 | 28.6 | 45.9 | 59.3 |
| Mean | 29.3 | 29.0 | 34.3 | 35.0 | 29.7 | 39.6 |

Table 5.1: Performance comparison on TS2 using random forest (RF), SVM and random savanna (RS) classifiers.

random forest and the SVM approach. For this purpose, a random forest and a SVM were trained using either 200 or 1,400 positive training images per concept. Both systems were trained on non-spam images only, but without using the clustering result of positive training data, i.e., there is only one classifier per concept and not an ensemble (such as the random savanna). The results for this experiment based on TS2 are presented in Table 5.1. It is shown that neither the random forest nor the SVM classifiers can benefit from the additional 1,200 positive training images (RF 1400 and SVM 1400) compared to the initial systems (RF 200 and SVM 200), whereas the random savanna improves the results significantly when additional training images are used (RS 1400) and outperforms the SVM 1400 system in terms of mean AP by 4.6% (relative improvement of 13%). The RS 1400 system can improve its initial system (RS 200) for all concepts except for “sky”; overall, it improves the initial performance by 10% mean AP corresponding to a relative improvement of 33%.

**Experiment D (Performance decrease with more images)**: The RF 1400 and SVM 1400 system have suffered from noticeable performance degradation for some concepts compared to the corresponding initial system. For the random forest system it has been tested whether this loss is due to the fact that many of the additional positive training images are either noisy or from another domain than the initial positive training images. To investigate this question, the positive training sets consisting of 1,400 images have been manually filtered for the concepts whose performance declined for the random forest approach on TS1 (“airplane”, “building”, “sky”, and “vegetation”). All noisy images as well as concept images that do not belong to the same domain have been removed. Then, another random forest was trained on this new training set and compared to the RF 200 and the RF 1400 system. The results in Table 5.2 demonstrate that the bad results for the RF 1400 system are indeed caused by noisy training images. Compared to the baseline RF 200 system, the results are slightly better for

| **AP** [%] | **RF 200** | **RF 1400** | **RF 1400 (manual)** |
|---|---|---|---|
| Airplane | 47.0 | 36.3 | 48.1 |
| Building | 68.3 | 41.5 | 62.5 |
| Sky | 60.3 | 44.4 | 61.8 |
| Vegetation | 50.6 | 43.7 | 52.2 |
| Mean | 56.6 | 41.5 | 56.2 |

Table 5.2: Impact of noisy positive training images (TS1).

the concepts “airplane”, “sky”, and “vegetation”, except for the concept “building”. This result shows that the proposed random savanna system successfully deals with the issue of noisy training data by clustering the training images and removing inadequate “spam” clusters, in contrast to the random forest system.

**Experiment E (Scalability)**: In this experiment, the scalability of the proposed updating strategy for the random savanna system is analyzed by testing a maximum limit of 400 positive training images per concept. For this purpose, the performance of the system using the scalability component is compared with the baseline random savanna system using TS2. The results in Table 5.3 show that the scalable version is quite effective in removing unnecessary positive training images and achieves a comparable mean AP for the nine concepts. The removal of bad training images works well and is an important component in order to be able to employ additional positive training images from the WWW without any noticeable additional computation efforts (except for the scalability manager itself, of course).

**Experiment F (Other test set: *VOC*)**: Furthermore, the incremental learning capability in conjunction with the scalability manager for five concepts of the *VOC* test set (TS-VOC) has been investigated. In this experiment, images from the *VOC* training set have been used as validation set in the random savanna system. The results displayed in Figure 5.9 demonstrate that incremental learning of the random savanna using web images improves the performance on the average. The best AP of 33.8% is achieved for 2,800 training samples, the average of the last five measurements (2,200 - 3,000 training samples) is 32.4%. In other words, the system using web training images obtains an AP that is more than 4 times better than a random retrieval system (the average frequency of

| **Mean AP** [%] | **RS 200** | **RS 1400** | **RS 2000** | **RS max** |
|---|---|---|---|---|
| No scalability | 29.7 | 39.6 | 40.7 | 42.6 |
| With scalability | 28.8 | 37.7 | 39.7 | 40.3 |

Table 5.3: Impact of the scalability manager which retains a limited number of 400 positive images.

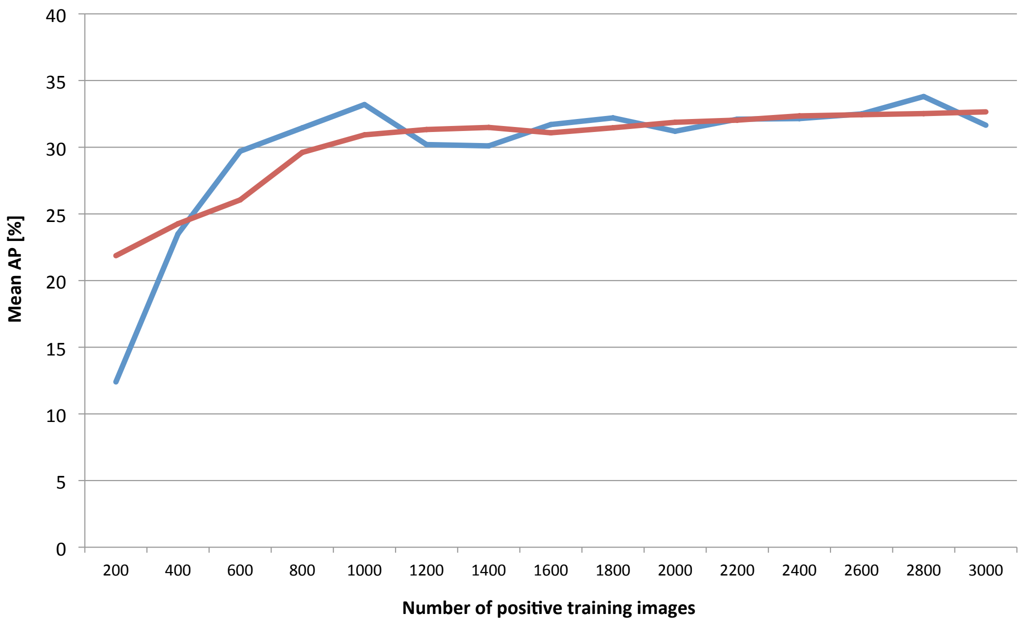


Figure 5.9: Web-supervised learning results for five concepts of the *VOC* test set (TS-VOC) based a varying number of positive training images (blue), averaged for five subsequent points of measurement (red).

the concepts is 7% in the TS-VOC test set). A SVM with an RBF kernel has also been trained and tested for 3,000 positive training samples on the TS-VOC test set. When textual and visual spam filtering of the random savanna system was not used to clean the training data, the SVM achieved an AP of 26.9%; the SVM using spam filtering achieved an AP of 31.4%. The random savanna system slightly outperforms the SVM approach by an absolute performance improvement of 2.4% mean AP, which is not as clear as on the test set TS2 in experiment C. Nevertheless, the usefulness of spam filtering is demonstrated by this experiment. In addition, the random savanna system offers the advantages that it is extensible due to its incremental learning capability, and it is easily parallelizable at the tree and forest level.

**Experiment G (Strategies for classifier combination)**: In this experiment, five different strategies for classifier combination as described in Section 5.3.3 have been investigated: the most-confident strategy, the majority-driven strategy, the idealistic strategy, the validation-driven strategy, and the average strategy. The results presented in Table 5.4 show that the most-confident and the majority-driven strategy achieve the best results, while the validation-driven strategy and the simple averaging strategy degrade performance by 1.0% and 1.7% mean AP, respectively. Interestingly, the idealistic strategy performed only slightly worse than the two best strategies.

| Ensemble Strategy | RS 3000 |
|---|---|
| Most-confident | 31.8% |
| Majority-driven | 31.4% |
| Idealistic | 31.1% |
| Validation-driven | 30.8% |
| Average strategy | 30.1% |

Table 5.4: Performance comparison of different ensemble strategies in terms of mean AP.

**Experiment H (Alternative clustering)**: This experiment examined the impact of another clustering algorithm for arranging the positive training data in clusters (subclasses) using visual features. Therefore, the EM clustering algorithm has been applied with a maximum number of ten iterations. Textual spam filtering was still performed using the K-means algorithm. The experimental results are presented in Figure 5.10. When using the EM clustering algorithm, similar results were achieved. However, the results are slightly worse in comparison to the results based on the K-means clustering algorithm and do not exceed a mean AP of 30%.

**Comparison with related work**: An overview of experimental results and settings of related approaches is presented in Table 5.5. Only approaches that have been tested on image data are considered. The table shows that the experimental settings are very different, image training and test data stem from different domains, different evaluation measures and scenarios have been used, and the number of investigated concepts ranges from 5 to more than 100. It is obvious that the annotation and retrieval quality of these approaches is hard to compare. This is also due to the fact that the approaches emphasize different aspects of image acquisition or annotation. The goal of Fergus et al. [2005], Li and Fei-Fei [2009], and Schroff et al. [2011] is to automatically build a large image database for the purpose of learning visual object categories by exploiting web data. Gong et al. [2009] and Xu et al. [2009] presented image annotation systems that correlate visual and textual information from the WWW. Xu et al. [2009] analyzed keyword correlations from Flickr and annotated images using a keyword correlation matrix. Wang et al. [2006] presented a similar annotation system that exploits information of semantically and visually similar search results. However, it requires the user to provide a keyword for an image to be annotated. This is also the case for another approach of Wang et al. [2008c]. Torralba et al. [2008] demonstrated that image classification can be realized by using a very large number of tiny (resized) images from the WWW. Ulges et al. [2011] concentrated on exploiting the context information in Flickr's image database. Tang et al. [2009] presented a semi-supervised approach for image annotation that is evaluated in the Flickr domain. Finally, Liu et al. [2009] and Datta et al. [2007] incorporated relevance feedback of users in their system and included annotations that

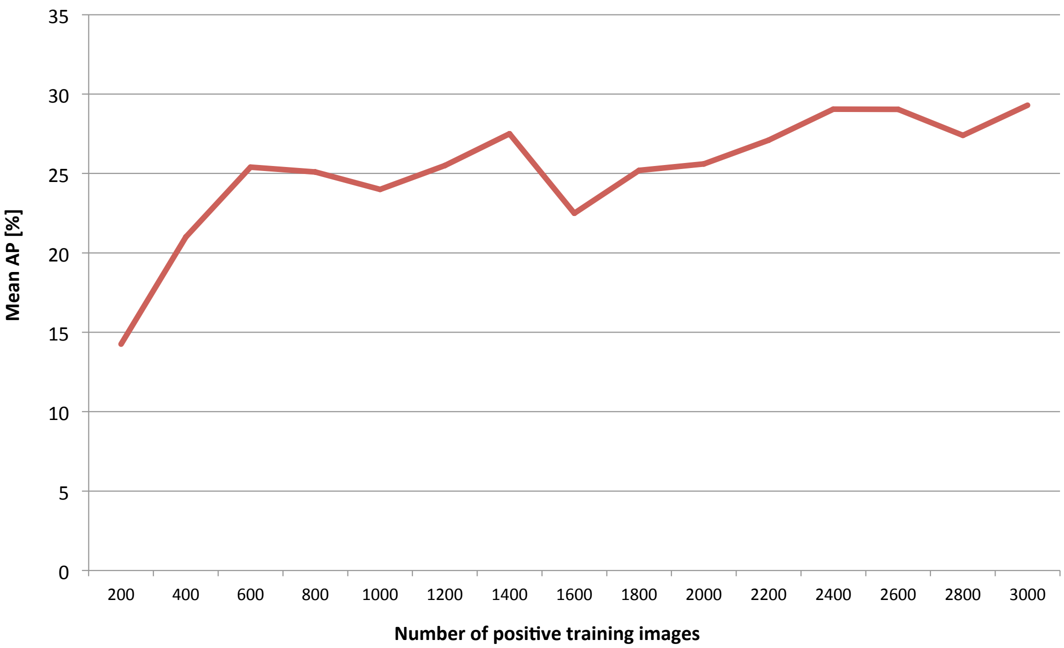


Figure 5.10: Results for five concepts on the test set TS-VOC using the EM algorithm for clustering positive training images.

changed over time, respectively. As described in Section 5.2, only few proposals address the task of incremental web-supervised learning of (heterogeneous) visual concepts. Datta et al. [2007] focused on image sharing sites with collaborative tagging. In contrast to our approach, they exploited user feedback during the learning process, whereas the proposed system works in a general manner and exploits information from arbitrary web sources. Zhang et al. [2009] proposed an incremental learning extension for SVMs, but it improves the baseline SVM only with respect to training time but not with respect to accuracy, in contrast to our approach. The OPTIMOL system [Li and Fei-Fei 2009] concentrates on automatically collecting a large image data set using a number of seed images. Due to these seed images, it is less suited to model heterogeneous image classes. Li and Fei-Fei [2009] conducted only a small classification experiment on seven *Caltech* classes. They did not use the learned models directly for classification, but trained a multi-class SVM using model scores for Caltech training data as features. Hence, this experiment is not directly comparable to our experiments.

The advantages of the random savanna approach can be summarized as follows: The system works without any explicit manual supervision and its learning process is based only on image data acquired from the WWW. In contrast to some other approaches, it is not targeted at a special domain. The proposed system is aimed at modeling heterogeneous image classes by modeling subclasses, and it does not use any seed training data from a target domain. Moreover, an

| Approach | Image Domain (Train/Test) | #Train Images | #Test Images | Mean AP | Mean Precision@ | Mean F1 | Accuracy | Remarks |
|---|---|---|---|---|---|---|---|---|
| Datta et al. [2007] | Corel/Corel | 26,000 | 10,000 | - | - | 0.49 | - | Static |
| Datta et al. [2007] | Alipr/Alipr | 20,000 | 16,000 | - | - | 0.43 | - | Static |
| Datta et al. [2007] | Alipr/Alipr | 1,000-10,000 | 10,000 | - | - | 0.45 | - | Incr. learn.: rel. feedback |
| Fergus et al. [2005] | Web/Web | 4,091 | 4,091 | - | 0.68 | 0.24 | - | @15% recall, 7 concepts |
| Fergus et al. [2005] | Web/Caltech | 4,091 | 2,148 | - | - | - | 0.72 | |
| Gong et al. [2009] | Web/Web | n/a | 200 | - | 0.22 | 0.1 | - | @position 1 |
| Gong et al. [2009] | Web/Web | n/a | 200 | - | 0.22 | 0.31 | - | @position 10 |
| Li and Fei-Fei [2009] | Web/Caltech | $\sim 2.4 \cdot 10^4$-$10^5$ | 2,148 | - | - | - | 0.75 | |
| Liu et al. [2009] | Web/Corel | 1,300,00 | 4,999 | - | 0.33 | - | - | @top-70, rel. feedback |
| Schroff et al. [2011] | Web/Web | 28,323 | 28,323 | - | 0.73 | 0.25 | - | @15% recall, 5 concepts |
| Schroff et al. [2011] | Web/Web | 28,323 | 28,323 | - | 0.69 | - | - | @15% recall, 18 concepts |
| Tang et al. [2009] | Flickr/Flickr | 27,807 | 27,808 | 0.16 | - | - | - | |
| Torralba et al. [2008] | Web/Web | $\sim 78,000,000$ | 1,125 | - | - | - | 0.79* | *Area ROC/curve |
| Ulges et al. [2011] | Flickr/Flickr | <8,000* | <8,000* | 0.22 | - | - | - | *Data split for train/test |
| Ulges et al. [2011] | Corel/Corel | <5,000* | <5,000* | 0.36 | - | - | - | *Data split for train/test |
| Wang et al. [2006] | Web/UWashgt. | 2,400,000 | 1,109 | - | - | 0.07 | - | |
| Wang et al. [2008c] | Web/UWashgt. | 2,400,000 | 1,109 | - | - | 0.29 | - | 100 queries+keyword |
| Wang et al. [2008c] | Web/UWashgt. | 2,400,000 | 1,109 | - | - | 0.21 | - | 50 queries+keyword |
| Wang et al. [2008c] | Web/Corel | 4,500 | 500 | - | - | 0.43 | - | Without keyword |
| Xu et al. [2009] | Web/Web | 1,000 | 4,000 | - | - | 0.29 | - | |
| Proposed method | Web/Web | 200-3,000 | 1,000 | 0.43 | - | - | - | Pos. train. imgs per concept |
| Proposed method | Web/VOC'07 | 200-3,000 | 4,952 | 0.34 | - | - | - | Pos. train. imgs per concept |

Table 5.5: Survey of experimental results of related approaches (best results are given for each approach).

incremental learning capability is realized that improves the system performance over time. Finally, the system can be parallelized easily and scales well: only a small and constant number of good training images has to be maintained during the learning process.

## 5.5 Summary

In this chapter, a novel purely web-supervised system for long-term learning of visual concepts that uses training data from the WWW has been presented. Only a single word describing the visual concept is required to initiate the learning process; it does not require any further human supervision. The system continuously updates its learned models while at the same time it deals with scalability by retaining only a small number of training images. Experimental results demonstrate the advantages of the proposed system, and in particular, the four learning capabilities of the system: 1.) multi-modal clustering is effective in removing spam images; 2.) the system is able to incrementally improve its performance without any (manual) supervision; 3.) the random savanna ensemble that realizes subclass learning is superior compared to random forests and SVMs; 4.) the scalability manager is very effective by means of keeping only a limited number of positive training images in the system. Another advantage is that the random savanna approach is easily parallelizable in the training and testing phases at the tree or forest level.

*"A good video can make all the difference."*

Brian Harold May

# 6 Applications in Psychology and Media Sciences

## 6.1 Introduction

In recent years, several technological innovations have fostered an enormous increase of multimedia data, including larger hard disk capacities, processor power, and network bandwidth, accompanied by the improvement of audio, image, and video compression technologies, and the proliferation of digital photo and video cameras. In the field of media sciences the annotation of images and videos is an ever-reoccuring, time-consuming, manual task. It has to be carried out to provide a basis for scientific film studies. The current chapter deals with video content analysis to support interdisciplinary research in the field of psychology and media sciences.

First, an automatic computer games video content analysis system was built to support interdisciplinary research efforts in the field of psychology and behavioral sciences. The psychological research question studied was whether and how playing violent content in computer games may induce aggression. Therefore, novel semantic concepts, most notably "violence", are detected in computer game videos to gain insights into the interrelationship of violent game events and the underlying neurophysiologic basis (brain activity) of a player. Beforehand, neither video content analysis methods have been applied to computer game recordings nor automatic video content analysis has been suggested for the field of behavioral sciences.

Second, the video retrieval component of the software toolkit *Videana* is presented. The integrated video content analysis system supports the scholarly analysis of audio-visual material and relieves media scholars from the task of annotating films manually.

The remainder of this chapter is organized as follows: Section 6.2 describes interdisciplinary research efforts in the field of psychology and behavioral sciences. A computer games video content analysis system has been built to automatically detect violent game events. In Section 6.3, the software toolkit `Videana` including the concept retrieval component is presented. Section 6.4 summarizes the chapter.

Parts of this chapter have been published by Mühling et al. [2007a] and Ewerth et al. [2007b].

## 6.2 Violence Detection in Computer Games

Computer games play a very important role in today's entertainment media and belong to the most popular entertainment products. Unfortunately, the number of computer games containing serious violence increases. There is an extensive ongoing debate about the question whether playing violent games causes aggressive cognition, aggressive affects or aggressive behavior, in particular with respect to teens and young adults. The neurophysiologic perspective of mass communication research concentrates on emotional responses to video game playing. Mathiak and Weber [2006] developed neurophysiologically grounded measures for the "human experience of media enjoyment". The study continues their prior work on video game playing in which functional magnetic resonance imaging (fMRI) scans were taken during video game playing [Weber et al. 2006]. From this neurophysiologic perspective, they demonstrated that a specific neurological mechanism is activated when playing a first-person-shooter game. One central finding is that cognitive areas seem to suppress affective areas during the (virtually) violent interactions. This mechanism helps to better understand a potential link between playing certain types of violent video games and aggressive cognition and affects. The experimental design presented by Weber et al. [2006] is based on the definition of certain game states and captures a player's brain activity via fMRI while (s)he is playing a violent computer game. Several semantic game events are distinguished: 1.) inactive, 2.) preparation, 3.) search and explore, 4.) danger, 5.) under attack, and 6.) fighting and killing. Once the game recordings are annotated with these semantic categories, the interrelationship of violent game events and the underlying neurophysiologic basis (brain activity) of the player can be investigated. Normally, human annotators are required to index such game content according to the current game state, but this is a very time-consuming task. In this context, computer-based automatic video content analysis of computer game recordings promises several advantages: Human annotation efforts can be reduced noticeably, and the annotation process is speeded up and is based on reproducible and objective criteria only. At the same time, researchers are enabled to investigate an increased number of computer game videos to gather more experimental data.

In the following, a semantic video analysis system is presented that supports the experimental design described above by automatically identifying the game states (i.e., categories). The system is aimed at minimizing the human annotation effort and thus requires manual annotations for a single video only. The content analysis relies on audio-visual low-level features as well as on mid-level features. The considered mid-level features are the results of shot boundary detection [Ewerth and Freisleben 2004], camera motion estimation [Ewerth et al. 2004], audio type classification [Stadelmann 2010], text detection [Gllavata and Ewerth 2004] and face detection [Viola and Jones 2004]. For each game category, a SVM is trained using the low-level and mid-level features. Only a single video sequence with a duration of twelve minutes is required to provide training data and hence, human annotation effort is kept at a minimum. Afterwards, new videos are automatically analyzed using these SVM models. Finally, the graphical user interface (GUI) of the software system `Videana` (Section 6.3) allows a human expert to refine or correct the annotation results, if needed. Experimental results demonstrate the excellent performance of the proposed approach.

This section is organized as follows: Related work concerning semantic analysis of videos for certain genres is discussed in Section 6.2.1. In Section 6.2.2, the concepts (semantic game categories) which must be recognized for the experiment conducted by Weber et al. [2006] are listed and explained. Section 6.2.3 describes the experimental design and the main processing steps of the proposed semantic video analysis system. Experimental results are presented in Section 6.2.4. Section 6.2.5 discusses the results.

### 6.2.1 Related Work

Beforehand, neither video content analysis methods have been applied to computer game recordings nor automatic video content analysis has been suggested for the field of behavioral sciences. Nevertheless, semantic video analysis systems exist which are specialized for a certain genre, e.g., sports or news videos. There are many approaches addressing the analysis of news videos. This emphasis might have been enforced by the *TRECVid* evaluation series Smeaton et al. [2006] in which comprehensive news video test collections have been provided and used for evaluation purposes. A summary of semantic concept detection approaches regarding news videos was presented by Naphade and Smith [2004].

In a way, sports videos can be considered as somewhat related to the genre of computer games: Since both genres are rule-driven, the amount of possibly appearing content is limited in both sports and computer games ("e-sport"). The automatic indexing of sports videos has been extensively studied in recent years. As noted by Sadlier and O'Connor [2005], many specific approaches exist for several sports domains, e.g., Formula-1, cricket, tennis, American football, or soccer. Apart from specific approaches, frameworks have been proposed that cover more than only a single type of sports. For example, Xu and Chua [2006] proposed a framework for event detection in team sports videos that is based on audio-visual features, domain knowledge, and external information sources.

Tong et al. [2005] suggested a framework for semantic shot representation of sports videos. This framework is applicable to field sports, and shots are classified based on the camera distance, the displayed subject and the edited video layout.

Sadlier and O'Connor [2005] presented an event detection system for field sports as well. The following features were employed in a supervised learning process: image crowd detection, speech-band audio activity, on-screen graphics tracking, motion activity measure, field line orientation and some other features. Sadlier and O'Connor [2005] argued that it is not feasible to build a generic supervised event detection system for any kind of sports and they found the limitation to field sports reasonable.

### 6.2.2 Semantic Concept Classes for the Computer Game Experiment

Participants of the experiment conducted by Weber et al. [2006] played the "mature" rated first-person-shooter game "Tactical Ops: Assault on Terror" (`http://www.tactical-ops.de`). As mentioned above, the experiment was aimed at gaining insight into the interrelationship of playing violent computer games and changes in the consumer's brain activities. Therefore, several game states were defined, and the dependence of the players' brain activity is set in relation to these game states. Brain activity was measured via fMRI scans. A system is presented that is able to classify the following semantic classes with an acceptable high accuracy:

1. "inactive":
   The player's avatar (PA) is dead or the game has not started yet.

2. "preparation":
   The PA is buying equipment at the beginning of a new round.

3. "search/explore/danger":
   The PA explores the virtual world and searches for hostages, enemies and weapons.

4. "violence":
   The PA is fighting and/or injured.

In the original study, the semantic game categories were distinguished and annotated more sophistically (Figure 6.1). Category 3 was further divided into "search" and "potential danger", and for category "violence" it is distinguished

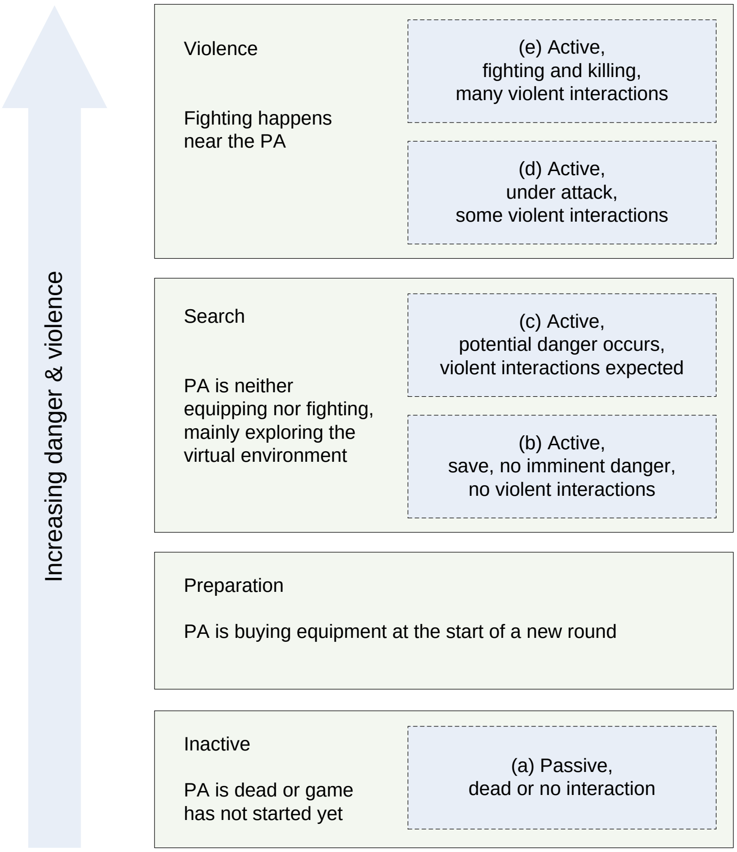
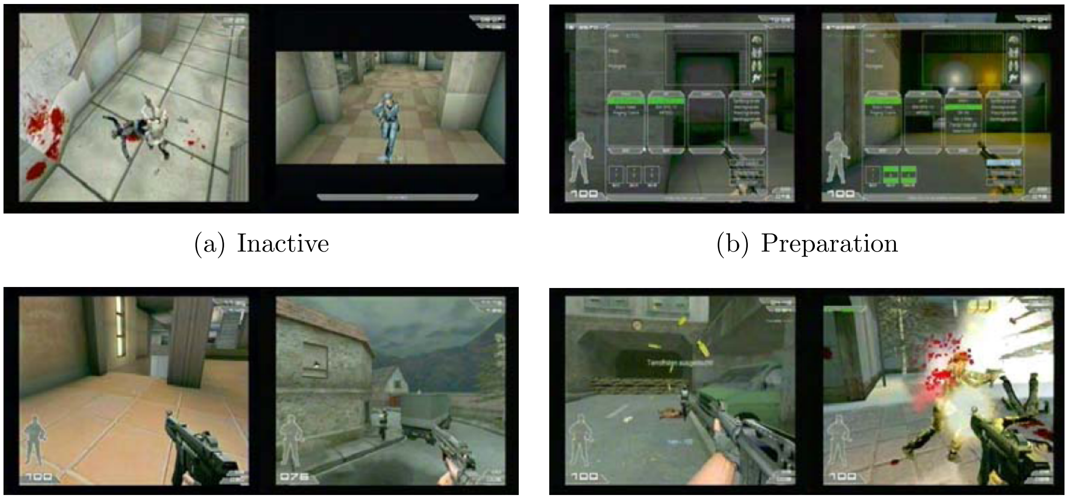

Increasing danger & violence
Violence
Fighting happens near the PA
(e) Active, fighting and killing, many violent interactions
(d) Active, under attack, some violent interactions
Search
PA is neither equipping nor fighting, mainly exploring the virtual environment
(c) Active, potential danger occurs, violent interactions expected
(b) Active, save, no imminent danger, no violent interactions
Preparation
PA is buying equipment at the start of a new round
Inactive
PA is dead or game has not started yet
(a) Passive, dead or no interaction

Figure 6.1: The four boxes explain the different semantic game classes used in this study and how they relate to the categories used by Mathiak and Weber [2006], which are displayed in the dashed nested boxes. The classes are ordered from bottom to top in terms of increasing violent content, where PA stands for "player's avatar".

whether the PA is injured/attacked or fighting actively. However, automatic distinction of these semantic classes would not be feasible without neglecting the target to have a generic video content analysis system. This can be for example considered regarding the highly abstract semantics concerning the distinction of the game states "search" and "danger". When the PA currently is in the state "search" (no imminent danger) and spots another character, its state switches to (potential) "danger". Now, according to whether this character is identified as an enemy or not, the state switches to "violence", because the PA shoots at the enemy, or back to "search" when the appearing character is harmless. Normally, the game state "danger" endures only for a few seconds before the state evolves further in the mentioned manner. Furthermore, the appearance of new characters in the PA's field of view often takes place near the horizon, where avatars are



(a) Inactive

(b) Preparation

(c) Search/Explore/Danger

(d) Violence

Figure 6.2: Visual impressions of the semantic game categories.

only a few pixels in size, and it is extremely difficult to perform the necessary friend-or-foe identification with a reasonable precision. Furthermore, the automatic system does not distinguish between "active" and "passive" violence. In practice, "passive" violence is a very short segment before either "active" violence or "inactive" (player's avatar is dead) take place. This is the reason for the definition of the four classes described above. In this way, an automatic and generic annotation system is feasible and the remaining manual revisions are minimized. Figure 6.2 shows example frames for each of the four semantic game categories.

### 6.2.3 Semantic Analysis of Computer Game Videos

In this section, a system to support interdisciplinary research in the field of psychology and behavioral sciences via automatic video content analysis is presented. It utilizes automatically extracted audio-visual low-level and mid-level features to infer about the semantic game classes via supervised learning. Two main targets have been pursued: First, the system is supposed to remain a generic video content indexing system and thus does not contain any specific content detectors (restricting its applicability to a certain computer game would offer a lot of tuning possibilities). Second, the annotation effort that is needed to apply a machine learning approach should be kept at a minimum, i.e., allowing the system to use a single labeled training video only.

The main processing steps of the system are discussed in more detail in the following sections: the audio-visual feature extraction and the classification of the semantic game states.

### Feature Extraction

The semantic content of computer games is present in all modalities of the recordings: fighting and killing, for example, is visible in the video domain by the presence of enemies, muzzle flash and blood; it is also audible in the accompanying soundtrack by means of shoots or explosive sounds as well as moans.

For each video frame the automatic content analysis system extracts several visual and audio features. In addition to color moments and texture features, several mid-level features were automatically extracted by utilizing camera motion estimation [Ewerth et al. 2004], audio type classification [Stadelmann 2010], face detection [Viola and Jones 2004] and text detection [Gllavata and Ewerth 2004]. These low-level features and the algorithms used to extract the mid-level features have already been described in Chapter 4. The camera motion estimation algorithm is used to compute confidence scores for the following camera motion types: translation (along the x-axis or y-axis), rotation (around the x-axis, y-axis, or z-axis), and zoom. The text detection approach is applied to derive the following features: the number of text elements, the distribution of text elements, and the text frame coverage. Furthermore, frontal faces are detected in each video frame and the number of detected faces and the face frame coverage are considered as additional feature values.

The camera motion features are useful to recognize the game state of searching and exploring, whereas text detection and texture features help to recognize the preparation state. A player steps into the preparation state with the intention to maintain his/her equipment. This screen contains several menus and is characterized by a high proportion of overlaid text. Thus, text features are assumed to be a very good criterion to detect preparation states. However, text detection in game videos is a challenging task, when the text is printed on complex background and the frames include many MPEG artifacts. For the first-person-shooter game “Tactical Ops: Assault on Terror” color moment features could be useful to detect the state “inactive” because of the mostly appearing black areas at the top and bottom of the screen.

Last but not least, the automatic content analysis system extracts a number of general audio low-level features which support the recognition of the semantic game states. The following features are extracted from non-overlapping 25 ms frames: 8th-order MFCCs to capture the broad envelope of the spectrum, zero crossing rate to measure the oscillation and intra-frame variation, short time energy that corresponds to loudness, sub-band energy distribution describing the loudness ratio for four successive frequency bands, brightness and bandwidth, spectrum flux to capture the inter-frame spectral variation, band periodicity, and noise frame ratio.

In addition, audio mid-level features are produced by feeding the audio low-level features into a content-based audio classification and segmentation system, as described in Section 4.6.3. This system produces mid-level features in form of

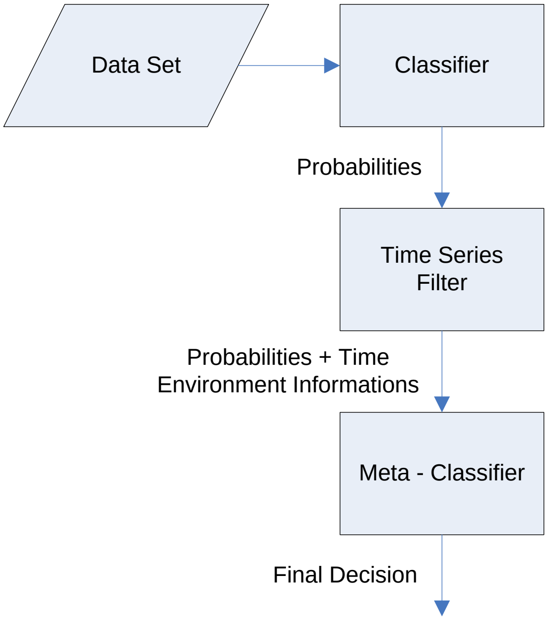


Figure 6.3: Concatenation of classifiers to employ temporal information.

acoustic class labels and related probabilities for silence, pure/non-pure speech, music, background, and action sounds.

### Semantic Classification

The goal of the proposed system is to learn concept models for the high-level semantic states of video games based on the extracted audio-visual low-level and mid-level features. As stated above, the system does not focus on special properties of the computer game under consideration ("Tactical Ops: Assault on Terror"). Instead of using a specific and narrow approach that only works for a single video game, a generic video content analysis system is built that can be easily adapted to other games or video genres.

The SVM algorithm is applied to learn the mapping between the extracted audio-visual features and the semantic game states using a RBF-kernel. An early fusion scheme is employed for the multi-modal analysis of the audio-visual low-level and mid-level features. The training of the SVM is realized by the SMO method [Platt 1999].

Since SVMs are binary classifiers, the baseline system uses several SVMs (one for each concept class) that have to be combined to solve the classification problem. To make a decision about the game state of a certain frame, the SVM models are employed to provide confidence scores for a test instance (frame). These scores are compared and the class with the highest score is chosen.

Besides the baseline system, two further classification strategies are investigated to predict the game states of the recordings.

It has been observed that the appearance of a certain class is reflected also by the probability scores which are assigned to neighbored frames by an initial SVM classifier. This is the motivation for the second strategy to classify the computer game content. In addition to the audio-visual features, some time series information is utilized. The basic idea of this strategy is to obtain information about the temporal neighborhood of a frame using the probability scores of the initial SVM classifier. Based on the classification results, the relative frequency of each class in the temporal environment is computed for the current frame. The relative frequency of class $c$ in the neighborhood of frame $k$ is calculated according to the following formula:

$$freq_c(instance_k) = \frac{1}{2w+1} \cdot \sum_{i=k-w}^{k+w} t_c(instance_i), \tag{6.1}$$

with

$$t_c(instance_i) = \begin{cases} 1 & \text{if frame } i \text{ is classified as class } c \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

where $w$ defines the window size. For example, if the relative frequency of violence is 0.5 for a frame, it follows that 50% of the neighboring frames are classified as violence. Furthermore, a smoothing filter is applied to the class probabilities obtained by the initial classifier. In both cases, a sliding window size of 25 frames is applied. The probability scores of the initial classifier, the frequencies and the smoothed values (four features each) are used as new features and then retrain another classifier that makes the final decision. The processing steps of the approach using temporal neighborhood information are displayed in Figure 6.3.

The third strategy uses semi-supervised learning to refine the classification results. In the setting of the addressed psychological experiment, the consumers always play the same game but at different levels and hence, they explore different virtual environments. Thus, it is possible that the SVM models learned from the training video are not suited well to distinguish between the different game classes in the test video. In order to achieve a more robust classification for a particular game video, a semi-supervised learning approach is proposed. A machine learning approach is called semi-supervised when unlabeled samples are incorporated in the training process. In our case, these are all frames of the test video since the class labels are unknown for them. The main processing steps of the semi-supervised learning approach are the following (Figure 6.4): First, the training video is used to build a classifier consisting of the initial game category models. The initial classifier is used to classify the instances (frames) of a test game video as described in the baseline system. Then, the instances are ranked separately for each game category based on the probabilities of the detected classes. These rankings are then used to choose the instances with the highest confidence for

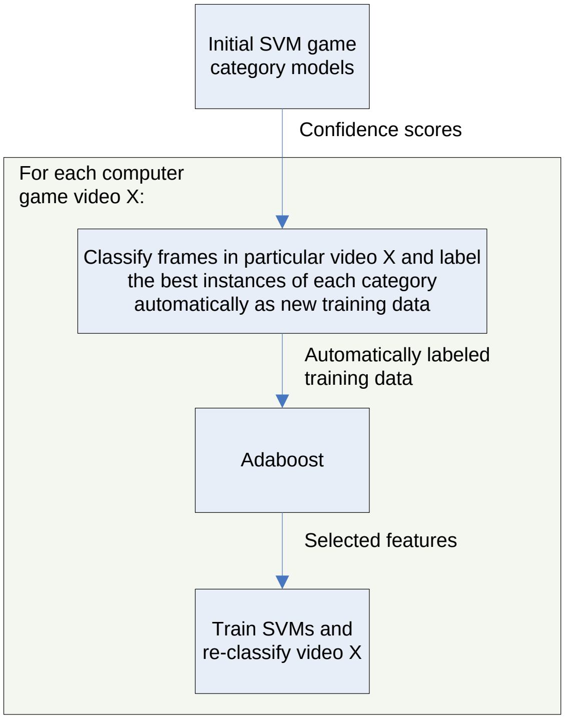

Initial SVM game category models
Confidence scores
For each computer game video X:
Classify frames in particular video X and label the best instances of each category automatically as new training data
Automatically labeled training data
Adaboost
Selected features
Train SVMs and re-classify video X

Figure 6.4: Main processing steps of the semi-supervised learning approach.

each class. The top 50% of each class of the automatically labeled instances are chosen as positive training samples. Based on these automatically labeled most relevant instances of the test video, relevant features are selected using Adaboost [Viola and Jones 2004]. The most relevant 77 features are chosen for subsequent use. An additional classifier is built using the previously chosen instances and selected features. Finally, this semi-supervised classifier, consisting of four newly trained SVMs, is used to classify the test video.

### 6.2.4 Experimental Results

In this section, several experiments are conducted to test the system's applicability for the psychological study. The main goal is to significantly reduce the human annotation effort while achieving an accuracy that is comparable to a manual annotation. In the original experimental setting, the human annotators needed 120 hours to label the entire video collection [Weber et al. 2006]. In addition, the goal was to keep the video content analysis approach generic.

Four computer game videos were used to evaluate the system performance. The computer game videos show a resolution of 352x288 pixels and a video frame rate of 25 frames per second. Table 6.1 presents the distribution of the semantic game categories for each of the used videos. The ground truth data were created by Weber et al. [2006].

| | **Inactive** | **Preparation** | **Search** | **Violence** | **Total** |
|---|---|---|---|---|---|
| game-vmj3_7 | 2,155 | 2,390 | 11,657 | 488 | 16,690 |
| game-vmj6_3 | 5,601 | 1,665 | 8,574 | 525 | 16,365 |
| game-vmj6_4 | 5,251 | 2,364 | 6,445 | 2,630 | 16,690 |
| game-vmj6_5 | 2,581 | 2,157 | 10,023 | 1,211 | 15,972 |

Table 6.1: Number of frames referring to semantic game categories for each of the used computer game videos.

| [%] | **Preparation** | **Search** | **Violence** | **Inactive** |
|---|---|---|---|---|
| Recall | 84.3 | 92.3 | 53.9 | 88.5 |
| Precision | 86.0 | 87.5 | 68.7 | 93.4 |
| F1 | 85.1 | 89.9 | 60.4 | 90.9 |

Table 6.2: "Baseline" system: Recall, precision and f1-measure for each of the four semantic classes.

| [%] | **Preparation** | **Search** | **Violence** | **Inactive** |
|---|---|---|---|---|
| Recall | 83.1 | 92.6 | 56.7 | 91.6 |
| Precision | 87.7 | 88.5 | 68.4 | 94.1 |
| F1 | 85.4 | 90.5 | **62.0** | 92.8 |

Table 6.3: "Baseline + Temporal Features": Recall, precision and f1-measure for each of the four semantic classes.

| [%] | **Preparation** | **Search** | **Violence** | **Inactive** |
|---|---|---|---|---|
| Recall | 92.2 | 94.9 | 55.3 | 92.0 |
| Precision | 96.0 | 90.0 | 66.0 | 97.6 |
| F1 | **94.1** | **92.4** | 60.2 | **94.8** |

Table 6.4: "Baseline + Semi-Supervised Learning": Recall, precision and f1-measure for each of the four semantic classes.

| | **Preparation (GT)** | **Search (GT)** | **Violence (GT)** | **Inactive (GT)** |
|---|---|---|---|---|
| Det. Preparation | 23,733 | 918 | 52 | 8 |
| Det. Search | 1,991 | 104,535 | 6,449 | 3,194 |
| Det. Violence | 0 | 3,623 | 8,046 | 520 |
| Det. Inactive | 4 | 1,021 | 15 | 43,042 |

Table 6.5: Confusion matrix for the semi-supervised learning experiment. For example, the most frequent error is that a "violence" frame is misclassified as "search" and vice versa.

| [%] | **Baseline** | **Temporal** | **Semi-Supervised** |
|---|---|---|---|
| Total recall | 87.5 | 88.5 | 91.0 |

Table 6.6: Total recall for each of the tested systems.

A “leave k-1 videos out” cross-validation scheme was applied for evaluation. Since the main goal is the reduction of the human annotation effort, only one video is used as training data in each round while the remaining three videos are used as test videos. The SVM implementation of the WEKA library [Witten and Frank 2005] together with a RBF-kernel has been used for training and classification. The recall and precision measure as well as the f1-score are used to evaluate the following three system variations:

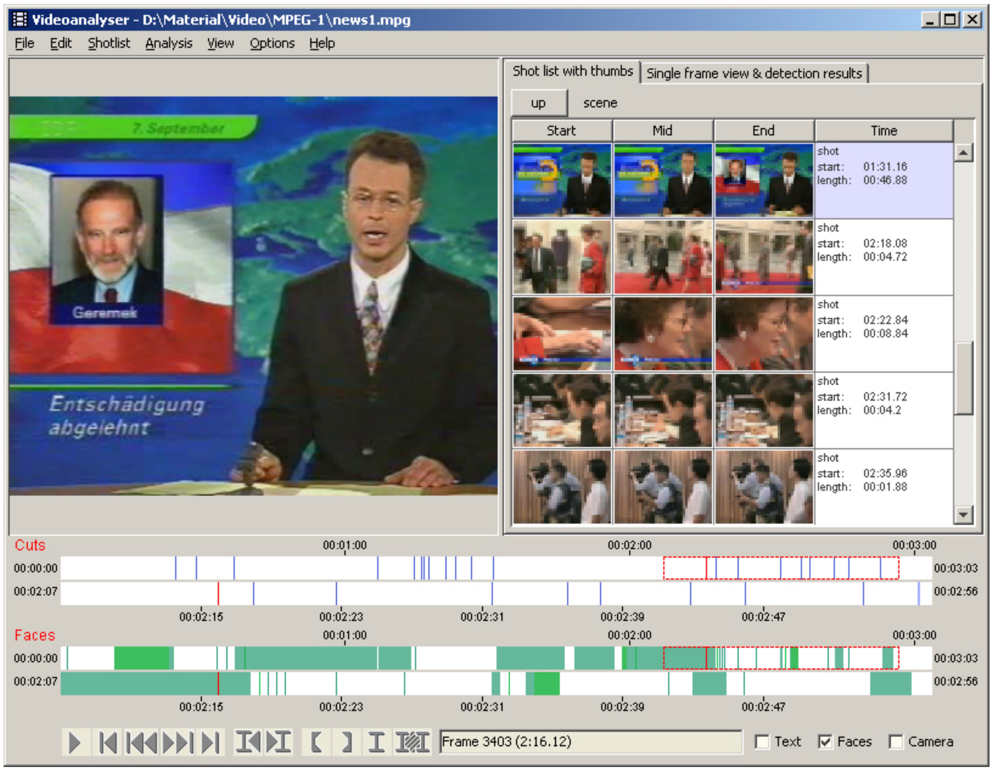
1. The baseline system
2. The system that captures temporal characteristics of the game states
3. The semi-supervised learning system

The results of these three experiments are presented in Table 6.2, 6.3, and 6.4. Several observations can be made. First, the automatic baseline system achieves a frame-based total recall of 87.5% on the average. This is a very good result considering that the inter-coder reliability in the original psychological experimental setting between the human annotators was 0.85 Cohen’s kappa [Weber et al. 2006]. In nearly any experiment “preparation”, “search”, and “inactive” states were recognized well, whereas the recognition of violent states was rather difficult. In terms of total recall, the semi-supervised approach gains the best results (Table 6.6). The approach using temporal neighborhood information achieved the best performance for the most difficult concept “violence” and recognizes more than half of the violent actions correctly while keeping the precision at nearly 70%. The confusion matrix in Table 6.5 allows gaining insight in the failures of the semi-supervised learning system. The diagonal represents the number of correctly classified frames. For example, the most frequent error is that a “violence” frame is misclassified as “search”, and vice versa, whereas, e.g., a “violence” frame was never classified as “preparation”.

Overall, the proposed system achieves a very satisfying performance demonstrating the ability to reduce human annotation efforts to a minimum. The system automatically determines relevant game events with high reliability.

### 6.2.5 Discussion

In this section, an automatic video analysis system has been presented that supports psychological experiments with respect to violence in computer games. In the addressed interdisciplinary study, annotations are required to find interrelationships between the consumer’s brain activity and the game events during the recorded game sessions, in particular with respect to violent actions. The proposed system automatically labels the videos and achieves a total recall of up to 91%. Considering the fact that Weber et al. [2006] observed an inter-coder reliability of 0.85 Cohen’s kappa for human annotators, our automatic system



| Time |
|---|
| shot<br>start: 01:31.16<br>length: 00:46.88 |
| shot<br>start: 02:18.08<br>length: 00:04.72 |
| shot<br>start: 02:22.84<br>length: 00:08.84 |
| shot<br>start: 02:31.72<br>length: 00:04.2 |
| shot<br>start: 02:35.96<br>length: 00:01.88 |

Frame 3403 (2:16.12)

□ Text ☑ Faces □ Camera

Figure 6.5: Screenshot of Videana's main window.

demonstrates an excellent performance. In addition, since this approach needs labeled training data for a single video only, the required human supervision in this interdisciplinary study could be kept at a minimum. The GUI of the software Videana enables a human expert to refine or correct the annotation results: As a basic requirement, the annotations must be as accurate as possible to investigate the interrelationship with a player's brain activity. However, such a correction step must also be applied when only human annotators label the videos. Overall, the experimental results demonstrate the applicability of our system for the interdisciplinary studies in the field of psychology and behavioral sciences.

## 6.3 Videana for Media Sciences

In the context of the interdisciplinary research project *Media Upheavals*, a comprehensive software toolkit, called Videana has been developed. This toolkit has been extended by a concept-based video retrieval component. Videana is aimed at providing computer assisted methods to support the scholarly analysis of audio-visual material. Its focus is on disburdening media scholars from typically very time-consuming manual annotation tasks. In the following, Videana and its video retrieval component are described in more detail.

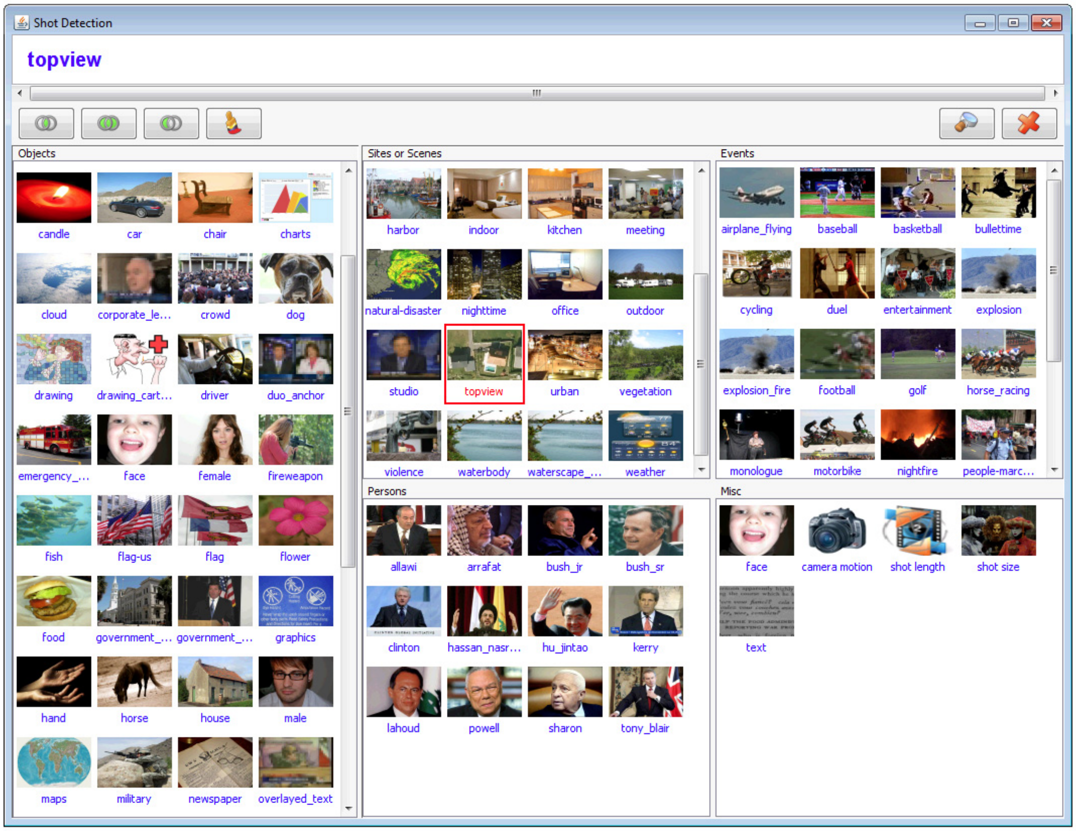
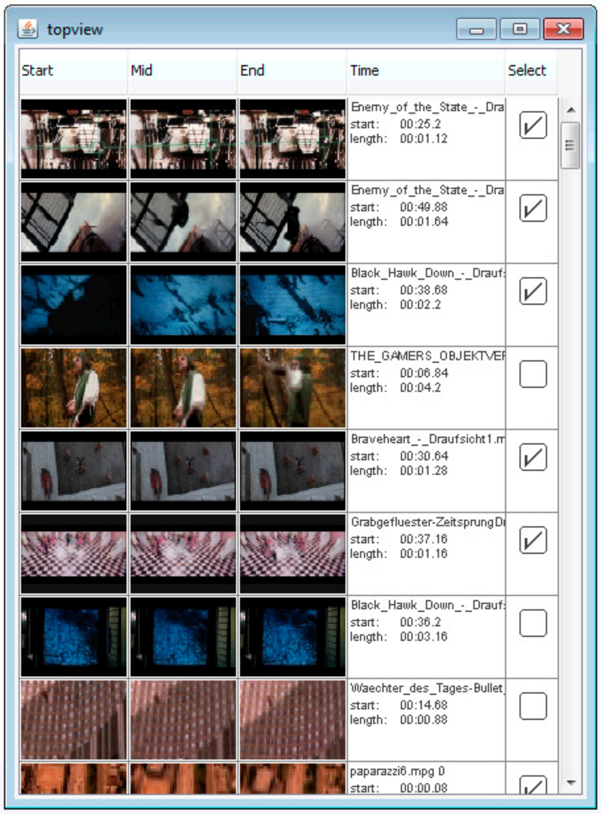


Figure 6.6: Screenshot of Videana's video retrieval GUI.

The main window of Videana is displayed in Figure 6.5. A video player is located on the left side of the window. The Videana GUI allows users to play back videos and to access particular video frames. Two timelines at the bottom of the main window visualize the analysis results for the temporal segmentation into video shots as well as for face detection results. Additional time lines appear as analysis results get available. The vertical lines in the "Cuts"-timeline represent abrupt shot changes, and the colored areas in the "Faces"-timeline mark the appearing face sequences in the video. On the right side, the temporal segmentation is presented in another way: shots are represented by three icons using the beginning, the middle, and the end frame of a shot. The related video frames are directly accessible by choosing the corresponding icon.

Videana supports the following tasks via automatic video content analysis: shot boundary detection [Ewerth and Freisleben 2004], text detection and recognition [Gllavata and Ewerth 2004], camera motion and shot size estimation [Ewerth et al. 2004], face detection [Viola and Jones 2004], dominant color extraction, audio segmentation and speaker clustering [Stadelmann 2010]. Based on a plug-in approach, any type of analysis algorithm can be easily updated, exchanged or removed. Additionally, the GUI allows to manually correct erroneous analysis



| Start | Mid | End | Time | Select |
|---|---|---|---|---|
| | | | Enemy_of_the_State_-_Dra<br>start: 00:25.2<br>length: 00:01.12 | ☑ |
| | | | Enemy_of_the_State_-_Dra<br>start: 00:49.88<br>length: 00:01.64 | ☑ |
| | | | Black_Hawk_Down_-_Drauf:<br>start: 00:38.68<br>length: 00:02.2 | ☑ |
| | | | THE_GAMERS_OBJEKTVEF<br>start: 00:06.84<br>length: 00:04.2 | □ |
| | | | Braveheart_-_Draufsicht1.m<br>start: 00:30.64<br>length: 00:01.28 | ☑ |
| | | | Grabgefluester-ZeitsprungDı<br>start: 00:37.16<br>length: 00:01.16 | ☑ |
| | | | Black_Hawk_Down_-_Drauf:<br>start: 00:36.2<br>length: 00:03.16 | □ |
| | | | Waechter_des_Tages-Bullet_<br>start: 00:14.68<br>length: 00:00.88 | □ |
| | | | paparazzi6.mpg 0<br>start: 00:00.08 | ☑ |

Figure 6.7: Screenshot of the video retrieval results for the concept “top view”.

results. The produced meta data is saved into XML files via the *Multimedia Data Description Interface* of the MPEG-7 standard [Martinez 2002] which facilitates interoperability with other multimedia applications.

Furthermore, `Videana` has been extended by a GUI that supports efficient search- and retrieval operations in large video databases. Therefore, our visual concept detection approach has been integrated. Initially, the videos are automatically annotated with semantic concepts like “indoor/outdoor”, “studio”, “anchor”, “nature”, “sky”, “city”, “car”, “crowd” or concepts rather related to media sciences such as “bullet time”, “top view”, or “duel”. Currently, an image or video database can be made accessible via approximately 150 visual concepts. These concept annotations allow the user to search for particular video shots via textual queries.

The retrieval GUI consists of two components: the query window (Figure 6.6) and the retrieval window (Figure 6.7). The query dialog is opened by choosing the menu item “Shot Retrieval” from the “Analysis” menu. The available concepts are represented by their names and icons. The icons are grouped into different areas according to the affiliation to the following categories: “objects”, “events”, “sites/scenes”, “persons”, and “miscellaneous”. The category “miscellaneous" allows to specify the shot size and the shot length (e.g., "find all shots longer than two minutes"). Furthermore, it provides queries to search for camera motion types and for text within the optical character recognition (OCR) results. Alltogether, the different conceptual queries can be combined via union, intersection and negation. Submitting the final query opens the retrieval window (Figure 6.7), where the shots are sorted according to their probability. The first shot has the highest probability to fulfill the query. The shots are again represented by the beginning, middle and end frames. A mouse click on one of the three icons opens a video player, that makes the related video shot directly accessible at the appropriate frame position. Additionally, a selection mechanism for evaluation purposes is integrated into the retrieval dialog.

In cooperation with the project *Media Narrations and Media Games* which is part of the collaborative research center *Media Upheavals* (SFB/FK 615), concept models have been developed for novel semantic concepts like "top view", "bullet time", "duel", and "tracking". This project investigated hybrid forms of game and narration, which are observable in computer games and feature films since the 1990ths. The aim was a formal-aesthetical and function-logical analysis of these sequences and a summarization into a typology. The retrieval window in Figure 6.7 shows the query results for the concept "top view".

## 6.4 Summary

Visual concept detection has been applied to support interdisciplinary research efforts in the field of psychology and media sciences. Using a generic concept detection approach several novel concepts have been detected in new application domains. The proposed concept detection approach has been utilized for example in the field of psychological research in an external cooperation, conducted together with Klaus Mathiak (RWTH University Aachen, Germany) and René Weber (University of California, Santa Barbara). The psychological research question studied was whether and how playing violent content in computer games may induce aggression. Therefore, novel semantic concepts, most notably "violence", have been detected in computer game videos to gain insights into the interrelationship of violent game events and the brain activity of a player. Furthermore, the software toolkit *Videana* has been extended by a video retrieval component. Besides about 150 concepts, like "indoor/outdoor", "sky", "waterscape", or "car", several novel concepts have been detected in cooperation with the project *Media Narrations and Media Games*.

The experimental results demonstrated the usefulness of the proposed generic concept detection approach for such interdisciplinary research. In the computer games experiment 91% of the game events were correctly recognized.

Intuitively, one might think that humans always achieve a recognition rate of nearly 100%, but subjectivity and diminishing attention seem to be limiting factors. The comparison of manual annotations against each other shows the performance of automatic software systems in a more favourable light. Considering the fact that Weber et al. [2006] observed an inter-coder reliability of 0.85 Cohen's kappa for human annotators, the automatic system demonstrates an excellent performance.

*"Now this is not the end. It is not even the beginning of the end.
But it is, perhaps the end of the beginning."*

Winston Churchill

# 7 Conclusions

## 7.1 Summary

Current approaches for image and video retrieval focus on semantic concepts serving as an intermediate description to bridge the "semantic gap" between the data representation and the human interpretation. Due to the large complexity and variability in the appearance of visual concepts, the detection of arbitrary concepts represents a very challenging task.

A major problem is to find robust feature representations for successful concept detection systems. This problem together with machine learning issues has been addressed in Chapter 3 and 4.

Chapter 3 focused on the enhancement of local descriptors for mid-level feature coding. The effect of different spatial extents in a state-of-the-art system for visual concept detection has been investigated. Based on the observation that SIFT descriptors with different spatial extents yield large performance differences, a novel concept detection system has been proposed that combines feature representations for different spatial extents using MKL. This system achieved superior performance on the *Mediamill* as well as on the *VOC Challenge* with 44.6% and 54.1% mean AP, respectively. This is a relative improvement of more than 100% compared to the baseline system provided by the *Mediamill Challenge*. To the best of our knowledge, the best reported result for the same color features on this challenge is approximately 42% mean AP. Furthermore, the Bag-of-Words approach has been leveraged for audio features to enhance video concept detection, and MKL has been again proposed as the appropriate fusion scheme for these BoAW and BoVW features. Experiments demonstrate the effectiveness of using

BoAW features: The system using BoAW features yielded a relative performance improvement of approximately 30% mean AP compared to a state-of-the-art audio approach relying on pLSA. In addition, it has been demonstrated that an early fusion scheme degrades detection performance, whereas the combination of auditory and visual Bag-of-Words features via MKL yields a clear performance improvement. Finally, a method for the SIFT-based integration of color information, called CMSIFT, has been presented. Color information can be very helpful to classify semantic concepts, like for example “sunset”, “meadow”, or “sky”. It has been shown experimentally, that CMSIFT descriptors achieve the same concept detection performance as state-of-the-art transformed color SIFT or RGB-SIFT features, while at the same time being considerably faster.

Novel object-based features for the task of visual concept detection have been presented in Chapter 4. Based on the observation that the use of face detection results improved the performance of several face related concepts, further object detectors have been incorporated. Additionally, MKL has been proposed to combine object-based and BoVW features. Extensive experiments on the *Mediamill*, *VOC* and *TRECVid Challenge* showed significant improvements in terms of retrieval performance not only for the object classes, but also in particular for a large number of indirectly related concepts. Moreover, it has been demonstrated that a few object-based features are beneficial for a large number of concepts. On the *VOC Challenge* the combination of BoVW and object-based features via MKL led to a superior performance for the image classification task of 63.8% mean AP, which is a relative performance improvement of 18.9% compared to using BoVW features alone. Thus, BoVW and object-based features seem to be very complementary. Top results were also achieved at the annual *TRECVid Challenge*. The experiments revealed that the approaches additionally employing object-based features significantly improved the overall performance. Some concepts, like “animal”, “bicycling” or “vehicle”, were improved by more than 100% infAP. For the concepts “vehicle”, “ground vehicle”, “overlaid text” and “two people” even the best result of all participating teams was achieved. Furthermore, the generalization capabilities of concept models have been investigated. Different source and target domains led to a severe loss in concept detection performance, both for different channels and genres. It has been shown that additionally used object-based features clearly improve the cross-domain concept detection performance and thus the generalization capabilities of concept models. Additionally, it has been demonstrated how a concurrent multi-class object detection system can be constructed to speed up the detection of many object classes in images since it is inefficient to run a large number of single-class object detectors. This multi-class approach, which is an extension of class-specific Hough forests, is capable of detecting 24 different object classes at a time instead of applying one object detector for each object class separately.

In Chapter 5, the task of gathering training data from the WWW to automatically build and update concept models has been addressed. A novel purely web-supervised system for long-term learning of visual concepts has been presented. Only a single word describing the visual concept is required to initiate the learning process; no further supervision is required. The system continuously updates its learned models while at the same time it deals with scalability by retaining only a small number of training images. Experimental results demonstrate the advantages of the proposed system and in particular the four learning capabilities of the system: 1.) the multi-modal clustering approach is effective in removing spam images; 2.) the system is able to incrementally improve its performance without any manual supervision; 3.) the subclass learning via random savannas is superior to random forests and SVMs; 4.) the scalability manager is very effective by means of keeping only a limited number of positive training images in the system. Furthermore, the random savanna approach is easily parallelizable in the training and testing stage at the tree or forest level.

Finally, visual concept detection has been applied to support interdisciplinary research efforts in the field of psychology and media sciences. This contribution has been presented in Chapter 6. An automatic concept detection system has been built to support psychological experiments. The psychological research question studied was whether and how playing violent content in computer games may induce aggression. Therefore, novel semantic concepts most notably "violence" were detected in computer game videos to gain insights into the interrelationship of violent game events and the brain activity of a player. Experimental results demonstrate the usefulness of the proposed generic concept detection approach for such research. In this experiment 91% of the requested game events were recognized correctly. This is an excellent result considering the inter-coder reliability between the human annotators in the original psychological experimental setting.

Overall, concept detection has reached a level that promises significant benefits for search and navigation in multimedia databases. The robustness of concept detection systems has been improved by novel and enhanced feature representations.

## 7.2 Open Issues and Future Research

Open issues and future directions in research of the addressed issues are discussed below:

**Mid-level feature extraction.** The concept detection performance of a state-of-the-art BoVW system has been significantly boosted by combining different spatial sizes of local descriptors using MKL – this was even more effective than using spatial pyramid representations. An area of future work could be to automatically find an optimal combination of these sizes. Concerning BoAW features, areas for future work are the integration of temporal information beyond the scope of audio frames and the investigation of features based on the matching pursuit method instead of MFCCs. Furthermore, the integration of other coding and pooling strategies, such as super-vector coding, could further improve the detection performance. Recently, super-vector coding seems to be very promising in the field of visual concept detection.

**Object-based features.** First, since it is infeasible to incorporate detection results for all possible objects, the next step is to determine a minimum set of reliable object detectors to achieve a maximum improvement in concept detection. Our experimental results suggest that it is reasonable to use object detectors for frequent object classes. Second, besides their benefits as context features, object sequences can be exploited for further temporal analysis. For example, person sequences can be used to improve the recognition of events, like "person sitting down" or "cycling". Furthermore, in the field of object detection the application of multi-class detectors for different subsets of object classes could be further investigated.

**Web-supervised learning.** First, the impact of other feature sets using color and other clustering methods, such as spectral clustering or latent semantic indexing, should be investigated. Second, different updating and boosting strategies for the ensemble are promising areas of further research. Third, it could be very interesting to use textual information not only for spam detection, but to further consider the correlation of keyword annotations for the detection of relevant concept subclasses. Also, the convergence behavior of the learning algorithm could be analyzed in more detail to determine under what circumstances learning stops and how this can be prevented. Finally, it could be very interesting to explore the possibility of cross-domain learning when using training data from the WWW.

**Concept detection in computer games.** In the given scenario, it could be interesting to detect further semantic concepts, like "danger". This concept depends on the detection of person occurrences and particularly on a "friend-or-foe" distinction. However, these persons appear in very small sizes in the game, and it is even hard for a human annotator to make a decision whether the situation is actually dangerous or not. Furthermore, temporal state transitions promise to entail additional useful information, e.g., the state "inactive" is always preceded by the state "violence". An integration of such temporal relationships could further improve the automatic annotation system.

**Other directions.** Some other aspects have to be mentioned. Currently, large scale image and video retrieval increasingly becomes a focus of research. In this context, the optimization of the memory usage and the computation time are very important factors. A possible solution that has been realized in our BMWi project *Cloud-Based Software Services for Semantic Search in Images and Videos* is the application of cloud computing and cloud service paradigms to build an efficient distributed concept detection system.

A further interesting direction in research are convolutional neural networks or convolutional deep belief networks [Lee et al. 2011]. This technique has been suggested recently in the field of image and video analysis. Similar to the BoVW approach, "features" are locally extracted and are combined in different network layers using a max-pooling strategy. The extracted features are in some way automatically learned during the training process instead of using predefined features. However, this approach is computationally very expensive but has the advantage that only one model has to be built for all concept classes. Promising results have been achieved at the *Large Scale Visual Recognition Challenge* [Deng et al. 2012]. Recently, the top results at the *TRECVid Challenge* 2013 were achieved using a combination of BoVW and convolutional neural network approaches [Snoek et al. 2014].

# Lists and Registers

# List of Figures

# List of Tables

# List of Listings

# Bibliography

A. Abdel-Hakim and A. Farag. CSIFT: A SIFT Descriptor with Color Invariant Characteristics. In *Proceedings of the 19th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 1978–1983, New York, New York, USA, 2006. IEEE.

P. Abend, T. Thielmann, R. Ewerth, D. Seiler, M. Mühling, J. Döring, M. Grauer, and B. Freisleben. Geobrowsing the Globe: A Geovisual Analysis of Google Earth Usage. *Linking GeoVisualization with Spatial Analysis and Modeling (GeoViz)*, 2011.

S. Ahmadi and A. Spanias. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE Transactions on Speech and Audio Processing*, 7(3):333–338, May 1999.

D. Arthur and S. Vassilvitskii. K-Means++: The Advantages of Careful Seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, New Orleans, Louisiana, USA, 2007.

J. A. Aslam, V. Pavlu, and E. Yilmaz. A Statistical Method for System Evaluation using Incomplete Judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 541–548, New York, New York, USA, 2006. ACM.

F. Bach and G. Obozinski. Sparse Methods for Machine Learning - Theory and Algorithms, 2010. URL http://www.di.ens.fr/~fbach/Cours_peyresq_2010.pdf.

F. Bach, G. Lanckriet, and M. Jordan. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. In *Proceedings of the 21st International Conference on Machine Learning (ICML'04)*, pages 1–8, Banff, Alberta, Canada, 2004. ACM.

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *Journal of Computer Vision and Image Understanding*, 110(3):346–359, 2008.

A. Ben-Hur and J. Weston. A User's Guide to Support Vector Machines. 2011. URL http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Y. Bengio, O. Delalleau, and N. L. Roux. Label Propagation and Quadratic Criterion. In O. Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-Supervised Learning*, pages 193–216. MIT Press, 2006.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2009.

A. Bosch, A. Zisserman, and X. Muoz. Image Classification using Random Forests and Ferns. In *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, 2007. IEEE.

A. Bosch, A. Zisserman, and X. Muñoz. Scene Classification Using a Hybrid Generative/Discriminative Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

J.-Y. Bouguet. Pyramidal Implementation of the Affine Lucas Kanade Feature Tracker. 2001. URL http://pages.slc.edu/~aschultz/mocap/Bouget_Affine.pdf.

Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning Mid-Level Features for Recognition. In *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 2559–2566, San Francisco, CA, USA, 2010. IEEE.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

H. Bredin, L. Koenig, and J. Farinas. IRIT @ TRECVid 2010 : Hidden Markov Models for Context-aware Late Fusion of Multiple Audio Classifiers. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, Gaithersburg, Maryland, USA, 2010. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

W. M. Campbell, D. E. Sturim, D. A. Reynolds, and W. Street. Support Vector Machines using GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.

C.-C. Chang and C.-J. Lin. {*LIBSVM*}*: A Library for Support Vector Machines*, 2012. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm.

S.-F. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D.-Q. Zhang. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'05)*, Gaithersburg, Maryland, USA, 2005. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, 2006.

K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The Devil is in the Details: an Evaluation of Recent Feature Encoding Methods. In *Proceedings of the 22nd British Machine Visision Conference (BMVC'11)*, pages 76.1–76.12, Dundee, Scotland, UK, 2011. British Machine Vision Association.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proceedings of the 18th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 886–893, San Diego, CA, USA, 2005. IEEE.

R. Datta, D. Joshi, J. Li, and J. Wang. Tagging Over Time: Real-World Image Annotation by Lightweight Meta-Learning. In *Proceedings of the 15th International Conference on Multimedia (MM'07)*, pages 393–402, Augsburg, Germany, 2007. ACM.

J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2012. URL `http://www.image-net.org/challenges/LSVRC/2012`.

C. Diou, G. Stephanopoulos, and A. Delopoulos. The Multimedia Understanding Group at TRECVID-2010. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, Gaithersburg, Maryland, USA, 2010. NIST. URL `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm`.

T. Dittrich, S. Kopf, P. Schaber, B. Guthier, and W. Effelsberg. Saliency Detection for Stereoscopic Video. In *Proceedings of the 4th ACM Multimedia Systems Conference (MMSYS)*, pages 12–23, Oslo, Norway, 2013.

N. Elleuch, M. Zarka, I. Feki, A. Ben Ammar, and A. M. Alimi. REGIMVID at TRECVID 2010 : Semantic Indexing. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, Gaithersburg, Maryland, USA, 2010. NIST. URL `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm`.

M. Ester and J. Sander. *Knowledge Discovery in Databases*. Springer Verlag, Berlin, 2000.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

R. Ewerth. *Robust Video Content Analysis via Transductive Ensemble Learning.* PhD thesis, Department of Mathematics and Computer Science, University of Marburg, Germany, 2008.

R. Ewerth and B. Freisleben. Video Cut Detection without Thresholds. In *Proceedings of the 11$^{th}$ International Workshop on Signals, Systems and Image Processing (IWSSIP'04)*, pages 227–230, Poznan, Poland, 2004.

R. Ewerth, M. Schwalb, P. Tessmann, and B. Freisleben. Estimation of Arbitrary Camera Motion in MPEG Videos. In *Proceedings of the 17$^{th}$ International Conference on Pattern Recognition (ICPR'04)*, volume 1, pages 512–515, Cambridge, UK, Aug. 2004. IEEE.

R. Ewerth, M. Mühling, and B. Freisleben. Self-Supervised Learning of Face Appearances in TV Casts and Movies. In *Proceedings of the 8$^{th}$ IEEE International Symposium on Multimedia (ISM'06)*, pages 78–85, Washington, DC, USA, 2006a. IEEE.

R. Ewerth, M. Mühling, T. Stadelmann, E. Qeli, B. Agel, D. Seiler, and B. Freisleben. University of Marburg at TRECVID 2006: Shot Boundary Detection and Rushes Task Results. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'06)*, Gaithersburg, Maryland, USA, 2006b. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

R. Ewerth, M. Mühling, and B. Freisleben. Self-Supervised Learning of Face Appearances in TV Casts and Movies. *International Journal of Semantic Computing*, 1(2):185–204, 2007a.

R. Ewerth, M. Mühling, T. Stadelmann, J. Gllavata, M. Grauer, and B. Freisleben. Videana: A Software Toolkit for Scientific Film Studies. In *Proceedings of the International Workshop on Digital Tools in Film Studies*, pages 1–16, Siegen, Germany, 2007b. Transcript Verlag.

R. Ewerth, M. Mühling, and B. Freisleben. Robust Video Content Analysis via Transductive Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):1–26, 2011.

R. Ewerth, K. Ballafkir, M. Mühling, D. Seiler, and B. Freisleben. Long-Term Incremental Web-Supervised Learning of Visual Concepts via Random Savannas. *IEEE Transactions on Multimedia*, 14(4):1008–1020, 2012.

J. Fan, C. Yang, Y. Shen, N. Babaguchi, and H. Luo. Leveraging Large-Scale Weakly-Tagged Images to Train Inter-Related Classifiers for Multi-Label Annotation. In *Proceedings of the 1$^{st}$ ACM Workshop on Large-Scale Multimedia*

*Retrieval and Mining (LS-MMRM'09)*, pages 27–34, New York, New York, USA, 2009. ACM.

G. Fanelli, J. Gall, and L. Van Gool. Hough Transform-Based Mouth Localization for Audio-Visual Speech Recognition. In *Proceedings of the 20th British Machine Visision Conference (BMVC'09)*, London, UK, 2009. British Machine Vision Association.

L. Fei-Fei, R. Fergus, and P. Perona. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In *Computer Vision and Image Understanding*, volume 106, pages 59–70. Elsevier, 2007.

I. Feki, A. B. Ammar, and A. M. Alimi. Audio Stream Analysis for Environmental Sound Classification. In *Proceedings of the International Conference on Multimedia Computing and Systems (ICMCS'11)*, Ouarzazate, Morocco, 2011. IEEE.

P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61(1):55–79, Jan. 2005.

P. F. Felzenszwalb, D. McAllester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. In *Proceedings of the 21st IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, Anchorage, Alaska, USA, June 2008. IEEE.

P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade Object Detection with Deformable Part Models. In *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 2241–2248, San Francisco, CA, USA, 2010a. IEEE.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection With Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sept. 2010b.

R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google's Image Search. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05)*, pages 1816–1823, Beijing, China, 2005. IEEE.

Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

J. Gall and V. Lempitsky. Class-Specific Hough Forests for Object Detection. In *Proceedings of the 22nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, pages 1022–1029, Miami Beach, Florida, USA, 2009.

H. J. Gans. *Deciding What's News: a Study of CBS Evening News, NBC Nightly News, Newsweek, and Time.* Vintage Books, New York, NY, USA, 1980.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993.

J. C. v. Gemert, J. M. Geusebroek, C. Veenman, and A. W. M. Smeulders. Kernel Codebooks for Scene Categorization. In *Proceedings of the 10th European Conference on Computer Vision (ECCV'08)*, pages 696–709, Marseille, France, 2008. Springer.

J. C. v. Gemert, C. G. M. Snoek, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek. Comparing Compact Codebooks for Visual Categorization. *Computer Vision and Image Understanding*, 114(4):450–462, Apr. 2010a.

J. C. v. Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010b.

J. Gllavata and R. Ewerth. Text Detection in Images Based on Unsupervised Classification of High-Frequency Wavelet Coefficients. In *Proceedings of 17th International Conference on Pattern Recognition (ICPR'04)*, pages 425–428, Cambridge, UK, 2004. IEEE.

Z. Gong, Q. Liu, and J. Guo. Deriving Semantic Terms for Images by Mining the Web. In *Proceedings of the 11th International Conference on Electronic Commerce (ICEC'09)*, number 2, pages 323–328, New York, New York, USA, 2009. ACM. ISBN 9781605585864.

D. Gorisse, F. Precioso, P. Gosselin, L. Granjon, D. Pellerin, M. Rombaut, H. Bredin, L. Koenig, H. Lachambre, E. E. Khoury, R. Vieux, B. Mansencal, Y. Zhou, J. Benois-Pineau, H. Jégou, S. Ayache, B. Safadi, G. Quénot, A. Benoît, and P. Lambert. IRIM at TRECVID 2010: Semantic Indexing and Instance Search. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, Gaithersburg, Maryland, USA, 2010. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

K. Grauman and T. Darrell. The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05)*, number October, pages 1458–1465, Beijing, China, 2005. IEEE.

G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. Technical report, California Institute of Technology, 2007. URL http://authors.library.caltech.edu/7694.

A. Hauptmann, R. Yan, and W.-H. Lin. How Many High-Level Concepts Will Fill the Semantic Gap in News Video Retrieval? In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR'07)*, pages 627–634, Amsterdam, The Netherlands, 2007. ACM.

C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A Practical Guide to Support Vector Classification. Technical Report 1, 2010.

Y. Hu and P. Loizou. Subjective Comparison of Speech Enhancement Algorithms. In *Proceedings of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, pages 153–156, Toulouse, France, 2006. IEEE.

N. Inoue and K. Shinoda. A Fast MAP Adaptation Technique for Gmm-Supervector-Based Video Semantic Indexing Systems. In *Proceedings of the 19th ACM International Conference on Multimedia (MM'11)*, pages 1357–1360, Scottsdale, Arizona, USA, 2011. ACM.

N. Inoue and K. Shinoda. A Fast and Accurate Video Semantic Indexing System Using Fast MAP Adaptation and GMM Supervectors. *IEEE Transactions on Multimedia*, 6(1):1–22, 2012.

N. Inoue, T. Saito, K. Shinoda, and S. Furui. High-Level Feature Extraction Using SIFT GMMs and Audio Models. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10)*, pages 3220–3223, Istanbul, Turkey, Aug. 2010a. IEEE.

N. Inoue, T. Wada, Y. Kamishima, K. Shinoda, I. Kim, B. Byun, and C.-H. Lee. TT + GT at TRECVID 2010 Workshop. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, Gaithersburg, Maryland, USA, 2010b. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

N. Inoue, T. Wada, Y. Kamishima, K. Shinoda, and S. Sato. TokyoTech + Canon at TRECVID 2011. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'11)*, Gaithersburg, Maryland, USA, 2012. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

N. Inoue, Y. Kamishima, K. Mori, and K. Shinoda. TokyoTechCanon at TRECVID 2012. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'12)*, Gaithersburg, Maryland, USA, 2013. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

S. Jeannin and B. Mory. Video Motion Representation for Improved Content Access. *IEEE Transactions on Consumer Electronics*, 46(3):645–655, 2000.

W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. Loui. Short-Term Audio-Visual Atoms for Generic Video Concept Classification. In *Proceedings of the 17th ACM International Conference on Multimedia (MM'09)*, pages 5–14, Vancouver, British Columbia, Canada, 2009a. ACM.

Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR'07)*, pages 494–501, Amsterdam, The Netherlands, 2007. ACM.

Y.-G. Jiang, C.-W. Ngo, and S.-F. Chang. Semantic Context Transfer Across Heterogeneous Sources for Domain Adaptive Video Search. In *Proceedings of the 17th ACM International Conference on Multimedia (MM'09)*, pages 155–164, Vancouver, British Columbia, Canada, 2009b. ACM.

Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann. Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study. *IEEE Transactions on Multimedia*, 12(1):42–53, 2010a.

Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, Gaithersburg, Maryland, USA, 2010b. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

T. Joachims. Text Categorization With Support Vector Machines: Learning With Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning (ECML'98)*, pages 137–142, Chemnitz, Germany, 1998. Springer.

T. Joachims. Transductive Inference for Text Classification using Support Vector Machines. In *Proceedings of the 16th International Conference on Machine Learning (ICML'99)*, pages 200–209, Bled, Slovenia, 1999.

F. Jurie and B. Triggs. Creating Efficient Codebooks for Visual Recognition. In *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05)*, pages 604–610, Beijing, China, 2005. IEEE.

L. Kennedy, S.-F. Chang, and I. Kozintsev. To Search or To Label?: Predicting the Performance of Search-Based Automatic Image Classifiers. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR'06)*, pages 249–258, Santa Barbara, CA, USA, 2006. ACM.

M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and Accurate Lp-Norm Multiple Kernel Learning. *Advances in Neural Information Processing Systems*, 22(1):997–1005, 2009.

M. Kloft, U. Rückert, and P. Bartlett. A Unifying View of Multiple Kernel Learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'10)*, number II, pages 66–81, Barcelona, Spain, 2010. Springer.

M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-Norm Multiple Kernel Learning. *Journal of Machine Learning Research*, 12(1):953–997, 2011.

P. Koniusz and K. Mikolajczyk. Spatial Coordinate Coding to Reduce Histogram Representations, Dominant Angle and Colour Pyramid Match. In *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP'11)*, pages 661–664, Brussels, Belgium, 2011. IEEE.

S. Kopf. *Computergestützte Inhaltsanalyse von digitalen Videoarchiven.* PhD thesis, Department of Computer Science, University of Mannheim, Germany, 2006.

J. Krapac, J. Verbeek, and F. Jurie. Modeling Spatial Layout with Fisher Vectors for Image Categorization. In *Proceedings of the 13th International Conference on Computer Vision (ICCV'11)*, pages 1487–1494, Barcelona, Spain, Nov. 2011. IEEE.

P. Kruizinga and N. Petkov. Nonlinear Operator for Oriented Texture. *IEEE Transactions on Image Processing*, 8(10):1395–407, Jan. 1999.

G. Kühne, S. Richter, and M. Beier. Motion-based Segmentation and Contour-based Classification of Video Objects. In *Proceedings of the 9th ACM International Conference on Multimedia (MM)*, pages 41–50, Ottawa, Ontario, Canada, 2001.

V. Kumar and I. Patras. A Discriminative Voting Scheme for Object Detection using Hough Forests. In *Proceedings of the 2nd British Machine Vision Conference Postgraduate Workshop*, pages 1–10, Aberystwyth, UK, 2010. British Machine Vision Association.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the 19th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 2169–2178, Washington, DC, USA, 2006. IEEE.

B. H. Lee, R. Grosse, R. Ranganath, and A. Ng. Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks. *Communications of the ACM*, 54(10):95–103, 2011.

B. Leibe, A. Leonardis, and B. Schiele. Robust Object Detection with Interleaved Categorization and Segmentation. *International Journal of Computer Vision*, 77(1-3):259–289, 2008.

H. Li, L. Bao, Z. Gao, A. Overwijk, W. Liu, L.-F. Zhang, S.-I. Yu, M.-Y. Chen, F. Metze, and A. Hauptmann. Informedia @ TRECVID 2010. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, Gaithersburg, Maryland, USA, 2010a. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

L.-J. Li and L. Fei-Fei. OPTIMOL: Automatic Online Picture Collection via Incremental Model Learning. *International Journal of Computer Vision*, 88 (2):147–168, July 2009.

L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS'10)*, pages 1–9, Vancouver, British Columbia, Canada, 2010b.

R. Lienhart, S. Pfeiffer, and W. Effelsberg. The MoCA Workbench: Support for Creativity in Movie Content Analysis. In *Proceedings of the 3rd IEEE Conference on Multimedia Computing and Systems*, pages 314–321, Hiroshima, Japan, 1996. IEEE.

R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video Abstracting. *Communications of the ACM*, 40(12):54–62, 1997.

R. Lienhart, W. Effelsberg, and R. Jain. VisualGREP: A Systematic Method to Compare and Retrieve Video Sequences. *Multimedia Tools and Applications*, 10(1):47–72, 1999a.

R. Lienhart, S. Pfeiffer, and W. Effelsberg. Scene Determination based on Video and Audio Features. In *IEEE International Conference on Multimedia Computing and Systems*, pages 685–690, Florence, Italy, 1999b.

R. Lienhart, L. Liang, and A. Kuranov. A Detector Tree of Boosted Classifiers for Real-Time Object Detection and Tracking. In *Proceedings of the 4th IEEE International Conference on Multimedia and Expo (ICME'03)*, number c, pages 277–280, Baltimore, Maryland, USA, 2003. IEEE.

L. Liu, L. Wang, and X. Liu. In Defense of Soft-assignment Coding. In *Proceedings of the 13th International Conference on Computer Vision (ICCV'11)*, pages 2486–2493, Barcelona, Spain, Nov. 2011. IEEE.

Y. Liu, D. Xu, I. W. Tsang, and J. Luo. Using Large-Scale Web Data to Facilitate Textual Query Based Retrieval of Consumer pPhotos. In *Proceedings of* *the 17$^{th}$ ACM International Conference on Multimedia (MM'09)*, pages 55–64, Vancouver, British Columbia, Canada, 2009. ACM.

D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

L. Lu and A. Hanjalic. Audio Keywords Discovery for Text-Like Audio Content Analysis and Retrieval. *IEEE Transactions on Multimedia*, 10(1):74–85, 2008.

L. Lu, H.-J. Zhang, and S. Z. Li. Content-Based Audio Classification and Segmentation by Using Support Vector Machines. *Multimedia Systems*, 8(6):482–492, Apr. 2003.

B. D. Lukas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, Vancouver, Canada, 1981.

S. Maji and J. Malik. Object Detection Using a Max-Margin Hough Transform. In *Proceedings of the 22$^{nd}$ IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, pages 1038–1045, Miami Beach, Florida, USA, 2009. IEEE.

S. Mallat and Z. Zhang. Matching Pursuits with Time-frequency Dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

B. S. Manjunath, J.-R. Ohm, and V. V. Vasudevan. Color and Texture Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6): 703–715, 2001.

J. M. Martinez. MPEG-7 Overview. Technical report, ISO/IEC, Klagenfurt, Austria, 2002.

K. Mathiak and R. Weber. Toward Brain Correlates of Natural Behavior: fMRI during Violent Video Games. *Human Brain Mapping*, 27(12):948–56, Dec. 2006.

E. Mbanya, S. Gerke, and P. Ndjiki-Nya. Spatial Codebooks for Image Categorization. In *Proceedings of the 1$^{st}$ ACM International Conference on Multimedia Retrieval (ICMR'11)*, Trento, Italy, 2011. ACM.

K. Mikolajczyk and C. Schmid. Indexing Based on Scale Invariant Interest Points. In *Proceedings of the 8$^{th}$ IEEE International Conference on Computer Vision (ICCV'01)*, pages 525–531, Vancouver, British Columbia, Canada, 2001. IEEE.

F. Moosmann, B. Triggs, and F. Jurie. Fast Discriminative Visual Codebooks using Randomized Clustering Forests. In *Proceedings of the 20$^{th}$ Annual Conference on Neural Information Processing Systems (NIPS'06)*, pages 1–7, Vancouver, British Columbia, Canada, 2006.

M. Mühling, R. Ewerth, T. Stadelmann, B. Freisleben, R. Weber, and K. Mathiak. Semantic Video Analysis for Psychological Research on Violence in Computer Games. In *Proceedings of the 6$^{th}$ ACM International Conference on Image and Video Retrieval (CIVR'07)*, pages 611–618, Amsterdam, The Netherlands, July 2007a. ACM.

M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, C. Zöfel, and B. Freisleben. University of Marburg at TRECVID 2007: Shot Boundary Detection and High-Level Feature Extraction. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'07)*, Gaithersburg, Maryland, USA, 2007b. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, and B. Freisleben. University of Marburg at TRECVID 2008: High-Level Feature Extraction. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'08)*, Gaithersburg, Maryland, USA, 2008. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

M. Mühling, R. Ewerth, and B. Freisleben. Improving Semantic Video Retrieval via Object-Based Features. In *Proceedings of the 3$^{rd}$ IEEE International Conference on Semantic Computing (ICSC'09)*, pages 109–115, Berkeley, CA, USA, 2009a. IEEE.

M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, and B. Freisleben. University of Marburg at TRECVID 2009: High-Level Feature Extraction. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'09)*, Gaithersburg, Maryland, USA, 2009b. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

M. Mühling, R. Ewerth, T. Stadelmann, B. Shi, and B. Freisleben. University of Marburg at TRECVID 2010: Semantic Indexing. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, Gaithersburg, Maryland, USA, 2010. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

M. Mühling, K. Ballafkir, R. Ewerth, and B. Freisleben. University of Marburg at TRECVID 2011: Semantic Indexing. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'11)*, Gaithersburg, Maryland, USA, 2011a. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

M. Mühling, R. Ewerth, and B. Freisleben. On the Spatial Extents of SIFT Descriptors for Visual Concept Detection. In *Proceedings of the 8$^{th}$ International Conference on Computer Vision Systems (ICVS'11)*, pages 71–80, Sophia Antipolis, France, 2011b. Springer.

M. Mühling, R. Ewerth, B. Shi, and B. Freisleben. Multi-Class Object Detection with Hough Forests Using Local Histograms of Visual Words. In *Proceedings of 14th International Conference on Computer Analysis of Images and Patterns (CAIP'11)*, pages 386–393, Seville, Spain, 2011c. Springer.

M. Mühling, R. Ewerth, J. Zhou, and B. Freisleben. Multimodal Video Concept Detection via Bag of Auditory Words and Multiple Kernel Learning. In *Proceedings of the 18th International Conference on Advances in Multimedia Modeling (MMM'12)*, pages 40–50, Klagenfurt, Austria, 2012. Springer.

M. Naphade and J. R. Smith. On the Detection of Semantic Concepts at TRECVID. In *Proceedings of the 12th Annual ACM International Conference on Multimedia (MM'04)*, pages 660–667, New York, New York, USA, 2004. ACM.

M. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann. A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005. Technical report, 2005.

M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE Multimedia Magazine*, 13(3):86–91, 2006.

K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learning*, 39(2): 103–134, 2000.

E. Nowak, F. Jurie, and B. Triggs. Sampling Strategies for Bag-of-Features Image Classification. In *Proceedings of the 9th European Conference on Computer Vision (ECCV'06)*, pages 490–503, Graz, Austria, 2006. Springer.

P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton. TRECVID 2006 - An Overview. Technical report, Gaithersburg, Maryland, USA, 2007. URL `http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6overview.pdf`.

P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. Smeaton, W. Kraaij, and G. Quéenot. TRECVID 2010 — An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'10)*, pages 1–34, Gaithersburg, Maryland, USA, 2011. National Institute of Standards and Technology (NIST). URL `http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/tv10overview.pdf`.

P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. Smeaton, W. Kraaij, and G. Quéenot. TRECVID 2011 — An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics. In *Proceedings of the TREC*

*Video Retrieval Evaluation Workshop (TRECVid'11)*, pages 1–56, Gaithersburg, Maryland, USA, 2012. National Institute of Standards and Technology (NIST). URL http://www-nlpir.nist.gov/projects/tvpubs/tv11.papers/tv11overview.pdf.

P. Over, J. Fiscus, G. Sanders, B. Shaw, G. Awad, M. Michel, A. Smeaton, W. Kraaij, and G. Quéenot. TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'12)*, pages 1–58, Gaithersburg, Maryland, USA, 2013. National Institute of Standards and Technology (NIST). URL http://www-nlpir.nist.gov/projects/tvpubs/tv12.papers/tv12overview.pdf.

Y. Peng, Z. Lu, and J. Xiao. Semantic Concept Annotation Based on Audio PLSA Model. In *Proceedings of the 17th ACM International Conference on Multimedia (MM'09)*, pages 841–844, New York, New York, USA, 2009. ACM.

F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *Proceedings of the 11th European Conference on Computer Vision (ECCV'10)*, pages 143–156, Heraklion, Crete, Greece, 2010. Springer.

J. C. Platt. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods*, pages 185–208. MIT Press, Apr. 1999.

M. Porter. An Algorithm for Suffix Stripping. *Program: Electronic Library and Information Systems*, 14(3):130–137, 1980.

G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative Multi-Label Video Annotation. In *Proceedings of the 15th International Conference on Multimedia (MM'07)*, pages 17–26, Augsburg, Germany, 2007a. ACM.

G.-J. Qi, X.-S. Hua, Y. Song, and H.-J. Zhang. Transductive Inference with Hierarchical Clustering for Video Annotation. In *Proceedings of the 8th IEEE International Conference on Multimedia and Expo (ICME'07)*, pages 643–646, Beijing, China, 2007b. IEEE.

M. Riley, E. Heinen, and J. Ghosh. A Text Retrieval Approach to Content-Based Audio Retrieval. In *Proceedings of the 9th International Conference of Music Information Retrieval (ISMIR'08)*, pages 295–300, Philadelphia, Pennsylvania, USA, 2008.

D. A. Sadlier and N. E. O'Connor. Event Detection in Field Sports Video Using Audio-Visual Features and a Support Vector Machine. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1225–1233, Oct. 2005.

K. E. A. v. d. Sande, T. Gevers, and C. G. M. Snoek. A Comparison of Color Features for Visual Concept Classification. In *Proceedings of the 7th ACM International Conference on Content-Based Image and Video Retrieval (CIVR'08)*, pages 141–150, Niagara Falls, Ontario, Canada, 2008. ACM.

K. E. A. v. d. Sande, T. Gevers, and C. G. M. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1582–96, Sept. 2010.

F. Schroff, A. Criminisi, and A. Zisserman. Harvesting Image Databases from the Web. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(4):754–66, Apr. 2011.

D. Seiler, R. Ewerth, S. Heinzl, T. Stadelmann, M. Mühling, B. Freisleben, and M. Grauer. Eine Service-Orientierte Grid-Infrastruktur zur Unterstützung medienwissenschaftlicher Filmanalyse. In *Proceedings of the Workshop on Gemeinschaften in Neuen Medien (GeNeMe'09)*, pages 79–89, Dresden, Germany, Sept. 2009.

A. F. Smeaton, P. Over, and W. Kraaij. Evaluation Campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR'06)*, pages 321–330, New York, New York, USA, 2006. ACM.

A. F. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In *Multimedia Content Analysis*, chapter Theory and, pages 151–174. 2009.

A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(12):1349–1380, 2000.

C. Snoek, D. Fonjne, Z. Li, K. V. D. Sande, and A. Smeulders. Deep Nets for Detecting, Combining, and Localizing Concepts in Video. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'13)*, Gaithersburg, Maryland, USA, 2014. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

C. G. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders. The Semantic Pathfinder: Using an Authoring Metaphor for Generic Multimedia Indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(10):1678–1689, Oct. 2006a.

C. G. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In *Proceedings of the 14th Annual ACM International*

*Conference on Multimedia (MM'06)*, pages 421–430, Santa Barbara, CA, USA, 2006b. ACM.

C. G. Snoek, K. E. A. van de Sande, O. D. Rooij, B. Huurnink, J. R. R. Uijlings, M. V. Liempt, M. Bugalho, I. Trancoso, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J.-M. Geusebroek, T. Gevers, M. Worring, A. W. M. Smeulders, and D. C. Koelma. The MediaMill TRECVID 2009 Semantic Video Search Engine. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'09)*, Gaithersburg, Maryland, USA, 2009. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.htm.

C. G. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, E. Gavves, D. Odijk, M. D. Rijke, T. Gevers, M. Worring, D. C. Koelma, and A. W. M. Smeulders. The MediaMill TRECVID 2010 Semantic Video Search Engine. In *Proceedings of the TREC Video Retrieval Evaluation Workshop (TRECVid'11)*, Gaithersburg, Maryland, USA, 2011. NIST. URL http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/mediamill.pdf.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large Scale Multiple Kernel Learning. *Journal of Machine Learning Research*, 7(1):1531–1565, 2006.

S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. Bona, A. Binder, C. Gehl, and V. Franc. The SHOGUN Machine Learning Toolbox. *Journal of Machine Learning Research*, 11(1):1799–1802, 2010.

T. Stadelmann. *Voice Modeling Methods for Automatic Speaker Recognition.* PhD thesis, Department of Mathematics and Computer Science, University of Marburg, Germany, 2010.

J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring Semantic Concepts from Community-Contributed Images and Noisy Tags. In *Proceedings of the 17th ACM International Conference on Multimedia (MM'09)*, pages 223–232, New York, New York, USA, 2009. ACM.

X. Tian, L. Yang, and J. Wang. Transductive Video Annotation via Local Learnable Kernel Classifier. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'08)*, pages 1509–1512, Hannover, Germany, 2008. IEEE.

X. Tong, Q. Liu, L. Duan, H. Lu, C. Xu, and Q. Tian. A Unified Framework for Semantic Shot Representation of Sports Video. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'05)*, pages 127–134, Singapore, Singapore, 2005. ACM.

A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing Visual Features for Multiclass and Multiview Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):854–869, 2007.

A. Torralba, R. Fergus, and W. T. Freeman. 80 Million Tiny Images: a Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(11):1958–1970, Nov. 2008.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support Vector Machine Learning for Interdependent and Structured Output Spaces. In *Proceedings of the 21st International Conference on Machine Learning (ICML'04)*, pages 104–112, Banff, Alberta, Canada, 2004. ACM.

T. Tuytelaars and K. Mikolajczyk. Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008.

T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine. Unsupervised Object Discovery: A Comparison. *International Journal of Computer Vision*, 88(2):284–302, July 2009.

A. Ulges, M. Worring, and T. Breuel. Learning Visual Contexts for Image Annotation From Flickr Groups. *IEEE Transactions on Multimedia*, 13(2):330–341, Apr. 2011.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

A. Vedaldi and B. Fulkerson. VLFeat — An Open and Portable Library of Computer Vision Algorithms. In *Proceedings of the 18th ACM International Conference on Multimedia (MM'10)*, pages 1469–1472, Firence, Italy, 2010. ACM.

V. Viitaniemi and J. Laaksonen. Experiments on Selection of Codebooks for Local Image Feature Histograms. In *Proceedings of the 10th International Conference on Visual Information Systems. Web-Based Visual Information Search and Management (VISUAL'08)*, pages 126–137, Salerno, Italy, 2008. Springer.

P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Scalable Search-Based Image Annotation of Personal Images. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR'06)*, pages 269–278, Santa Barbara, CA, USA, 2006. ACM.

G. Wang, T.-S. Chua, and M. Zhao. Exploring Knowledge of Sub-Domain in a Multi-Resolution Bootstrapping Framework for Concept Detection in News Video. In *Proceedings of the 16th ACM International Conference on Multimedia (MM'08)*, pages 249–258, Vancouver, British Columbia, Canada, 2008a. ACM.

J. Wang, Y. Zhao, X. Wu, and X.-S. Hua. Transductive Multi-Label Learning for Video Concept Detection. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (MIR'08)*, pages 298–304, Vancouver, British Columbia, Canada, 2008b. ACM.

J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-Constrained Linear Coding for Image Classification. In *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 3360–3367, San Francisco, CA, USA, 2010.

X. Wang, L. Zhang, X. Li, and W. Ma. Annotating Images by Mining Image Search Results. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(11):1919–1932, 2008c.

R. Weber, U. Ritterfeld, and K. Mathiak. Does Playing Violent Video Games Induce Aggression? Empirical Evidence of a Functional Magnetic Resonance Imaging Study. *Media Psychology*, 8:39–60, 2006.

J. V. D. Weijer and C. Schmid. Coloring Local Feature Extraction. In *Proceedings of the 9th European Conference on Computer Vision (ECCV'06)*, pages 334–348, Graz, Austria, 2006. Springer.

I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, second edition, 2005.

W. Wojcikiewicz, A. Binder, and M. Kawanabe. Enhancing Image Classification with Class-Wise Clustered Vocabularies. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10)*, pages 1060–1063, Istanbul, Turkey, Aug. 2010. IEEE.

J. Wu, D. Ding, X.-S. Hua, and B. Zhang. Tracking Concept Drifting with an Online-Optimized Incremental Learning Framework. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'05)*, pages 33–40, Singapore, Singapore, 2005. ACM.

H. Xu and T.-S. Chua. Fusion of AV Features and External Information Sources for Event Detection in Team Sports Video. *Transactions on Multimedia Computing, Communications, and Applications*, 2(1):44–67, Feb. 2006.

H. Xu, X. Zhou, M. Wang, Y. Xiang, and B. Shi. Exploring Flickr's Related Tags for Semantic Annotation of Web Images. In *Proceedings of the 8th ACM*

*International Conference on Image and Video Retrieval (CIVR'09)*, page 1, Santorini, Fira, Greece, 2009. ACM.

J. Yang and A. G. Hauptmann. (Un)Reliability of Video Concept Detection. In *Proceedings of the 7th ACM International Conference on Image and Video Retrieval (CIVR'08)*, pages 85–94, Niagara Falls, Canada, 2008. ACM.

J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C. Ngo. Evaluating Bag-of-Visual-Words Representations in Scene Classification. In *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR'07)*, pages 197–206, Augsburg, Bavaria, Germany, 2007a. ACM.

J. Yang, R. Yan, and A. G. Hauptmann. Cross-Domain Video Concept Detection Using Adaptive SVMs. In *Proceedings of the 15th International Conference on Multimedia (MM'07)*, pages 188–197, Augsburg, Germany, 2007b. ACM.

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In *Proceedings of the 22nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, pages 1794–1801, Miami Beach, Florida, USA, June 2009. IEEE.

A. Yao, J. Gall, and L. Van Gool. A Hough Transform-Based Voting Framework for Action Recognition. In *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 2061–2068, San Francisco, CA, USA, 2010. IEEE.

E. Yilmaz and J. Aslam. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*, pages 102–111, Arlington, Virginia, USA, 2006. ACM.

E. Yilmaz, E. Kanoulas, and J. Aslam. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610, Singapore, Singapore, 2008. ACM.

K. Yu, T. Zhang, and Y. Gong. Nonlinear Learning using Local Coordinate Coding. *Advances in Neural Information Processing Systems (NIPS)*, 22:2223–2231, 2009.

J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

X. Zhang, Y.-C. Song, J. Cao, Y.-D. Zhang, and J.-T. Li. Large Scale Incremental Web Video Categorization. In *Proceedings of the 1st Workshop on Web-Scale*

*Multimedia Corpus (WSMC'09)*, pages 33–40, New York, New York, USA, 2009. ACM.

W.-L. Zhao, X. Wu, and C.-W. Ngo. On the Annotation of Web Videos by Efficient Near-Duplicate Search. *IEEE Transactions on Multimedia*, 12(5): 448–461, Aug. 2010.

D. Zhou and C. J. C. Burges. Spectral Clustering and Transductive Learning with Multiple Views. In *Proceedings of the 24th International Conference on Machine Learning (ICML'07)*, pages 1159–1166, Corvallis, Oregon, USA, 2007. ACM.

D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with Local and Global Consistency. *Advances in Neural Information Processing Systems (NIPS)*, 16:321–328, 2004.

X. Zhou, K. Yu, T. Zhang, and T. Huang. Image Classification using Super-Vector Coding of Local Image Descriptors. In *Proceedings of the 11th European Conference on Computer Vision (ECCV'10)*, pages 141–154, Heraklion, Crete, Greece, 2010. Springer.

# Erklärung

Ich versichere, dass ich meine Dissertation

**Visual Concept Detection in Images and Videos**

selbständig, ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen und Hilfen bedient habe. Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den 24.03.2014

Markus Mühling