Department of Information and Computer Science

Visual category detection: an experimental perspective

Ville Viitaniemi

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 45/2012

Visual category detection: an experimental perspective

Ville Viitaniemi

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Science for public examination and debate in Auditorium TU2 at the Aalto University School of Science (Espoo, Finland) on the 9th of May 2012 at noon.

Aalto University
School of Science
Department of Information and Computer Science

Supervisor
Prof. Erkki Oja

Instructor
Dr. Jorma Laaksonen

Preliminary examiners
Dr. Gabriela Csurka, Xerox Research Centre Europe, France
Prof. Serkan Kiranyaz, Tampere University of Technology, Finland

Opponent
Prof. Alan Smeaton, Dublin City University, Ireland

Aalto University publication series
DOCTORAL DISSERTATIONS 45/2012

ISBN 978-952-60-4585-6 (printed)
ISBN 978-952-60-4586-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2012

Finland

The dissertation can be read at http://lib.tkk.fi/Diss/

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Ville Viitaniemi

Name of the doctoral dissertation
Visual category detection: an experimental perspective

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 45/2012

Field of research Computer and Information Science

Manuscript submitted 9 September 2011 **Manuscript revised** 12 January 2012

Date of the defence 9 May 2012 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Nowadays huge volumes of digital visual data are constantly being produced and archived. Automatically distilling useful information from such information masses requires one to somehow answer the grand long-standing question of computer vision: how to make computers understand images?

In this thesis the visual content analysis problem is looked at as a category detection problem. In the category detection formulation, the general image content understanding task is partitioned into a number of small binary decision tasks. In each of the sub-tasks, one decides whether an image belongs to some pre-defined category. A category could be defined, for example, to consist of images taken indoors. By defining an appropriate set of categories, the visual content of an image can be described on a desired level of granularity by determining the image's membership in each one of the categories.

This thesis discusses a framework for visual category detection that consists of three major components: feature extraction, feature-wise detection and fusion of the detection results. The point of view in the discussion is empirical: the framework is validated by the good levels of performance systems implementing it have demonstrated in various benchmark tasks of visual analysis. A body of experiments is described that compare various technological alternatives for implementing the three major components of the framework. In addition to comparing implementation techniques, the experiments demonstrate that the discussed generic category detection architecture is very versatile: a set of diverse visual analysis problems can be addressed using the same visual category detection system as a backbone component by equipping the system with a task-specific front-end.

From the experiments and discussion in the thesis, one can conclude that the category detection formulation is a useful way of approaching the general image content understanding problem. In category detection, performances reaching the state-of-the-art can be realised using the presented fusion-based system architecture and implementation technologies of the system components.

Keywords computer vision, image analysis, visual category, feature fusion, local image descriptor

ISBN (printed) 978-952-60-4585-6 **ISBN (pdf)** 978-952-60-4586-3

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2012

Pages 322 **The dissertation can be read at** http://lib.tkk.fi/Diss/

Tekijä
Ville Viitaniemi

Väitöskirjan nimi
Visuaalisten kategorioiden tunnistaminen: kokeellinen näkökulma

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietojenkäsittelytieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 45/2012

Tutkimusala Informaatiotekniikka

Käsikirjoituksen pvm 09.09.2011 **Korjatun käsikirjoituksen pvm** 12.01.2012

Väitöspäivä 09.05.2012 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Nykyaika tuottaa jatkuvasti valtavia määriä visuaalista digitaalista aineistoa. Jotta näistä suurista tietoaineistoista voitaisiin automaattisesti löytää käyttökelpoista informaatiota, olisi löydettävä jonkinlainen vastaus tietokonenäön pitkäaikaiseen peruskysymykseen: kuinka saada tietokoneet ymmärtämään kuvien sisältöä?

Tässä väitöskirjassa visuaalisen sisällön luonnehtimista tarkastellaan kategorioiden tunnistamisen näkökulmasta. Yleinen kuvan sisällön luonnehtimistehtävä pilkotaan lukuisiksi pieniksi kyllä-ei -päätöstehtäviksi. Kussakin yksittäisessä päätöstehtävässä vastataan kysymykseen, kuuluuko tarkasteltava kuva johonkin ennalta määrättyyn kategoriaan. Voitaisiin esimerkiksi määritellä, että sisätiloissa otetut kuvat muodostavat yhden kategorian. Kuvien sisältöä voidaan kuvailla halutulla yksityiskohtaisuustasolla määrittelemällä sopiva joukko kategorioita ja tunnistamalla kunkin kategorian kohdalla, mitkä kuvat siihen kuuluvat.

Väitöskirjassa käsitellään mallia, jossa kategoriantunnistusjärjestelmä koostuu kolmesta pääosasta: piirreirrotuksesta, piirrekohtaisesta tunnistuksesta sekä näiden tunnistustulosten fuusiosta. Tekstin näkökulma on kokeellinen: tämän järjestelmäarkkitehtuurin toimivuus perustellaan hyvillä suorituskykyarvoilla, joita siihen perustuvat järjestelmät ovat saavuttaneet erilaisissa visuaalisen analyysin suorituskykyä mittaavissa tehtävissä. Väitöskirjassa kuvataan lukuisia kokeita, joissa arvioidaan eri tekniikoita järjestelmän kolmen pääkomponentin toteuttamiseksi. Toteutustekniikoiden vertailemisen lisäksi kokeet myös osoittavat, että esitetty yleiskäyttöinen kategoriantunnistusmalli on hyvin joustava: joukko erilaisia visuaalisia analyysitehtäviä on voitu ratkaista järjestelmällä, jonka ydinosan kaikissa tapauksissa muodostaa sama kategoriantunnistinkomponentti. Eri tehtäviä varten ydin on ympäröity tehtäväkohtaisilla sovitinosilla.

Väitöskirjassa esitettyjen kokeiden ja analyysien perusteella voidaan päätellä, että kategorioiden tunnistaminen on käyttökelpoinen tapa lähestyä yleistä kuvien sisällön tulkitsemistehtävää. Voidaan myös todeta, että esitetyllä piirrefuusiota hyödyntävällä järjestelmäarkkitehtuurilla ja esitetyillä järjestelmän osien toteutustekniikoilla saavutetaan tämänhetkinen huipputaso kategoriantunnistuksessa.

Avainsanat konenäkö, kuva-analyysi, visuaalinen kategoria, piirrefuusio, paikallinen kuvapiirre

ISBN (painettu) 978-952-60-4585-6 **ISBN (pdf)** 978-952-60-4586-3

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2012

Sivumäärä 322 **Luettavissa verkossa osoitteessa** http://lib.tkk.fi/Diss/

Preface

This work has been done while working under the guidance of D.Sc. Jorma Laaksonen and Professor Erkki Oja. The work started in the Laboratory of Computer and Information Science at Helsinki University of Technology, and was finished in the Department of Information and Computer Science in the School of Science of Aalto University. Despite the apparently numerous changes in the organisation, the research environment has been very stable and supportive from the point of view of an individual researcher. For this I am very grateful towards the wise leadership of the department/laboratory.

Working with Jorma Laaksonen has been an enjoyable experience. Always has the door to his room been open, and stupid questions about the departmental computer system, LaTeX trickery and inner workings of the PicSOM system have often received his immediate attention. After a few minutes of hands-on experimentation, the solution has often been found right away. I also want to thank the other members of our research group, D.Sc. Markus Koskela and M.Sc. Mats Sjöberg in particular, for making the working environment to be a friendly and immediate one.

Finally, I wish to thank my parents for their continuous support over the years.

Espoo, March 28, 2012,

Ville Viitaniemi

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Ville Viitaniemi and Jorma Laaksonen. Techniques for still image scene classification and object detection. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN 2006)*, Part II, pages 35–44, Athens, Greece, September 2006.

II Ville Viitaniemi and Jorma Laaksonen. Techniques for image classification, object detection and object segmentation. In *Proceedings of the 10th International Conference on Visual Information Systems (VISUAL 2008)*, pages 231–234, Salerno, Italy, September 2008.

III Ville Viitaniemi and Jorma Laaksonen. Evaluating the performance in automatic image annotation: example case by adaptive fusion of global image features. *Signal Processing: Image Communications*, Volume 22, issue 6, pages 557–568, July 2007.

IV Ville Viitaniemi and Jorma Laaksonen. Improving the accuracy of global feature fusion based image categorisation. In *Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies (SAMT 2007)*, pages 1–14, Genova, Italy, December 2007.

V Mats Sjöberg, Markus Koskela, Ville Viitaniemi and Jorma Laaksonen. Indoor location recognition using fusion of SVM-based visual classifiers. In *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 343–348, Kittilä, Finland, August-September 2010.

VI Ville Viitaniemi, Mats Sjöberg, Markus Koskela and Jorma Laaksonen. Concept-based video search with the PicSOM multimedia retrieval system. *Technical report TKK-ICS-R39*, Aalto University, December 2010.

VII Ville Viitaniemi and Jorma Laaksonen. Experiments on selection of codebooks for local image feature histograms. In *Proceedings of the 10th International Conference on Visual Information Systems (VISUAL 2008)*, pages 126–137, Salerno, Italy, September 2008.

VIII Ville Viitaniemi and Jorma Laaksonen. Combining local feature histograms of different granularities. In *Proceedings of the 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, pages 636–645, Oslo, Norway, June 2009.

IX Ville Viitaniemi and Jorma Laaksonen. Spatial extensions to bag of visual words. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR 2009)*, pages 636–645, Fira, Greece, July 2009.

X Ville Viitaniemi and Jorma Laaksonen. Region matching techniques for spatial bag of visual words based image category recognition. In *Proceedings of the 20th International Conference on Artificial Neural Networks (ICANN 2010)*, Part I, pages 531–540, Thessaloniki, Greece, September 2010.

XI Ville Viitaniemi and Jorma Laaksonen. Representing images with χ^2 distance based histograms of SIFT descriptors. In *Proceedings of the 19th International Conference on Artificial Neural Networks (ICANN 2009)*, Part II, pages 694–703, Limassol, Cyprus, September 2009.

List of symbols

$a_w^{\mathrm{gt}}, a_w^{\mathrm{pred}}$	binary indicator variables of keyword w
$\mathbf{a}^{\mathrm{gt}}, \mathbf{a}^{\mathrm{pred}}$	image-wise vectors of a_w^{gt} and a_w^{pred}
AP	average precision
$\mathrm{AP}_{\mathrm{int}}$	interpolated average precision
AUC	area under ROC curve
$c(i)$	number of correctly predicted keywords for ith image
C,C_n	(nth) image category
C	C-SVC cost function penalty parameter
$d(\cdot,\cdot), d_l(\cdot,\cdot)$	distance function
$d_{\chi^2}(\cdot,\cdot)$	χ^2 distance
DTMI	de-symmetrised termwise mutual information
$E_i[\cdot]$	average over test set images
E_i	ith early fusion combined feature
$E_w[\cdot]$	average over keywords
f_{F}	mapping function of fusion method F
F_i	ith elementary visual feature
F_α	F-measure of retrieval quality
$g_{\mathrm{T}}(\mathbf{x}_i, \mathbf{x}_j)$	SVM kernel function of type T
$\mathrm{GT}(i)$	binary ground truth function of the ranked list
$H(X)$	entropy of X
$H(X\|Y)$	conditional entropy of X given Y
$\hat{H}(X\|Y)$	de-symmetrised conditional entropy of X given Y
i,j,l,m,n	indexing variables
$I(i)$	ith image in a ranked list
I,I_i	(ith) image
I_i,I_s	binary indicator variables
$I(X;Y)$	mutual information between random variables X and Y
J	C-SVC cost function
k, k_i	dimensionality of feature vector
L_i	i:th location (room)

$\hat{L}$	prediction of location
M	number of (retrieved) images
M_{C}	number of correct retrievals
M'	number of images taken from the beginning of the list
$n(i)$	number of incorrectly predicted keywords for ith image
N	size of (image) collection
N_{C}	number of relevant images in the collection
N_{D}	number of detectors
N_{E}	number of early fusion features
N_{F}	number of elementary visual features
N_{P}	number of partitions in partitioning of training data
N_{train}	training set size
NS	normalised score
$p(X)$	probability of X
$p(X\|Y)$	conditional probability of X given Y
P	precision
$P(m)$	precision at depth m
P_w	precision for keyword w
q_i, Q	counts of variables being selected in search
R	recall
$R(m)$	recall at depth m
R_w	recall for keyword w
$\widetilde{Rank}$	normalised average rank
s, s_i	detection score
$\mathbf{s} = [s_1\ s_2\ \cdots\ s_{N_{\mathrm{D}}}]$	vector of detector outcomes
$\mathbf{S}$	matrix of detector outcomes
w	keyword
$w(i)$	annotation length of the ith image in the ground truth
$\mathbf{w}, b$	C-SVC weights
W	size of vocabulary
$\|w_c\|$	number of correct predictions of word w
$\|w_{\mathrm{gt}}\|$	number of occurrences of word w in the ground truth
$w_{\max}$	maximum allowed annotation length
$\|w_{\mathrm{pred}}\|$	number of predictions of word w
x	scalar variable
X, Y	(discrete,binary) random variables
$\mathcal{X}, \mathcal{Y}$	support of X and Y
$\mathbf{x}_i = [x_{i1}\ x_{i2}\ \cdots\ x_{ik}]^T$	ith feature vector

y_i	binary class label associated with $\mathbf{x}_i$
α	balance factor in the F-measure
γ, γ_l	kernel smoothness parameter
λ, λ_l	weighting factor, combination weight
ξ_i	internal slack variable in the C-SVC cost function
$\phi(\mathbf{x}_i)$	mapping of $\mathbf{x}_i$ to SVM feature space
$\varphi(x)$	logistic sigmoid

Abbreviations

AP	average precision
ANMRR	averaged normalised modified retrieval rate
AUC	area under curve
BBR	Bayesian binary/logistic regression
BoV	bag of visual words
CBIR	content based image/information retrieval
CIE	Commission international de l'eclairage, International Commission of Illumination
CD	compact disc
CMRM	cross-media relevance model
CRM	continuous relevance model
DCT	discrete cosine transform
DD-SVM	diverse density support vector machine
DFT	discrete Fourier transform
DoG	difference of Gaussians
DTMI	de-symmetrised termwise mutual information
EM	expectation maximisation
EMD	earth mover's distance
GLOH	gradient location and orientation histogram
GMM	Gaussian mixture model
HLFE	high-level feature extraction
HMM	hidden Markov model
HMMD	hue-max-min-diff
HOG	histogram of oriented gradient
HSV	hue-saturation-value
infAP	inferred average precision
IRM	integrated region matching
KL	Kullback-Leibler
KLT	Kanade-Lucas-Tomasi
LBG	Linde-Buzo-Gray

LESH	local energy based shape histogram
LSCOM	large scale concept ontology for multimedia
LVQ	learning vector quantisation
MAP	maximum a posteriori
MAP	mean average precision
MARS	multimedia analysis and retrieval system
MBRM	multiple Bernoulli relevance model
MDL	minimum description length
MFCC	mel-frequency cepstral coefficient
MF-KDA	non-sparse multiple kernel Fisher discriminant analysis
MIAP	mean inferred average precision
MIL	multiple-instance learning
MI-SVM	support vector machine for multiple instance learning
MKL	multiple kernel learning
MoSIFT	motion SIFT
MPEG	Moving Picture Experts Group
MSE	mean square (quantisation) error
MSER	maximally stable extremal region
mSFBS	multifold-SFBS
NIST	National Institute of Standards and Technology
NS	normalised score
PR	precision-recall
QBIC	query by image content
RBF	radial basis function
RGB	red-green-blue
ROC	receiver operating characteristic
SBS	sequential backward search
SFBS	sequential forward-backward search
	sequential floating backward search
SFFS	sequential floating forward search
SFS	sequential forward search
SIFT	scale-invariant feature transform
SOM	self-organising map
SR-KDA	kernel discriminant analysis using spectral regression
SURF	speeded up robust features
SVM	support vector machine
TREC	text retrieval conference
TSIS	text space to image space

TS-SOM	tree-structured self-organising map
TV	television
VOC	visual object classes

1 Introduction

1.1 Background

In the world of today, huge volumes of digital visual content are produced every day. Media professionals create some of this digital content, but also ordinary people have become content producers. This content is supplemented by the data produced by automatic means such as satellites, surveillance cameras and webcams. The amount of visual data one can access is increased by the fact that nowadays much of the data is shared through the Internet. All in all, one is left with enormous collections of visual data from which lots of potentially useful and interesting information could be distilled.

In order to draw full benefit from the collections, one needs to be able to index, search and browse the collections by their content. An example scenario could be searching the archives of a TV broadcaster in search for particular types of entertainment programs or news clips about some particular event in history. One could also think of some data mining applications, for example in a futuristic scenario the surveillance camera data from a certain part of a city could be analysed and the types of people moving in that area at certain times of day profiled, so that the stores in that area could better target their special offers towards these people groups.

One possible approach facilitating the content-based accessing of collections of visual information would be to attach detailed textual annotations to each piece of visual content. However, typically such annotations are not provided by the content generation phase or the annotations are very incomprehensive. Usually one would need to analyse the content of the images and videos afterwards.

Due to its laboriousness, careful visual analysis by human inspection and manual annotation is limited to some highest-priority applications, such as medical diagnostics and uses in military reconnaissance. The

range of applications for which visual analysis is practicable broadens immensely if the analysis can be performed with automatic methods. This provides a strong incentive for developing such automatic visual analysis methods, which is also the topic of this thesis. If reliable automatic methods would be available, one could imagine, for instance, letting the computer automatically organise and index the digital images one takes with his mobile phone.

In addition to being potentially very useful, automatic visual analysis is such a challenging goal that a completely satisfactory solution is not even in sight in the foreseeable future. This is not because of lack of trying. In fact, the area has continuously attracted intense research attention since the invention of modern electronic computers. Further research is still needed to answer the question how to best implement an automatic visual analysis system.

Examples of well-working visual analysis systems can be found in the brains of animals. If one only could emulate the visual processing of the brain, that would be an excellent solution. Indeed, some parts of early visual processing are understood well enough so that they can be used as an inspiration for components of an artificial visual analysis system. However, the system-level workings of the brain remain yet too unclear for one to be able to form an effective vision system out of elementary components by simply replicating the biological example.

Another route to the synthesis of a visual analysis system would be first to formulate a mathematical model of the process that generates natural images and then to invert the model. Currently this approach does not seem promising either. Generation of natural images is such a complex process that modelling it in a general case seems a formidable task. Furthermore, even if a model were available, inverting it even approximately would be infeasible.

Because of the lack of a good and complete underlying theory and the inability to decipher workings of complete biological systems, the synthesis of visual analysis systems is currently an engineering discipline that proceeds by trial and error. The systems are typically complex and consist of multiple components. A large number of researchers worldwide are continuously experimenting with new system architectures and component technologies. Not all the components in any particular system are usually of equally good quality. Still, through comparison of the systems and technologies, for instance in international evaluations and standard

benchmark tasks, well-working component technologies gradually gain popularity in the research community. As a consequence, the performance of the systems progressively improves.

1.2 Contents of the thesis

In this thesis, the visual analysis problem is addressed by converting the general problem into a series of binary visual category detection tasks. Figure 1.1 schematically shows the main components of a prototypical category detection system. This generic system architecture has emerged from years of experimentation in the research community—including the decade-long investigations by the PicSOM research group—as the prevailing method of constructing systems for detecting visual categories.

The presentation in the thesis formalises the architecture as a concise framework—the PicSOM category detection framework. Within this framework, various technological alternatives for implementing the system components are discussed and compared in the light of experiments.

In this thesis, selecting a certain way to construct a category detection system is justified empirically: if some category detection technique has consistently produced good performance in practical category detection

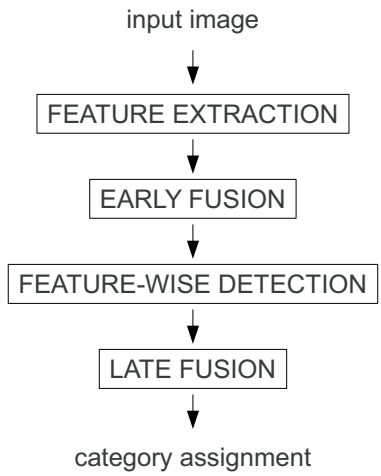


Figure 1.1. Components of a category detection system.

experiments, then using it is well-founded. Furthermore, among alternative category detection techniques, the one that produces the best performance in experiments can be chosen. Explicitly formulating this empirical principle may appear simplistic and self-evident. Yet it may be worth stating in order to remind of the reason for choosing this purely empirical point of view instead of one supported by theory: there currently exist no adequate applicable theories of complete vision systems or of visual categories, as pointed out already in Section 1.1.

In the empirical spirit, this thesis collects together and discusses a body of experimental knowledge. Some of the experimental results are from the publications of the thesis, while a large part of them originates from the published reports of experiments made by other researchers. As one type of empirical results the publications of the thesis demonstrate that the PicSOM category detection architecture can be successfully applied to a wide variety of visual analysis benchmark tasks. The presented experiments also empirically justify the use of some component technologies instead of others. On basis of the diverse category detection experiments, the thesis provides an extensive discussion of various technical alternatives.

This thesis consists of an introductory part and eleven original publications, whose contents will be detailed in Section 1.3. The introductory part is structured as follows. After this introduction, Chapter 2 provides necessary background information for the forthcoming detailed discussion. It introduces the formulation of visual analysis problems as category detection problems. It also prepares the reader for Chapter 3 by discussing the issue of measuring category detection performance, as well as by surveying some related work in the area of visual analysis.

Chapter 3 forms the methodological essence of the thesis. There a general-level system architecture for performing the visual category detection task is first defined in Section 3.1. The components of that architecture are then discussed in detail. Separate Sections 3.2, 3.3 and 3.4 are devoted for feature extraction, supervised detection and fusion techniques, respectively. The point of view of the discussion is empirical: different techniques for implementing the components are compared in the light of how well they perform in practical category detection experiments. The performance of various techniques in the TRECVID 2009 benchmark is used as an important example.

Subsection 3.2.2 of the feature extraction section is an important part

of the thesis. It discusses the Bag-of-Visual-Words (BoV) feature extraction paradigm. In recent years the BoV features—i.e. histograms of local image features—have turned out to be very effective features for visual category detection. The details and extensions of the BoV paradigm form one specific focus of the experiments in a number of publications of the thesis—and also of the discussion in this introductory part.

Chapter 4 complements the previous discussion by demonstrating how the architecture—implemented in the PicSOM framework—is applied to various benchmark tasks of visual analysis. These include automatic image annotation (Section 4.1), object detection, localisation and segmentation (Section 4.2), semantic multimedia search (Section 4.3) and robot navigation (Section 4.4). Chapter 5 summarises the introductory part.

Most of the experimental material as well as its analysis in Chapters 3 and 4 is reproduced from the publications of the thesis. On one hand, this saves the reader the burden of collecting the material piece by piece from the publications. On the other hand, gathering together the somewhat fragmentary material helps one to see how the individual experiments are related to each other and how they contribute to the general understanding of category detection systems. However, some additional experimental results that were omitted from the publications of the thesis, mainly for space reasons or for the sake of compactness of representation in the publications, are also reported in the introductory part of the thesis.

1.3 Content of the publications and the author's contributions

Publications I–VI demonstrate the application of the PicSOM category detection framework to diverse image analysis tasks. Publication I describes how the PicSOM category detection system has been applied for the task of object detection and localisation in photographic images of the PASCAL VOC 2006 benchmark collection. In the publication, algorithms are proposed and validated for using unsupervised image segmentation as a basis for object localisation. Similarly, Publication II describes the application of an improved version of the PicSOM framework to the 2007 edition of the PASCAL VOC benchmark.

In Publication III the PicSOM framework with global image features is applied to the automatic image annotation task, in particular to the widely-used Corel benchmark. The article discusses the performance

metrics of image annotation and introduces the novel de-symmetrised termwise mutual information (DTMI) measure. In Publication IV the PicSOM framework in an improved form is applied to a slightly different annotation task of the Corel images.

Publication V shows how the PicSOM framework can be used in mobile robot localisation. Publication VI describes the use of the PicSOM methodology in semantic multimedia indexing and search in the TRECVID 2008–2009 benchmark setting.

Publications VII–XI address the details and extensions of the BoV methodology. They all include a set of experiments that utilise a subset of VOC 2007 benchmark images and object categories. Publication VII describes methods and experiments in selecting BoV codebooks with various clustering algorithms and by supervised learning. In Publication VIII techniques for combining different codebook granularities in the BoV paradigm are described and experimentally compared. In Publication IX spatial extensions to BoV are described and compared in experiments, including the novel spatially soft tiling technique. Publication X compares region matching schemes in spatial BoV extensions. Publication XI investigates the idea of basing comparison of SIFT descriptors on the χ^2 distance instead of the Euclidean one.

The methods and experiments for Publications I–IV and VII–XI have been designed and implemented by the author of the thesis. The co-author has assisted in evaluating and reporting the results. The present author participated in performing the experiments for Publication V and wrote a part of the text. The writing of Publication VI was coordinated by the author of the thesis. The design of the methods and experiments, as well as writing of Publication VI was a collaborative effort by all the authors.

1.4 Contributions of the thesis

The author considers the following to be the most important contributions of this thesis:

1. Providing a unified viewpoint to the decade-long PicSOM research under the framework of visual category detection.
2. Surveying the research in the area of category detection and collecting together experimental evidence—both that reported by other

researchers and that provided by the experiments with the PicSOM system—that can be used for selecting the most suitable techniques for implementing a visual category detection system.

3. Demonstrating that the generic PicSOM category detection framework provides a basis on which task-specific systems can be built, resulting in competitive performance in many visual analysis benchmark tasks and competitions.
4. Investigating details of the BoV representation and its extensions in numerous experiments.
5. Introducing the spatially soft tiling technique for combining spatial scales in BoV systems.
6. Introducing the DTMI measure of image annotation quality.

The first two items in the list can be considered to be the contribution of this introductory part of the thesis. Otherwise, the scientific contributions of this thesis are contained within its publications.

2 Visual category detection task

Visual information processing represents a major share of a human's cognitive tasks. We can be assured of this being the case by remembering that a considerable proportion of the brain's volume is devoted solely for this purpose. It is therefore natural that from very early on, the developers of thinking machines—computers—have been interested in automating also this branch of information processing. Solving the computer vision problem would have opened immense possibilities of relieving humans from most of their routine tasks, thereby revolutionising the world.

Partially this has already happened. For many applications, computerised vision systems reach acceptable levels of accuracy so that human labour can be replaced in these tasks. For example, mail can be automatically sorted in postal sorting centres by the envelopes' zip codes with automated systems whose accuracies approach that of human workers (e.g. [101]). In another application, surveillance cameras mounted in parking garages can automatically recognise the license plates of cars parked there [7, 162, 213]. This facilitates—for instance—the identification of frequent or privileged customers, automatic time-based billing of the customers' pre-paid accounts, or contacting the car owners if rules of parking times are not followed.

In the seemingly simple examples mentioned above, the systems' domain of application is narrow. It is well regulated what can appear in the images. In the case of the license plate recognition application, for instance, the system would be totally unable to deliver any usable information of anything else in the parking area apart from the license plates, e.g. humans moving in the area. The imaging conditions are also rather controlled, for example the relative position of the camera and the objects to be recognised is always the same.

Despite the encouraging examples mentioned above, computer vision is still far from a solved problem. Generally, problems emerge when one moves towards broader domains by relaxing the restrictions on the image

content and imaging conditions. The accuracy of the present-day computer vision systems still deteriorates to an unusably low level if the domain is made general enough. Such general-domain applications are of growing importance, however, due to the huge volumes of digital visual content produced in today's world. For example, thousands of terrestrial, satellite and internet television channels send and archive their broadcasts routinely 24 hours a day. The availability of relatively cheap digital cameras—including those in mobile phones—and webcams has made also virtually everyone in the modern society a content producer. Inexpensive surveillance cameras monitor many private properties and public places nowadays. Common to these examples is that the produced image and video streams are very uncontrolled, consisting of practically all sorts of imagery. These are just the kind of general image domains that are very challenging for automatic analysis. However, the ability to perform visual analysis automatically would be very valuable as the high data volumes practically rule out careful manual analysis, except possibly in some high-priority applications such as military reconnaissance and medical diagnostics.

In this thesis the visual analysis problems are looked at from the category detection point of view. In category detection, the general question of the contents of an image or a video is turned into a series of small binary decision problems: does the image belong to a particular pre-defined visual category, for example to the category "outdoors". The category detection approach provides a route to generic image and video content analysis if only the elementary category membership problems can be solved efficiently.

The field of visual object and category detection has seen rapid progress during the last 15 years. The research group in which the author has worked has been a part of the development. Since the latter half of 90s the group has been investigating matters related to image and video retrieval. The research has revolved around a system called PicSOM [94, 95], the group's image analysis and information retrieval platform. The system was originally inspired as a visual counterpart of the WEBSOM text document management system [84] that was built to demonstrate the power of the self-organising map (SOM) algorithm [89]. First, the emphasis was on interactive content-based image retrieval (CBIR). However, gradually the repertoire was widened and the PicSOM system has been used for analysing various category detection benchmark data sets and for partic-

ipating in a number of benchmark competitions. This thesis describes a part of the PicSOM research since year 2001.

Chapter 3 of the thesis is devoted to the definition and discussion of a generic feature fusion based framework of a visual category detection system while the remainder of this chapter provides some background for this discussion. In this thesis, the point of view is empirical: the whole category detection architecture and its component technologies are justified in the light of the performance in category detection benchmarks. Therefore, Sections 2.3 and 2.4 discuss the category detection performance measures and benchmark data sets in some length. Before that, however, Section 2.1 elaborates on the issue of expressing visual analysis problems in the form of category detection, and Section 2.2 lists some examples of visual analysis tasks that have been approached by applying the category detection methodology. Concluding this chapter, Section 2.5 provides some historical background against which the techniques presented in the remaining chapters of thesis can be compared.

2.1 Visual analysis as category detection

Visual analysis seeks to answer the question "What can one see in the given image I". Similar question can be posed about the contents of videos. Even for just one specific image, there exists a very large variety of correct answers to the question. To make the problem and the answers manageable, the single open-ended question is turned in the category detection approach into a series of binary questions: "Does image I belong to category C_n", where the index n enumerates all the categories of interest. By defining an appropriate set of categories, an image's content can be described on a desired level of granularity by determining the image's membership in each one of the categories. In practice, instead of the binary formulation, the question is usually formulated probabilistically: "What is the likelihood of the image belonging to category C_n". Alternatively, one may just want to order a given set of images $\{I_1, \ldots, I_M\}$ according to their likelihood to belong to the category.

Figure 2.1 displays a set of images and several different image categories defined in the set. The categories can be used for indexing and searching the set of images. For example, based on the defined categories one can easily identify images that correspond to the description "A person

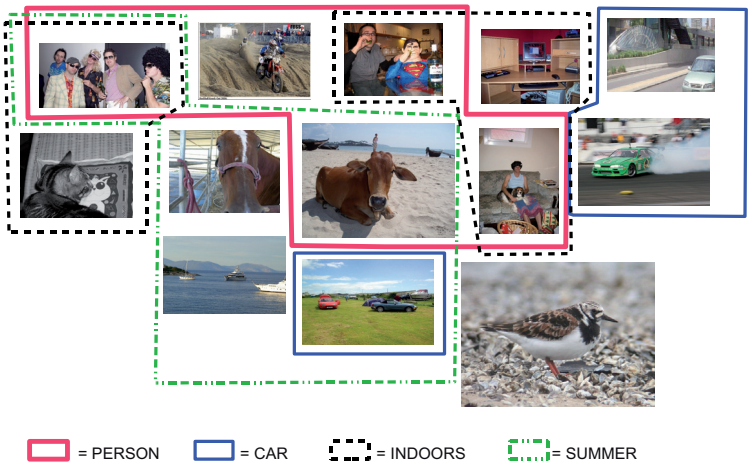


Figure 2.1. A collection of images with some image categories marked.

indoors in summer."

One can see from the example that the problem formulation is very generic: the image categories can be very different in nature. The category types can be grouped to the following classes:

1. Instances of specific object or person, e.g. "the Eiffel tower".
2. Instances of an object class, e.g. "car".
3. Scene types, e.g. "outdoors".

The category types are listed in the order of increasing generality, type 1 being the most specialised. If one is able to restrict to the more specialised category types, more specialised and therefore potentially more effective image analysis methods can be applied. For instance, one may try to fit a geometric transformation of a template image to the images to be tested if one knows that one is looking for a specific object. When the target is an object class, one may be able to benefit from the typical part composition and geometric layout of the objects, even though there may be variability within the object class. This is exactly what the so-called constellation models for object recognition do [22, 54, 55].

In the methods discussed in this thesis, however, all the above mentioned types of image categories are handled in similar fashion. By sacrificing some of the accuracy brought by the possibly existing specialised

detection methods for the type 1 and 2 categories, the same methodology can be applied to all category types. The thought of compromising any accuracy may not sound very appealing. However, there are some justifications that can be given. In practice, the application of generic visual concept detection methodology to some rather specialised tasks has shown to produce very competitive performance (e.g. Publications I–VI of this thesis). The last decade has seen significant improvement in the area of generic object and category detection methods. There seems to be no reason to believe that the development would have reached a standstill yet. By pushing the limits of the generic category detection methods as far as possible, one can simultaneously gain improvement in many different specialised tasks. Only when a saturation point is reached of no more easy improvement by generic methods would it appear necessary to try to reach further gains by delving into the particulars of the more specialised individual problems.

One driving force promoting the development of straightforward, generic methods applicable for many different types of categories comes from practice. On one hand, the huge data volumes speak for straightforward and rather lightweight category detection. The geometric object detection methods such as constellation models can scale badly with increasing data volumes. At least with the current computer resources, computational complexity is an issue. On the other hand, in practice it has turned out to be beneficial to detect hundreds of different categories, for example for the purpose of indexing multimedia collections. Given the number of categories, hand-crafting a specialised detector for each concept is out of the question. Generic category detection is therefore called for. Interestingly, also psycho-visual experiments have inspired frameworks capable of learning a large number of categories. It has been shown that humans can recognise in the order of ten thousand visual categories with ease [15]. Furthermore, learning to recognise a novel, previously unencountered category requires only a minimal number of training examples to be seen. This would seem to favour simpler generic models as learning more complex models generally requires more training and examples.

In this thesis, the visual category detection task is formulated as a supervised learning problem. The likelihood of novel test images to belong to a certain category is assessed on basis of the test images' visual similarity with positive example images, contrasted to corresponding similarity

with negative example images. The visual properties of images are naturally correlated with the category memberships in varying degree. Some categories might be extremely visual, for example category “red”. On the other hand, the category “scene before 1957” might be impossible to recognise with certainty based on visual cues alone.

2.2 Example applications of visual category detection

In this section some example applications of visual category detection are introduced. Common to all these diverse visual tasks is that the generic category detection system discussed in this thesis has been applied to the tasks rather successfully. The example applications are briefly addressed here because the discussion of the category detection system in Chapter 3 includes some experimental results obtained in these tasks. More complete description of the experiments in some of the tasks will be presented in Chapter 4.

Application examples of the generic category detection system discussed in this thesis include:

- automatic image annotation ([193, 195], Publications III and IV),
- mobile robot navigation (Publication V),
- semantic multimedia indexing and search ([166, 197], Publication VI),
- object localisation (Publications I and II),
- defect type classification on paper quality control ([74]),
- interpretation of satellite images ([123, 124]),
- real-time face recognition for a wearable augmented reality device ([3, 4]),
- interactive browsing of electronic fashion catalogues ([192]).

In automatic image annotation the goal is to associate keywords with images. In the supervised formulation, the keywords for a training portion of images are given to the system. The system is then expected to label an independent set of test images similarly. Automatic image annotation can be regarded as a category detection problem by regarding each of the keywords of the fixed vocabulary as a separate visual category.

In the mobile robot navigation setting a robot moves through an office environment capturing pictures continuously with cameras mounted on

it. The task of the automatic system is to name the room in which the robot moves on basis of the captured images. Given some manually labelled training material with pictures captured in known locations, the problem is a supervised category detection problem with different rooms defining the categories.

Semantic multimedia indexing and search addresses the problem of finding data relevant to the users' information needs from multimedia databases. Lately, it has become common to build semantic representations of multimedia content by applying machine learning techniques for detecting mid-level semantic concepts (events, objects, locations, people, etc.) on basis of the content's low-level visual and aural features [92, 130, 175]. In recent studies it has been observed that, despite the far-from-perfect accuracy of concept detectors, the representation is often very useful in supporting high-level indexing and querying of multimedia data [69]. The application of the visual category detection framework to the multimedia indexing problem is straightforward: the different mid-level semantic concepts form the categories that are to be detected.

In the object localisation application the goal is to pinpoint the locations of objects within images, not only whether they are present somewhere in the image. This requires one to detect the visual categories of parts of images, instead of the categories of the whole images. Sliding window methods are a common approach to object localisation. There a rectangular window is moved over the image area, with steps small enough to result in adequate spatial resolution of object localisation. It is straightforward to determine for each window location whether it belongs to the category of object or no-object. An alternative approach that has been realised using the PicSOM system, is to employ an unsupervised segmentation algorithm on the images. The category detection system is used for predicting the likelihood of each image segment to present the location of the object.

In the defect type classification application the image material comes from line cameras mounted over the production line in a paper mill, producing a continuous real-time feed of images of paper surface. The categories to be detected are the various types of defects that can appear on the surface.

In the satellite image interpretation application multi-spectral satellite images are partitioned into small sub-images—imagelets. The category

detection system is then used for deciding, which of the imagelets contain man-made structures, such as buildings. Comparing images taken from the same area at different times, one is also able to detect areas where changes have occurred.

In the face recognition application for a wearable augmented reality device, image material is captured from cameras worn or carried by the user. In order to augment the environment the user sees in a see-through display, objects of interest are recognised in real-time from the video feed by comparing the images to known objects in a database. Currently the detected categories correspond to faces of different persons, facilitating the augmentation of the faces with some potentially useful background information of the persons currently seen, such as their names and scientific publication histories.

Interactive browsing of mail-order catalogues of fashion houses is an example of interactive content-based image retrieval. There the user defines the image category of interest by selecting some of the images the system shows her. The system then displays more images from the catalogue that are likely to belong to the same category. The category definition evolves dynamically as the user tags newly shown images as interesting, thereby adding them to the set of on-line training examples of the category.

Chapter 4 details the application of the PicSOM system to those tasks that have been addressed in the original publications of this thesis.

2.3 Performance measures

Category detection performance can be measured in various ways. In many cases, category detection is used as an ingredient of a system that solves a particular image content analysis problem. A category detection module is used for producing an intermediate result that is processed further to determine the final output of the complete system. In specific applications, it often makes sense to use performance measures tailored to the applications so that the measure directly evaluates the success of the entire system in the application. Such application-specific performance measures are briefly described in connection with the respective applications in Chapter 4 that gives examples of the application of the PicSOM category detection system to various real-world tasks.

However, there are also more generic measures of category detection ac-

curacy. Here one can separate two main cases based on the output format of the category detection system. In the first case, the output of the system is a ranked list of images, ordered according to decreasing likelihood to belong to the category. In the second case, a binary decision is made for each image whether it belongs to the category to be detected. Generalising to simultaneous detection of multiple categories, the category detection outcome can be thought as a binary decision matrix where each row corresponds to one image, and each column to a category. Figure 2.2 shows examples of decision matrices. Common to these two groups of performance measures is that the detection outcomes are compared against a ground truth. The ground truth might originate—for example—from human annotation of the images in question, or it might be known a priori by the construction of the category detection experiment. For the measures of the following sections, the ground truth judgements must be binary.

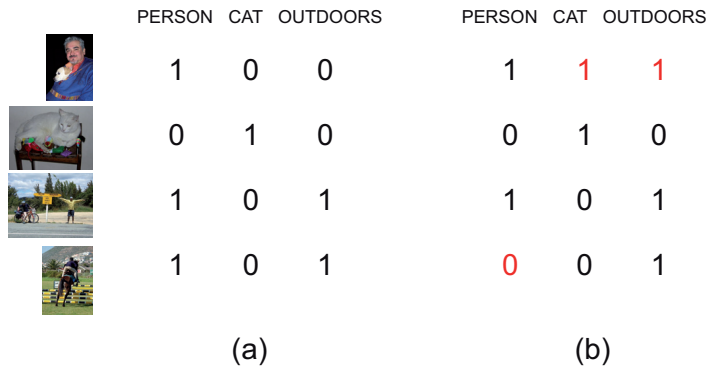


Figure 2.2. An example of binary decision matrices. The matrix in (a) corresponds to a manually specified ground truth. The matrix in (b) represents the output of a category detection system. The system makes three mistakes in its detections.

To help in the forthcoming discussion, two general statistics measuring the quality of retrieval are defined: precision and recall [110]. Suppose a set contains N items and N_{C} of them belong to the category of interest. A retrieval system retrieves a subset of this set of size M and M_{C} of the retrieved items belong to the category. Precision P is defined as the fraction of retrievals that are correct:

$$P = \frac{M_{\mathrm{C}}}{M}. \tag{2.1}$$

Recall R measures how large a fraction of the items belonging to the cat-

egory are contained in the retrieval result:

$$R = \frac{M_{\mathrm{C}}}{N_{\mathrm{C}}}. \tag{2.2}$$

2.3.1 Performance measures of ranked retrieval

In ranked retrieval, the category detection system outputs a ranked list of images. Let the items in this list be denoted as $I(i)$, where the index $i \in [1, M]$ indicates the image's position in the ranking. Index 1 corresponds to the image most likely to belong to the category. In the following it is assumed that the list gives ranking to all images in the collection, i.e. $M = N$.

If the binary ground truth function $\mathrm{GT}(i) \in \{0, 1\}$ of the category membership is given, one can measure the quality of ranking against this ground truth. In the following two popular measures are described in detail: the average precision (AP) measure [110] and the area under curve (AUC) measure [47] extracted from the receiver operating characteristic (ROC) curve. There exists naturally a number of other measures of ranking quality, for example the averaged normalised modified retrieval rate (ANMRR) that is used in the MPEG-7 standard [109], and the normalised average rank ($\widetilde{Rank}$) introduced in [127]. However, these measures are not described in detail here as they have not been used in the experiments reported in this thesis.

The average precision statistic can be derived from the recall-precision curve that describes the degree of agreement between the ranking I and the ground truth GT. This curve can be constructed by considering the sub-rankings that consist of the m first images of the full ranking I. For each of the depth m sub-rankings, one can evaluate the recall

$$R(m) = \frac{1}{N_{\mathrm{C}}} \sum_{m'=1}^{m} \mathrm{GT}(m'). \tag{2.3}$$

and precision

$$P(m) = \frac{1}{m} \sum_{m'=1}^{m} \mathrm{GT}(m'). \tag{2.4}$$

Here

$$N_{\mathrm{C}} = \sum_{m=1}^{M} \mathrm{GT}(m) \tag{2.5}$$

is the total number of relevant images in the collection. The $(R(m), P(m))$ pairs can be interpreted as coordinates of points in recall-precision space. One obtains the recall-precision curve by connecting the points that result

from letting m sweep the range $[1, M]$. The AP statistic is defined to be the average of the precision values of those sub-rankings whose last element $I(m)$ belongs to the category of interest, i.e. $\mathrm{GT}(m) = 1$. AP closely approximates the area under the recall-precision curve. It can be can be evaluated using the formula

$$\mathrm{AP} = \frac{\sum_{m=1}^{M} \mathrm{GT}(m) \cdot P(m)}{N_{\mathrm{C}}}. \tag{2.6}$$

The above definitions can be modified slightly so that the recall-precision curve is forced to be non-increasing [110]. The statistic giving area under the modified curve is called the interpolated average precision:

$$\mathrm{AP}_{\mathrm{int}} = \frac{\sum_{m=1}^{M} \mathrm{GT}(m) \cdot \max_{m' \geq m} P(m')}{N_{\mathrm{C}}}. \tag{2.7}$$

It is also common that the index m does not sweep every image in the ranked list, but is stepped to include only precision values occurring at specific levels of recall R. For this kind of sampling, it is better to use the interpolated measure, since interpolating smoothens the local variations in the curve, making the sampled measure less dependent on the alignment of the sampling points with local recall-precision curve details. For example, some of the recent VOC evaluations (cf. Section 2.4.2) measure ranking quality with an interpolated average precision measure sampled on eleven evenly spaced recall levels. Sometimes it is not feasible to obtain the ground truth GT for the whole test set but only for a sample. One can still estimate the AP rather accurately, as is done in the case of the inferred average precision (infAP) measure [212] that has lately been used in the TRECVID evaluations.

AP measures the ranking quality in the case of a single category that is detected. In the case of multiple detected categories, one often averages the AP values over the categories. One then speaks of mean average precision (MAP) [110]. Similarly, mean inferred average precision (MIAP) results from averaging infAP values.

The AUC statistic derives from describing the ranking quality with a different kind of a curve. The receiver operating characteristic (ROC) curve can be constructed from the ranked image list by once again considering the sub-rankings including the first m images of the list. In this case, one interprets the fraction of false positives

$$\mathrm{FP}(m) = \frac{m - \sum_{m'=1}^{m} \mathrm{GT}(m')}{M - N_{\mathrm{C}}} \tag{2.8}$$

and the fraction of true positives

$$\mathrm{TP}(m) = R(m) = \frac{\sum_{m'=1}^{m} \mathrm{GT}(m')}{N_\mathrm{C}} \tag{2.9}$$

as coordinate pairs. The ROC curve results from connecting the points when m sweeps the interval $[1, M]$. The area under the ROC curve

$$\mathrm{AUC} = \frac{\sum_{m=1}^{M}(1 - \mathrm{GT}(m))\mathrm{TP}(m)}{M - N_\mathrm{C}} \tag{2.10}$$

can be used as a summary statistic that indicates the ranking quality.

Given a recall-precision curve, the corresponding ROC curve can be constructed and vice versa. However, the same does not apply for the summary statistics. There is no one-to-one correspondence between the AP and AUC measures.

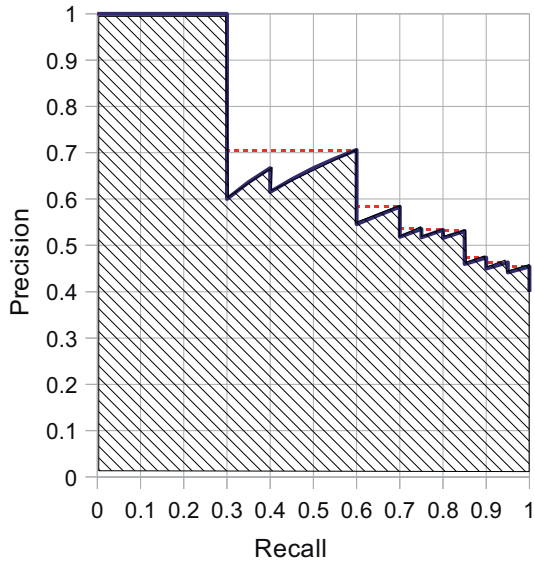
Figure 2.3 shows an example of a ranked list and the corresponding recall-precision and ROC curves. In this example 20 of the retrieved images do belong to the category of interest and 30 do not, resulting in an a priori probability of 0.400. In the ranked list (a), the images are judged to be either correct (C) or incorrect (I) retrievals. The leftmost letter corresponds to the highest ranked retrieval result. In the subfigure (b) the solid line corresponds to the non-interpolated recall-precision curve and the dashed line to the interpolated version. In the subfigure (c) the diagonal dashed line corresponds to the expected ROC curve of purely random retrieval. The diagonally hatched areas under the curves correspond to the AP (b) and ROC AUC (c) statistics. For this ranking, the value of the non-interpolated AP is 0.707 and that of the interpolated AP is 0.720. The ROC AUC has value 0.740. The ROC AUC value 0.500 of a random retrieval defines a zero-level for this statistic, from which any useful ranking method must differ significantly. Similarly, the a priori probability of a category can be used as a zero-level value for AP.

2.3.2 Measures of binary decision matrix quality

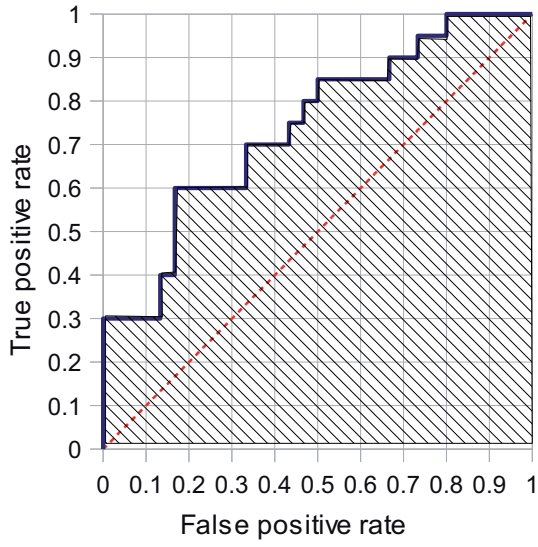
In this thesis, the binary decision matrix output format is used mainly in the image annotation application. The measures are thus discussed in greater length in Section 4.1 of the thesis that demonstrates the use of the PicSOM category detection system in that particular application. Only a brief overview of decision matrix quality measures is given here because the practical examples presented in Section 4.1 will support the in-depth discussion on the measures' properties there.

CCCCCCIIIICCICCCCIIIIICCIIICICICIIIIICIICIICIIIII

(a)



(b)



(c)

Figure 2.3. An example of a ranked list (a). The corresponding recall-precision (b) and ROC (c) curves are plotted with blue solid line. The hatched areas under the curves correspond to the AP and ROC AUC summary statistics.

In this thesis, the following binary decision matrix measures are used:

1. accuracy percentage,
2. average normalised score (NS),
3. precision-recall statistics (PR),
4. de-symmetrised termwise mutual information (DTMI).

Common to all the measures above is that larger numerical values correspond to better retrieval accuracy. All the measures attain their maximum values when and only when the decision matrix exactly matches the ground truth.

Accuracy percentage is the simplest of the measures. It tells the fraction of the decision matrix entries that match with the ground truth. The measure is most useful in the case when there is only one category to detect. In contrast, the normalised score is meant to be used in multi-category cases. The score for an image (row in the decision matrix) is obtained by a weighted sum of the number of correct detections and false alarms. The weighting scheme ensures proper normalisation when the average over all the images is taken, so that images exhibiting different number of categories in the ground truth are weighted equally in the average.

The PR statistics denote the averages of precision and recall values, these values first being determined separately for each category (column in the decision matrix). Here the non-zero values in the column indicate that corresponding image (row) belongs to the set of retrieved images. Precision and recall can be combined to a single statistic—the balanced F measure—by geometric averaging (see page 104).

In both the NS and PR measures different decision matrix entries receive different weights, but the weighting is somewhat arbitrary. In the DTMI measure, proposed in Publication III, the weighting is set on information theoretic grounds. The measure is essentially a modified version of mutual information between the entries of the decision matrix and the ground truth.

2.4 Benchmarking category detection systems

In this thesis, the empirical point of view is taken for justifying visual category detection techniques and systems. This means that techniques are

compared on the basis of how well they perform in category detection experiments. The principle is straightforward, but there are still difficulties in practice.

The category detection problem itself entails so many variable components that evaluations employing different experimental setups are not comparable. To obtain meaningful results, one has to compare a series of alternative techniques in the same setup. However, typically preparing such a setup is laborious. Realistic experimental visual category detection setups must involve large numbers of images or videos, making the data collection and preparation a considerable effort. One must also not forget the effort that goes into the careful planning of the experimental task to avoid all sorts of biases and other undesired effects that might result from improper choice of data.

Also the execution of the experiments can be tedious, both in terms of computational complexity and the researchers' effort involved. As a consequence, the experiments of a single research group rarely can cover all the aspects of the multi-component visual category detection system extensively enough so that a comprehensive picture could emerge of the system operation as a whole. Co-operation of research groups with the same benchmark data sets can therefore be very fruitful, giving the possibility to implement a more diverse set of experiments.

One additional argument speaking for the cross-group co-operation with benchmark data is that of impartiality. Often researchers are willing to see that a technique developed by them outperforms other techniques in experiments. Therefore, if the researchers themselves implement and execute also the methods that they compare against, there might exist the risk of bias towards their own method, intentional or unintentional. Rarely are the methods so parameter-free and their application independent of implementation choices that there would be no tuning involved in the process. It is not unthinkable that the researchers' own method receives more tuning effort, or the researchers could simply understand the workings of their own method better than that of the others. Common benchmark problems eliminate this issue as researchers all implement their own methods. Related issues, also addressed by common benchmarks, are the choice of the problem and reference methods. In isolated experiments, the choices can almost always be made to favour the proposed method, possibly totally unintentionally. In a reasonably popular benchmark problem that is addressed with many different methods this

issue can be avoided.

The above arguments show that benchmark experiments using common experimental setups are clearly advantageous. But even when using such common setups, interpreting the results of benchmark experiments is not always straightforward. Section 2.4.1 discusses some of the issues that have to be taken into account. After discussing benchmarking in a general level, Section 2.4.2 lists a selection of widely used benchmark data sets and evaluation campaigns for visual category detection. The list is in no means exhaustive. Rather, it includes benchmarks for which the PicSOM system has been applied.

2.4.1 Challenges in interpreting benchmark results

It often happens that some experimental results are contradictory. With the benchmarks evolving all the time, it is difficult to keep the experimental setup so controlled that every factor causing the discrepancies could be identified with certainty. Therefore, one should take the results of any single experiment with caution. One can be more certain of the beneficiality of a certain technique if it consistently outperforms other alternatives in slightly varying experimental setups by independent researchers.

Category detection systems are complex, consisting of multiple components that are not isolated from each other. As an example of the dependencies between the components, it would be possible that some given supervised learning technique is compatible with one particular fusion technique (cf. Chapter 3). However, some other fusion mechanism may work better with another learning algorithm. If one considered only the first of the learning techniques, one might be mislead to the conclusion that the former fusion technique is in general better than the latter. This is not to say that component-level testing would not be important. The results must always be interpreted with caution, however. System-level tests must be performed from time to time, evaluating different combinations of techniques. In practice there seems to be quite much diversity in the techniques researchers have included in their systems participating in the benchmarks. Therefore, one can at least hope that among the different alternatives, the best-performing techniques finally emerge as winners.

There exists also the opposite danger. If one looks only how well complete systems work, it may be difficult to deduce which of the system com-

ponents actually work exceptionally well, and which components perform only on the average level, without compromising the overall system performance too much. Very seldom is a situation encountered where all the components of a system would be essentially better than their competitors.

A related challenge is that component-level conclusions may lose their validity as the system performance improves drastically. For example, fusion technique A might be clearly better than technique B when the supervised detectors operate with 1% accuracy level. However, it is not self-evident that A would still be better than B with detector accuracies in order of 10%.

2.4.2 Examples of benchmark settings

TRECVID video retrieval evaluation

TRECVID [167] is an annual video retrieval evaluation campaign and workshop series organised by the National Institute of Standards and Technology (NIST). It is arguably the leading venue for evaluating research on content-based video analysis and retrieval. TRECVID originates from the video track organised in conjunction with TREC information retrieval conference series in years 2001 and 2002. Since 2003, TRECVID has been an independent workshop. The TRECVID campaign provides the participating organisations large training and test video collections, uniform scoring procedures, and a forum for comparing the results. Lately, TRECVID has attracted several dozens of research groups to submit their results in the evaluation each year.

Yearly, the TRECVID evaluation defines a set of video analysis tasks, such as high-level feature (i.e. concept or category) extraction, video search, video summarisation, and content-based copy detection, along with a video collection. Of the tasks of recent years, the high-level feature extraction (HLFE) has been the task that measures performance in supervised concept detection. The other defined tasks are less relevant to this thesis on category detection, although it has turned out that successful concept detection is an essential prerequisite to the video search tasks. Video search in the TRECVID setting will thus be discussed in Chapter 4 that shows some application examples of concept detection. The TRECVID evaluations have measured performance in the HLFE and search tasks by mean average precision (MAP) or alternatively its

Figure 2.4. A sample TRECVID search topic. Adapted from Publication VI.

sampled version MIAP (cf. Section 2.3.1).

The video collection used in TRECVID has evolved over the years. In 2003–2006 different compilations of captured broadcast news videos were in use. First all the videos came from English speaking television channels, later broadcasts in also Chinese and Arabic language were included. In 2007–2009 the type of video material used in TRECVID consisted of documentaries, news reports and educational programming from Dutch TV. The 2010–2011 evaluations are based on videos from the Internet Archive.

The TRECVID video data has always been divided into separate development and test sets. Parts of the video material have been re-used from year to year. For example, the test material from 2003 was re-used as the labelled training set in the following year. In order to be able to compare the concept detection results of different participants, a common reference segmentation of videos temporally into shots has been provided by the organisers [141]. A shot in a video is a sequence of frames captured continuously by one camera. In TRECVID, concepts have been detected on the temporal resolution of one shot. The TRECVID data volume has slowly increased over the years. In 2003, the data consisted of approximately 130 hours of news broadcasts. In 2009, the combined volume of training and test collections was about 380 hours, which translates to approximately 130 000 shots. To obtain training data for the high-level feature extraction task, collaborative annotation efforts [10] have been organised where the TRECVID participants label the training videos. As an illustration of the type of concepts involved, Table 2.1 lists the high-level features that were to be detected in 2009.

Due to the size of the test corpora, evaluation of the concept detection results represents a challenge. It is infeasible within the resources of

Table 2.1. The 20 concepts detected in TRECVID 2009 high-level feature extraction task. Adapted from Publication VI.

Classroom	Person playing a musical instrument	Hand
Chair	Person playing soccer	People dancing
Infant	Airplane flying	Nighttime
Traffic	Person riding a bicycle	Boat or ship
Doorway	Female human face closeup	Telephone
Cityscape	Person eating	Singing
Bus	Demonstration or protest	

the TRECVID initiative to perform an exhaustive examination of the test data in order to determine the concept-wise ground truth. Therefore, a pooling technique is used instead. First, a pool of possibly relevant shots is obtained by gathering the sets of shots returned by the participating teams. These sets are then merged, duplicate shots are removed, and the relevance of only this subset of shots is assessed manually. It should be noted that the pooling technique results in the underestimation of the performance of new algorithms and, to a lesser degree, new runs, which were not part of the official evaluation, as all unique relevant shots found by them will be missing from the pooled ground truth.

Apart from the slight performance evaluation bias, there exist some other issues with TRECVID. The TRECVID evaluation is not totally open. The video collections have been available only to the participants that have officially registered to NIST and participated in the evaluation competition. The collaborative annotations have been accessible only to research groups that have participated in the annotation effort.

Another issue is that the benchmark has not been kept stable over the years. The video collections and the sets of concepts to be detected have constantly been evolving. This naturally must happen so that innovation and new developments can be supported. At the same time, comparing different approaches presented in different years is difficult as the results are not comparable over the years. The high data volumes involved make this issue more serious by the sheer computational effort required for testing a method, originally applied to one year's task, on the task version of another year, even if all the needed data sets are available. It can be argued that the tediousness of transferring results between years results in wasteful use of computer and researcher resources. On the other hand,

because of the tediousness there is the obvious temptation not to re-apply an old method to new data, leading to less systematic comparison of category detection techniques.

As an example of the above problem, one can look at the results of the semantic indexing task of TRECVID 2010. The task was essentially similar to the HLFE task of TRECVID 2009, with the exception of a new, larger data set and a larger lexicon of concepts to be detected. The processing of this larger data set even with readily existing methods was such a laborious effort that not much time and other resources were left for developing and testing new techniques. The best performance in the 2010 task was obtained by applying again exactly the same system that provided the best performance in 2009 [171]. Also the PicSOM group managed to improve its relative performance by using only a subset of methods of 2009, although with the addition of improved multi-keyframe analysis [164].

The MediaMill challenge [173] is an attempt to partially overcome the issues related to the evolving nature of the TRECVID benchmark. It has fixed the used video collection to be 85 hours of international broadcast news videos from the TRECVID 2005/2006 benchmark. Furthermore, it provides a stationary set of 101 concepts to be detected. The challenge also addresses the issues related to the large workload of implementing and running a whole category detection system capable of the TRECVID HLFE tasks. Instead of having to implement all the system components, the challenge defines a set of experiments that address the separate components of a category detection system. Pre-computed reference implementations run on the challenge data can be used to simulate the other system components. Despite these advantages, the MediaMill challenge seems not to have attracted very large attention in the research community. The researchers tend to be more willing to evaluate their methods each year with the fresh edition of TRECVID tasks.

PASCAL Visual Object Classes

The Visual Object Classes (VOC) Challenge [45] and its related image collection, by the European PASCAL Network of Excellence, has become the de-facto standard benchmark in the area of object detection and localisation in still images. Since 2005, the challenge has been organised yearly. The VOC Challenges have not attracted quite as large a number of participants as the TRECVID campaign, but still over a dozen research groups have been submitting their results to the challenge each year (with an

Figure 2.5. Examples of VOC Challenge 2007 images and their annotations with bounding boxes. From Publication II.

exception of the starting year).

The challenge started with four object classes that had to be recognised and localised in the collection of 1578 photographic images. Since then, the number of object classes has increased gradually to 20 and also the image collection has grown to 19 740 images by 2010. The organisers manually annotate the images with bounding boxes of the objects to be detected. This centrally organised annotation procedure arguably results in more consistent annotation quality and criteria than e.g. the community annotation model employed in TRECVID. In recent years, some of the images have also been annotated with pixel-wise object segmentations. The challenge organisers partition the images into development and test sets, and specify also a suggested further partitioning of the development set into training and validation halves. The annotations for the development set are made public.

From the beginning, the challenge has included two tasks: "classification" and "detection". Classification concerns the decision whether any instance of the specified object class is visible anywhere in an image. In the object detection task, the goal is to detect bounding boxes of the objects. In recent years, two smaller-scale competitions (called "tasters" by the organisers) have also been included in the challenge: pixel-level object segmentation and person layout detection.

The VOC benchmark has been designed to be open, i.e. the image collections and annotations for the development sets are publicly available to anyone. This has made it possible for numerous researchers to apply their methods to these standardised image collections and tasks even if they have not taken part in the challenges. Figure 2.5 shows example images from the VOC 2007 Challenge collection along with their annotations. Table 2.2 lists the object classes that were to be detected.

Table 2.2. The 20 object classes of VOC Challenge 2007

aeroplane	bus	dining table	potted plant
bicycle	car	dog	sheep
bird	cat	horse	sofa
boat	chair	motorbike	train
bottle	cow	person	tv/monitor

Corel images

The third example of a benchmark data set differs from the two previous examples in that it does not originate from a performance evaluation competition. The images from a commercial Corel illustration image collection have been convenient material for researchers to demonstrate their automatic image annotation systems since the images come with annotating keywords, five keywords per image at most. The images have been packaged as different commercial products. In total, more than 80 000 images exist. Many researchers first demonstrated their methods by using various subsets of the images and keywords, leading to incomparable results as demonstrated in [127]. However, the experimental setups of [43] and [11]—especially the former with 5 000 images—have later risen into the position of a de-facto standard benchmark where many researchers, including the current author, have tested their methods.

Another possibility of benchmarking stems from the organisation of the Corel image collection. The images come on thematic CDs, each containing 100 images exhibiting that theme. With the themes as category labels, one conveniently obtains a labelled data set with thousands of images. This benchmark setup has been particularly popular among the multiple-instance learning (MIL) researchers for demonstrating their systems [8, 31, 148]. Figure 2.6 shows examples of the Corel images and their annotations. Table 2.3 lists some of the thematic categories.

A practical problem in using the Corel images as benchmark is that the images are commercial and thus not freely distributable. Furthermore, the product series is nowadays discontinued and the images cannot be purchased any more.

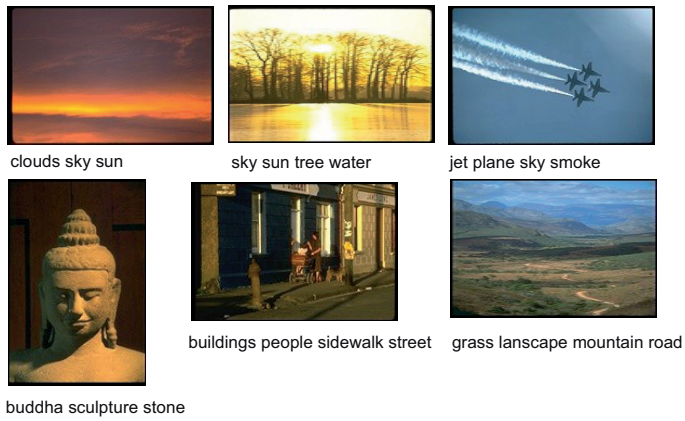


Figure 2.6. Examples of Corel images and their annotations.

Table 2.3. 20 of the thematic categories of Corel images.

Beach	Cars	Historical buildings
Buses	Dinosaurs	Elephants
Flowers	Horses	Mountains and glaciers
Food	Dogs	Lizards
Fashion	Sunsets	African people and villages
Waterfalls	Antiques	Battle ships
Skiing	Desert	

2.5 Some related work and historical background

Computer vision problems have been intensively studied from almost the advent of digital computers. The first early computer vision systems appeared during the sixties. Since then, a huge volume of computer vision literature has naturally appeared. There is no possibility nor desire to extensively review the literature here. However, some interesting historical developments are mentioned. The discussion here largely follows [54], that provides a thorough review of the historical development of object and category recognition.

In the early work, the focus was on detecting a particular object instance in images, possibly in a slightly varying imaging geometry or illumination conditions. Only later were the tasks generalised to concern categories of objects with intra-category variance. Much of the early work during 70s and 80s (e.g. [2, 16, 20]) was based on range images, providing a straight-

forward route to three-dimensional observations of the world. Then the attention was turned increasingly towards ordinary two-dimensional images. According to [54], however, during the 80s many approaches limited the imaging conditions and geometry so that the structure of the images would be easier to interpret, for example by imaging objects on a light table or against a uniform background. By the 90s there was progress towards more and more natural imaging environments and broader object categories instead of individual object instances.

Many of the early methods for specific object recognition were based on constructing a precise 3D-model of the object and then fitting this model to the image. The approaches often relied on the geometric properties of the objects. This made the methods invariant to illumination changes and possibly also to affine and projective transforms of images. Such methods were typically invariant also to the texture of objects, making the methods very well suited for some objects well described in terms of geometric forms (such as bottles), whereas some other object types (such as faces) could not be satisfactorily handled.

Prominent types of early geometric specific object detection methods include alignment techniques and techniques based on geometric invariants. In alignment methods (e.g. [73, 107, 184]), a set of possible transformations is first hypothesised that could project the object model into the image, based on lines and points detected in the image. After this, the space of tentative transformations is searched to find transformations that receive the most support from the image. In geometric invariant methods (e.g. [96, 150, 151, 152]), the objects are modelled as a collection of primitive geometric structures, such as sets of points, lines or conics. They are described in a way that is invariant to some suitable geometric transformations, such as affine transformations. During recognition, similar geometric primitives are detected from the images. The detected primitives then vote between various candidate object models. The procedure can be interpreted as a visual counterpart of text hashing.

In addition to geometric analysis of images, the object and category recognition has also been approached using methods based on either global image appearance or the appearance of local image regions. The global appearance methods have been applied since the late 90s. In those methods, exact models of the possible contents of the images are not explicitly prepared. Instead, one attempts to exploit the correlations between object classes and easily computable image properties. In the

beginning, many methods (e.g. [158, 177]) were based on simple global image statistics such as colour histograms, other properties of the colour distribution, and statistics of the image texture measured by Gabor filter outputs. Some other methods (e.g. [128, 183]) projected the pixel patterns of the images into low-dimensional subspaces, along the lines of early work in [90], facilitating straightforward machine learning based on the projections.

One particular application area of appearance-based image content analysis has been in interactive content-based image retrieval (CBIR) systems. In such systems—introduced first in the early 90s—a user interactively performs queries in image databases. The system answers the queries on basis of the database images' elementary visual properties. Examples of influential early CBIR systems are QBIC [57], Photobook [139] and MARS [72]. Also the PicSOM architecture—the subject of this thesis—was originally devised with the CBIR application in mind.

In methods based on the appearance of local image regions, one first selects a set of interesting image regions and then describes each region with a descriptor that characterises the appearance of that region. Depending on the implementation of the region detector, the regions can be very local or have a considerable spatial extent. The local region approaches share the idea with the earlier specific object recognition methods based on geometric invariants in that the image is described as a collection of primitives. However, the local appearance based methods generalise the idea by allowing a larger variety of descriptors of local appearance, such as the texture of the regions. The last 15 years have witnessed enormous progress in the field of object and category detection based on local appearance methods. There is a large body of work of proposing both different types of region detectors and descriptors. Some of these will be discussed in Section 3.2.2.

First the local appearance based methods were employed for specific object detection, an early example being [159]. A landmark paper in the area is [108]. There the image locations of interest are detected as scale-space maxima of the difference-of-Gaussians (DoG) operator. The paper proposes the regions around interest points to be described with SIFT (scale-invariant feature transform) descriptors which are histograms of local image intensity gradient directions. In that paper, the focus was on specific object recognition. Later on, similar techniques have been used very successfully also for detection of object and scene categories. In cat-

egory detection, one key to success has been to aggregate the statistics of local appearance properties in order to form a global image feature. [34] was a pioneering work in implementing this idea in a category detection system.

It might not be intuitively self-evident that the global statistics of local image appearance would correlate strongly enough with practical object and scene categories in images. One could imagine there to be many interfering effects: the effect of background against which the objects of interest appear, variations in imaging geometry and illumination, occlusions and effects of distracting objects that might appear in the images. Practice has shown, however, that often the correlations indeed are adequately strong. Apart from the object composition of images, the approach is also suitable for recognising more general visual image categories such as “nighttime” or “outdoors”. As can be seen from the experimental results presented in this thesis (the results of Section 3.2.4 in particular) the local appearance based methods—also known by the name bag of visual words (BoV)—form the backbone of the performance of the modern visual category detection systems.

The appearance-only models work surprisingly well for category detection. Intellectually this seems to be inconvenient to many. Intuitively it is clear that also spatial information should be taken into account, instead of regarding the images just as orderless collections of appearance primitives. Recently, there have been many attempts to combine geometric information with the models of local appearance, and not completely without success. For example, [22, 48, 54, 205] re-visit and refine the constellation model that was originally proposed already in the 70s [55]. Constellation models explicitly model the geometric ordering of parts of objects, which in turn are described in terms of their local appearance. More indirect approaches—briefly discussed also Section 3.2.2—are proposed in [99, 112].

3 System for visual category detection

This chapter discusses the ways of implementing a visual category detection system. Section 3.1 first lays out the overall system architecture of the PicSOM category detection framework that consists of several subsystems. This kind of system architecture has emerged over the years from the experiments of numerous researchers in the research community as the practical consensus on a feasible way for performing category detection. The PicSOM system has in essence implemented this system architecture throughout its whole history of more than ten years.

After defining the overall structure of the category detection framework, Sections 3.2 through 3.4 then discuss the components of the framework in detail. Included are descriptions of various alternative techniques for implementing the components, along with discussion of a body of experiments that compare the different techniques and thus justify the use of some of them. Section 3.2 starts with feature extraction techniques, followed by Section 3.3 on supervised detection. Finally, the techniques for fusion are covered in Section 3.4. Early and late fusion are discussed together, even though in the following prototypical system architecture there are separate modules for both. There are several reasons for this. Both early and late fusion can be seen as alternative means for achieving the same goal. The whole division is also somewhat artificial, as fusion can occur in any intermediate stage of a category detection system.

3.1 Architecture of a category detection system

This section outlines the overall architecture of the PicSOM category detection framework. In the architecture, visual categories are detected one-by-one, independently of each other. For each of the categories one wants to detect, the detection is performed in supervised mode: the system is first trained using a set of images for which it is known which of the im-

ages belong to the category of interest. After training, the system is ready to be used for predicting whether the images of a test set belong to the category. The following discussion reverses the time order of the actual system usage and starts from the testing phase in order to first introduce the parts the system. After that, the training of the system components is briefly outlined, followed by some remarks of the general level properties of the architecture.

3.1.1 Workflow in testing phase

Figure 3.1 schematically shows the workflow in the category detection system when it is used for deciding how likely a given test image is to belong to a pre-defined category. The system operation in the testing phase comprises of four stages:

1. feature extraction,
2. early fusion,
3. feature-wise supervised detection,
4. late fusion.

The processing starts by extracting a set of N_{F} low-level visual features $\{F_i\}_{i=1}^{N_{\mathrm{F}}}$ from the test image. In practice, the representation for feature F_i of the image is a k_i-dimensional feature vector of real values. One of the extracted features could be, for instance, the colour histogram of the image. The set of low-level features can be augmented in the early fusion phase. In early fusion, a set of N_{E} new features $\{E_i\}_{i=1}^{N_{\mathrm{E}}}$ are synthesised by combining two or more low-level features. This can, for instance, be accomplished by concatenating the respective feature vectors.

The third stage in the architecture is feature-wise supervised detection. In this stage, each of the $N_{\mathrm{D}} = N_{\mathrm{F}} + N_{\mathrm{E}}$ supervised detectors maps the corresponding feature vector into a partial detection score $s_i, \;\; i = 1, \ldots, N_{\mathrm{D}}$. Any supervised vector-space decision estimation algorithm can be used for implementing the detectors, the supervision being provided by the training phase of the category detection system. Support vector machines (SVM) [33] are a popular and well-performing choice. In the late fusion stage, the feature-wise detection scores $\{s_i\}_{i=1}^{N_D}$ are finally combined into a single score s, for instance by computing their average. This score is the final output of the category detection system. It reflects the estimated

likelihood of the test image to belong to the category the system has been trained to detect.

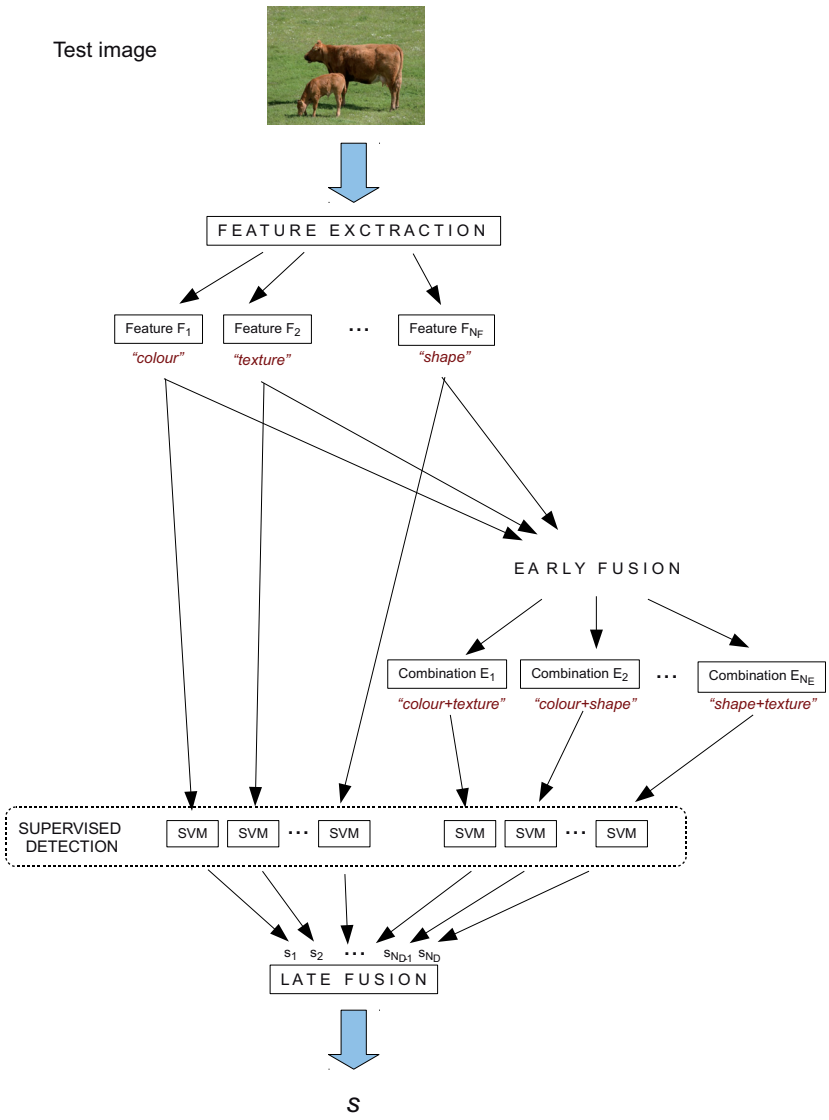


Figure 3.1. The test phase workflow in the system for detecting categories of images.

3.1.2 Training phase

The training phase of the category detection system sets the internal parameters of the system components to such values that good performance can be expected when the category membership of a test image is decided. On the general level, this is done by looking at the properties of the training images whose category memberships are known and selecting such parameter values that would realise the estimated mapping from the properties to the values of category membership.

In the architecture under discussion, the feature extraction phase is considered to be pre-determined. The training phase starts by extracting the same set of N_{D} features and their early fusion combinations from each of the training images as will be eventually done in the test phase.

A major task of the training phase is to prepare the feature-wise detectors. This is a prototypical supervised classification task: for each feature, one has the feature vectors of all the training images and their binary category membership labels. The learning algorithm is to estimate the feature–label mapping as well as it can. Any vector space learning algorithm can be used for this purpose.

Another training phase task is the setting of internal parameters of the late fusion stage if the fusion mechanism is supervised, i.e. contains any free parameters. In order to estimate the optimal parameters, one needs to apply the feature-wise detectors for the features extracted from the training images, resorting to cross-validation techniques for obtaining unbiased estimates.

3.1.3 General discussion of the overall architecture

Characteristic to this framework is that several different visual features are extracted from images. Analysis is first performed individually on basis of each feature, after which the results are combined in a fusion step. A further characteristic of the framework is that the analysis is generic, not specialised to any particular visual category. This genericness results from the feature–category mappings being automatically learnt from the provided example images during the training phase.

When using the framework, the properties of images are summarised with a rather compact set of visual features. Inevitably lots of visual information is lost when moving from the images to their feature represen-

tations, and consequently every single piece of information from each and every image is not analysed as carefully as would be possible. This is counterbalanced by the fact that the when using this framework, one is able to exploit the statistical dependencies of visual features in very large collections of images as the per-image processing effort is kept modest.

The framework is not limited to analysing just images, even though they are shown as the data objects in Figure 3.1. Any object type can be handled, as long as meaningful fixed-dimensional numerical feature vectors can be extracted to represent the objects. In addition to still images, the framework can be used for object types such as sub-parts of images, video footage, text and audio documents, and multimedia messages sent from mobile phones [163].

The justification for the outlined architecture is empirical. The architecture—in its basic form or with some extensions—is behind almost all successful category detection systems that have recently been used for tackling international benchmark evaluations and benchmark data sets, including the TRECVID and PASCAL VOC Challenge benchmarks discussed in Section 2.4. Accordingly, the architecture has gradually diffused in the category detection systems of researchers. It is safe to say that this development has occurred during the last 15 years, with the emphasis on the most recent years.

3.2 Feature extraction

Historically, in image analysis one has first tried to recognise a very limited range of objects from constrained images, a specific item in different poses in a known environment, for example. The recognition can in this case be implemented by exactly modelling the geometry of the objects to be recognised. However, exact models of the possible contents of images would become impossibly complex when the images become more generic. With the added temporal dimension, modelling contents of videos would be even more complex.

One may well ask, is it really necessary to infer all the geometric details of an image in order to decide whether it belongs to a certain category. In appearance-based approaches to image analysis, a stand is taken that much can be said about an image by analysing the correlations between image category labels and easily extractable features of the image

appearance, such as simple summary statistics. This idea is supported by the psycho-visual observation that humans can recognise scenes holistically, without needing to recognise the individual objects the scene consists of [14]. The category detection framework of this thesis takes a viewpoint on features in the appearance-based spirit: any type of feature can be accommodated as long as it is reasonably easy to extract from numerous data objects, and possibly bears a correlation with categorisations of interest.

In the discussed category detection system architecture, feature extraction is the interface to data objects that has to be adapted according to the particular type of the objects. For example, if one has to detect categories of images, then the feature extraction stage must implement methods for extracting visual features out of images. Only the feature extraction stage needs to be different in the system if one decides to detect categories of data objects of a different type instead, for instance categories of audio documents.

Feature extraction is also the system component that can be adapted to a specific narrow domain within a data object type. For instance, if a collection of face images are to be handled, a specialised set of face recognition features can be extracted that discriminate well between the faces of different persons.

In the remainder of this section, a look is taken at features that are applicable in generic image (Section 3.2.1) and video domains (Section 3.2.3). An exhaustive review of feature extraction is not even attempted, due to high volumes of existing work on the subject. Rather, some interesting and important feature extraction technologies are mentioned, with emphasis on methods that have been used in the PicSOM category detection system. Section 3.2.2 discusses a particular type of image features: the BoV features. The reason for devoting a separate section for the BoV features is two-fold. Firstly, the detailed investigations of BoV features form a considerable part of the material of the publications of this thesis, and Section 3.2.2 summarises the main results of those publications and sets them into context of some related work. Secondly, in many experiments in the literature the BoV feature extraction paradigm has proven to provide very competitive categorisation accuracy when compared against other feature extraction techniques. In the empirical spirit of this thesis, the final section on feature extraction puts the discussed feature extraction techniques in a practical test bench: Section 3.2.4 reviews feature extrac-

tion methods employed in the most successful systems of the high-level feature extraction task of TRECVID 2009 evaluation.

3.2.1 Image features

Local image properties

A global image feature characterises an image as a whole. In the current category detection architecture, all the image features must be global if one wants to detect the categories on the level of whole images. On the other hand, image feature extraction necessarily has to consider local image primitives, pixels being the most extreme example of locality. Image feature extraction therefore is a process where a global image feature is aggregated from the images' local features. Simple examples of local properties include pixel colour and outputs of simple edge detectors, such as the Canny edge detector [25].

In addition to the RGB colour space natural to most computer representations, one can consider also other colour spaces, in which the numerical colour coordinates possess more desirable properties. For example, the Munsell colour space [122] is designed to be perceptually as uniform as possible. Also CIE L*a*b* [32] and HSV [70] colour spaces are perceptually more uniform than the RGB space. CIE L*a*b* is an example of an opponent colour space where the effects of illumination are isolated into a separate component of the representation so that invariant colour descriptors are easier to synthesise. Other opponent colour spaces are those used in [64] and [185].

Besides colour, another local image property that can be characterised is the texture. This can be done by means of filtering (steerable filters [59] and/or wavelets [37], Gabor filters/wavelets [19, 102] in particular) or by local comparison of binarised grey levels of the image [136], for instance. Local regions may also have a definite spatial extent. In this case the shape of the local area can be described, e.g. via Zernike moments [85] or Fourier descriptors of the contour [9].

Aggregation of local features

There are several possible mechanisms for aggregating local features within an image in order to end up with a global feature. This can be interpreted as a prototypical data fusion problem. It is therefore somewhat arbitrary which of the aggregation techniques are discussed here in the

context of feature extraction and how much of the material is postponed to Section 3.4 that discusses fusion more generally in a category detection system.

One possible aggregation approach is to regard the values of the local features at different image locations as samples coming from a distribution of feature values. One can then characterise the distribution. A straightforward method is taking the average of the property, but also higher order moments of the distribution can be considered, as is often done with the colour values [176]. The distribution of the property can naturally be characterised by other means, too, e.g. by choosing a fixed number of representative values from the distribution or by forming a histogram of the distribution. The histogram aggregation can be modified to employ soft assignment (cf. page 66), resulting in better performing features. In [63] it is demonstrated that natural image texture statistics follow the integrated Weibull distribution. It is thus possible to represent the image with the parameters of the distribution, as is done in the Wiccest features of [62]. The distribution can also be modelled as a mixture of Gaussians. It is possible to estimate part of the parameters using all the images in the image collection, and adapt part of the parameters separately for each image, as is done in the very efficient SIFT-Bag kernel method [217]. In addition to characterising the spatially orderless distribution, one can also calculate statistics that take the spatial correlations of feature values into account, such as autocorrelograms [71] and Markov stationary features [103], which are a set of statistics computed from the autocorrelogram.

In the above discussion the local features were taken to be aggregated over the area of the whole image. However, a common alternative is to geometrically partition the aggregation area into parts and describe each sub-part separately. The case where the description of sub-parts is based on orderless statistics within the regions is denoted as “divide-and-disorder” in [100]. Such approaches have been popular in computer vision over the years (e.g. [65, 178, 182]). A rigorous formulation of the idea is given in [88], where one defines the concept of locally orderless images. There the idea is implemented in the form of a set of Gaussian apertures in various scales and locations over the image, with a histogram of image features aggregated within each aperture. In that work, the locally orderless images are claimed to have an important role in perception. Within the “divide-and-disorder” paradigm, a question still remains,

how to exactly partition the image area and how to utilise the features of sub-images for describing the whole image. This issue is touched in Publication IX and also in Section 3.2.2 below. A straightforward alternative is to partition the image area into a set of non-overlapping tiles with a tiling pattern, such as a rectangular grid. The features are aggregated from each tile separately, and then the feature vectors of the areas are concatenated. Such features can be called grid features. Figure 3.2 shows an example of a tiling mask that is used in the PicSOM system for feature extraction. In [99] the idea is extended by using several tiling masks in different scales in parallel. Extraction of practical features may involve several aggregation and mask-concatenation steps organised in a hierarchical structure so that the first layers operate on small sub-regions of the image. The subsequent layers accumulating information from ever larger part of the image, the final layer combining information from the area of the whole image.

3.2.2 BoV features

The Bag-of-visual-words (BoV) paradigm implements the general framework of image feature extraction that was outlined previously. The BoV paradigm is nowadays the most successful feature extraction paradigm for visual category detection, as will be demonstrated in Section 3.2.4. In the following, the general BoV principle is first described. Then aspects of the BoV are discussed that have been the subject of experimental investigations of Publications VII–XI. In the following, the results of these experiments are interleaved with the discussion of results of other researchers. The publications are explicitly mentioned when the contributions of this thesis are discussed.

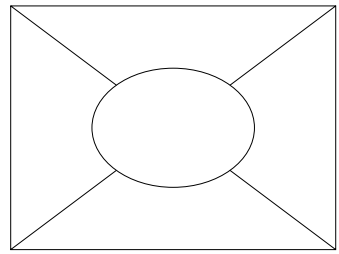


Figure 3.2. The 5-part centre-surround mask (cs-5) that is used in many features of the PicSOM system.

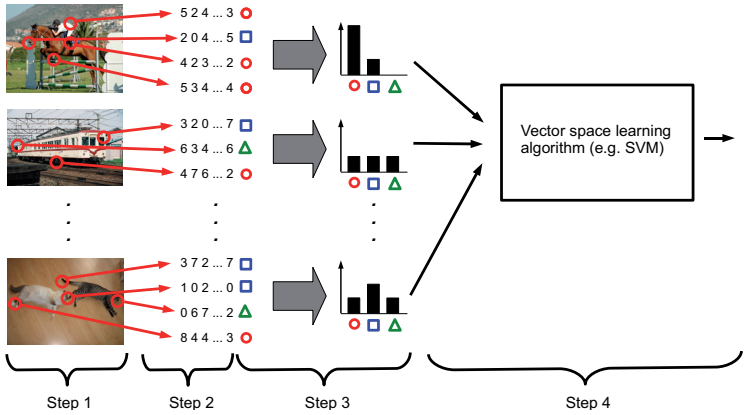


Figure 3.3. Steps in the supervised BoV pipeline. Adapted from Publication IX.

BoV principle

In their basic form, BoV features are based on the idea of representing images as orderless collections of their parts. This is analogous with the successful bag-of-words paradigm in the area of text retrieval [1]. In this analogy, different types of image parts correspond to different words in text documents. Figure 3.3 illustrates a prototypical BoV-based category detection system that was introduced in [34]. In such a system the processing of images can be divided into the following four steps:

Step 1 Selecting image locations of interest.

Step 2 Describing each location with a suitable visual descriptor.

Step 3 Characterising the distribution of the descriptor values within each image with a histogram.

Step 4 Using the histograms as feature vectors of the images in a supervised vector space learning algorithm, such as the support vector machine (SVM).

In this setup, the analogy to text retrieval is closely followed by quantising the visual descriptors to a finite set of descriptor types in visual vocabulary, i.e. the histogram codebook. The codebook is often selected by applying a clustering algorithm. Currently, the codebook transform is the most common way of synthesising an image-level feature out of local descriptors.

Codebook transform is not the only alternative, however. The descriptor distributions can be compared without quantisation to common vocabulary. This approach is implemented in [215], where k-means clustering

is first used within each image to obtain signatures for the images. The signatures are then compared via the earth mover's distance (EMD) [153] SVM kernel.

Methods employing Gaussian mixture models (GMM) [46, 97] are closely related to the codebook transform methods. There the GMM parameters (means and covariances of the components) are estimated jointly for all the images—for example with the EM-algorithm [38]—and the descriptor distribution in each image is then approximated with a weighted sum of the Gaussian components. The weights can then be used for comparing images, for example via the special SIFT-Bag kernel [217] that approximates a bound of the KL-divergence [93] between two distributions that are represented as GMMs with the same components. A more advanced approach employing Fisher kernels is taken in [140]. The Fisher kernel includes up to second order statistics of the model fit into individual statistics, in contrast with just the zeroth order statistics of histogram bin counting.

In the GMM approach, the image-collection-wide model estimation step is related to the selection of the codebook by clustering, only that in this case the codebook contains the covariances of the Gaussian components in addition to their means. Likewise, the image-wise weight estimation step is closely related to the assignment of descriptors to histogram bins. In this case, the assignment is no longer hard: one individual descriptor can affect the weights of several Gaussian components. This is qualitatively rather similar to the kernel codebook method of [186] where soft assignment is rather heuristically combined with the traditional codebook transform via kernels placed in the descriptor space. However, in the kernel codebook method the kernels smoothen the descriptor space equally around each codebook vector, whereas the different covariances of the Gaussian components are taken into account in the two-stage GMM method. It has been empirically found that soft histogram assignment significantly boosts performance of BoV systems based on codebook transform. Consequently, the technique is included in many such systems nowadays. A related idea is the use of Hamming embedding in conjunction of hard histograms [77] so that as an end result, the fine structure within the histogram cells gets taken into account when matching images.

BoV and point-wise matching

The BoV paradigm is related to another application area of local image descriptors: point-wise matching. In point-wise matching, one tries to identify corresponding locations in different images or image areas. Point-wise matches can be used, for example, for stereo correspondence problems [12], camera parameter estimation [155], object tracking in videos [181] or 3D reconstruction [137]. The point-wise matching methodology employs similar interest point detection and description steps as the BoV processing pipeline (steps 1 and 2).

Historically, many detector and descriptor algorithms have first been devised primarily point-wise matching in mind, and then used as such also for BoV processing. The assumption that descriptors and detectors performing well for point-wise matching would produce good performance also as part of the BoV pipeline is at least reasonable, if not optimal. However, point-wise matching sets stricter requirements for the detector–descriptor combination. For point-wise matching to work, the interest point detector has to be able to detect exactly the corresponding location in each of the images that are to be compared. This can be rather challenging if there is a severe geometric or photometric transformation between the images. In addition, detectors are limited to detect such salient image details that differ from their surroundings in appearance—for instance corners. For example, a point in the middle of a large non-textured or repetitively textured surface cannot be detected as an interest point, as it would be impossible to locate the corresponding point in the middle of the same surface on a transformed image. For BoV, the requirement of point-wise repeatability does not exist. It is enough that the statistical properties of the set of chosen image locations are similar to those of another set from a similar image.

Implementations of local image descriptor methods

A landmark paper in the area of local image descriptors is the paper by Lowe [108]. The paper considers techniques of point-wise matching. In the paper, the image locations of interest are detected as scale-space maxima of the difference-of-Gaussians (DoG) operator. This detector finds blob-like structures in images. The paper proposes to use SIFT (scale-invariant feature transform) descriptors for describing the regions around interest points. SIFT descriptors are histograms of local image intensity gradient directions. The interest point's neighbourhood is partitioned

with a 4×4 mask. For each part, a gradient direction histogram is collected using eight quantisation levels and soft assignment. The histograms are concatenated for the final 128-dimensional descriptor. Figure 3.4 illustrates the principle of the SIFT descriptor.

A number of other interest point detectors and descriptors have also been proposed. Examples of other interest point detector types include Harris-Laplace [119], Hessian-Laplace [121], Kadir-Brady [83], MSER [114] and SURF [13] detectors. The GLOH [120], HOG [35], LESH [154] and SURF [13] descriptors are examples of other proposed descriptors.

In [118] and [215], the performance of various detector–descriptor combinations is systematically evaluated in some category detection tasks. The studies of performance of various detectors and descriptors in pointwise matching (e.g [120, 206]) also bear some relevance to category detection, although the results are not directly applicable. As a conclusion of the studies, one can state that although some detectors and descriptors consistently show improved performance, the performance gain over the original SIFT methodology is not drastic. For example, variants of SIFT features still form the core of the BoV features of the top systems in the TRECVID 2009 evaluation.

Notable improvements in descriptor performance have been obtained, however, by extending the descriptor to take use of also colour information instead of the intensity-only original SIFT. In [185], several slightly different versions of such ColorSIFT descriptors are proposed. Another improvement that has proven useful in BoV context is to use image-independent sampling of image locations instead of interest point operators. The sampling can be performed either randomly [82, 111, 134] or using a regular grid [26, 49, 207]. The usefulness of the dense sampling approach is confirmed by many experiments (e.g. [134]), including those performed in the TRECVID 2009 HLFE setting that are reported in Section 3.2.4. In practice, good results have been obtained by performing the sampling in a single scale only, although multi-scale sampling has been reported to improve performance somewhat [86].

Selection of the visual vocabulary

An essential question in codebook-transform-based BoV systems is the selection of the visual vocabulary. The size of the codebook has to be determined and the codebook vectors selected. A traditional solution for the lat-

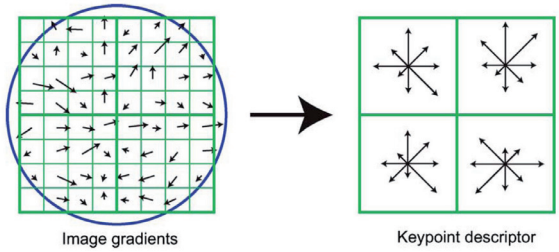


Figure 3.4. The principle of the SIFT descriptor. Image gradient magnitudes around an interest point are weighted with a circular Gaussian window. The interest point neighbourhood is partitioned into sub-regions. Each sub-region is described with a histogram of gradient orientations. The illustration shows a 2×2 partitioning, whereas the actual SIFT descriptor employs a 4×4 partitioning grid. The figure is from [108].

ter problem has been to employ an unsupervised clustering algorithm on the local descriptors. One may use, for instance, the k-means [156] or the Linde-Buzo-Gray (LBG) [105] clustering algorithms. The clustering approach has the fundamental shortcoming that optimising a clustering criterion, for example mean square quantisation error (MSE), does not guarantee optimal category detection performance. This can be seen in the experiment taken from Publication VII, where different clustering algorithms are compared (Figure 3.5). In that experiment, the self-organising map (SOM) clustering algorithm produces clusterings with significantly smaller MSE than the trivial alternative of choosing the codebook vectors randomly among the descriptor values. Still, SOM-based category detection performance is clearly inferior to such random codebooks.

It has been considered if one could use the knowledge of image categorisation task at hand, i.e. the training examples of categories, to make the codebook as distinctive as possible. Given the large number of local descriptors in a typical image collection, it is infeasible to use actual category detection performance as an optimisation cost function when selecting the codebook. One has to resort to shortcuts. A straightforward supervised codebook generation method is first to form a separate codebook for each category to be detected via unsupervised clustering and then to merge the codebooks [46, 215]. Alternatively, one can use a combination of category-specific and background codebooks when detecting a category (e.g. [140]). By doing this, one hopes to find codewords that are specific to particular visual categories. This was experimented with in Publica-

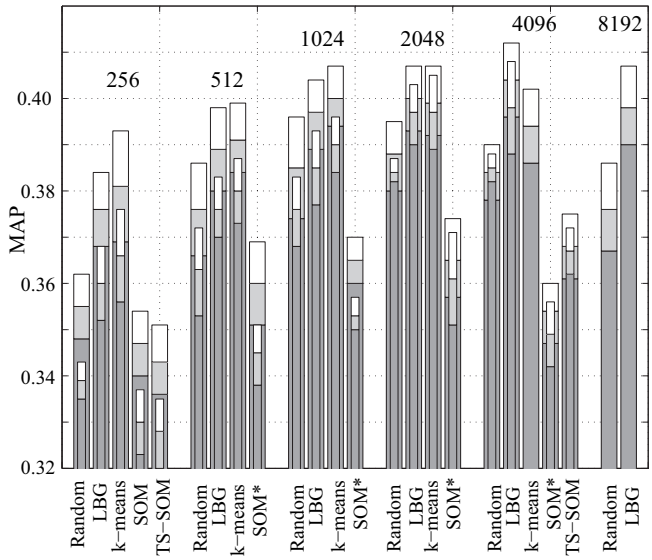


Figure 3.5. Category detection performances resulting from various unsupervised codebook selection algorithms in the experimental setting of Publication VII. The bars are grouped by the size of the codebooks. See the publication for more details.

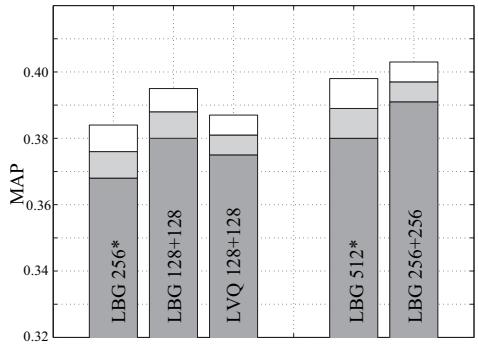


Figure 3.6. MAP performances resulting from supervised codebook selection The performance of unsupervisedly selected LBG codebooks of the same sizes are included for comparison and are marked with an asterisk (*). The image is from Publication VII.

tion VII. Even a simple implementation of this approach can result in some gain over unsupervised codebooks, as demonstrated by the results in Figure 3.6.

Other tractable alternatives include methods that consider the correlation of individual descriptor types and the category labels. A useful ap-

proach has been to first use clustering methods to generate tentative codebooks and then filter out codewords that have too mixed a category label distribution, measured in terms of entropy. Another alternative, followed in Publication VII, is to use the category label information for guiding the codebook search via the learning vector quantisation (LVQ) algorithm [89] that considers the correlations of individual codewords with category labels. This approach did not prove to work very well, although in [17] promising results have been reported for a similar approach. Of course, these methods build on the assumption that individual descriptor types are enough to distinguish between the categories. However, in reality the categories may be recognisable rather by the co-occurrence of different descriptor values.

Histogram granularity

In the histogram representation, the continuous distance between two visual descriptors is reduced to a single binary decision: whether the descriptors are deemed similar (i.e. fall into the same histogram bin) or not. Selecting the number of bins used in the histograms—i.e. the histogram size—directly determines how coarsely the visual descriptors are quantised and subsequently compared. In this selection, there is a trade-off involved. A small number of bins leads to visually rather different descriptors being regarded as similar. On the other hand, too numerous bins result in visually rather similar descriptors ending up in different histogram bins and regarded as dissimilar. The latter problem is not caused by the histogram representation itself, but the desire to use the histograms as structureless feature vectors in conventional machine learning algorithms.

The histogram granularity issues were experimentally studied in Publication VII. The results of those experiments indeed demonstrate that given a particular category detection task—a subset of PASCAL VOC Challenge 2007 images—the performance can vary markedly depending on the selected histogram size. The optimal size was about 4 000 in those experiments. In [209] it is demonstrated that the optimal size can vary between category detection problems by several orders of magnitude. The image corpora used there are the 1578 image PASCAL VOC Challenge 2005 data set and the TRECVID 2005 corpus with approximately 30 000 keyframes. There the optimal histogram sizes were approximately 5 000 and 300 000 (larger histograms were not tried), respectively. The em-

ployed learning algorithms can have a large effect on the optimal size. For example, with radial basis function (RBF) SVM instead of linear SVM the TRECVID performance peaked with just 20 000 keywords.

The various factors affecting the optimal histogram size are not well understood. In addition to the number of images, one would expect that the nature of the images together with the number and nature of the categories would be of importance. There seems to be no widely accepted principled methodology for selecting the size of histograms. Often one performs experiments with a few histogram sizes, possibly using cross-validation. The use of a cluster number selection metrics based on the minimum description length (MDL) principle has been proposed [86]. However, these metrics balance the complexity with the ability of the codebook to reproduce the source data, not with the resulting category detection performance. There seems to be no convincing experimental evidence that these size-selection metrics would lead to good category detection performance. For practical systems, one often resorts to selecting a fixed codebook size that has worked decently in some earlier category detection tasks. For example, many TRECVID 2009 participants have just used a codebook size of order of one thousand (e.g. [41, 138, 200]) with no obvious reason for just those choices.

Selecting the optimal size of histogram codebooks can be re-worded as selecting the optimal granularity or scale in the descriptor space. In Publication VIII, the scale selection problem is turned into scale-fusion problem following similar lines of thought as in [66]. Different methods for multi-granularity fusion are experimentally compared in the PASCAL VOC Challenge 2007 data set. In the following, the fusion is discussed in the context of BoV features. For the technical description of the fusion algorithms in general, the reader is referred to Section 3.4 of this thesis.

In the first of the methods considered in that publication, histograms of different granularities are concatenated with weights, corresponding to a multi-granularity kernel function in the SVM. This is an example of a distance combination kernel approach discussed in Section 3.4.1. It is closely related to the pyramid matching kernel method of [66].

Publication VIII also investigates two ways of modifying the histograms so that the descriptor-space similarity of the histogram bins and descriptors of the interest points are better taken into account. These approaches are called the post smoothing and soft histogram techniques. The latter of the methods is essentially the same as the one proposed in [188] if

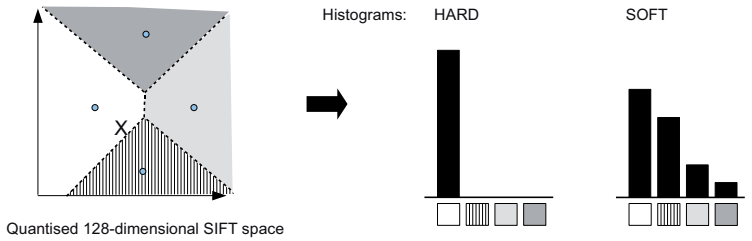


Figure 3.7. The principle of soft histograms contrasted with traditional hard histograms. In the feature space, an interest point descriptor (X) falls into one particular (white) hard histogram bin. When forming the soft histogram, however, the bin counters of the neighbouring bins are also incremented in a lesser degree.

some particular design choices are made. Figure 3.7 illustrates the principle of soft assignment. In Publication VIII, the use of the soft histogram technique is inspired by multi-granularity concerns, whereas in [188] the technique is motivated from the codeword ambiguity point of view. As an alternative to these domain-specific mechanisms for combining information from different levels of histogram granularity, also the general purpose post-classifier fusion by Bayesian binary regression (BBR) [60] (see Section 3.4.1) is evaluated in the publication.

Figure 3.8 summarises the results of the scale-fusion experiments of Publication VIII for codebooks of different sizes. In the figure, the bin counts on the abscissa directly indicate the number of bins in the baseline single-granularity hard histograms, as well as in the soft histograms. Histograms of all sizes from 128 up to the indicated number are used for BBR and multi-granularity kernel fusion. One can observe that by using the best one of the methods, a significant improvement of MAP from 0.404 to 0.451 was obtained in comparison with the best-performing histogram of a single granularity. Of the techniques, the soft assignment of descriptors to histogram bins resulted in clearly the best performance. Histogram smoothing as post-processing improved the performance only slightly over the baseline single-granularity histograms. The multi-granularity kernel technique was better than the baseline of single-granularity histograms with maximum MAP 0.425, but clearly inferior to soft histograms. Combining soft histograms with the multi-granularity kernel technique did not result in further performance gain, supporting the conclusion that both techniques leverage on the same information and are thus redundant. The soft histogram technique adds some computational cost in

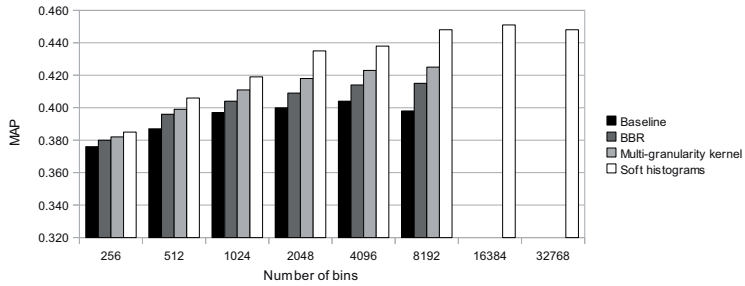


Figure 3.8. Category detection performances resulting from the use of various codebook granularity fusion algorithms. The baseline method uses single-granularity histograms with no fusion.

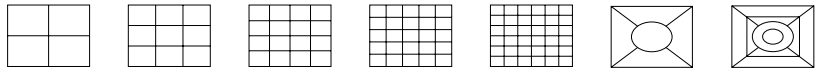


Figure 3.9. Tiling patterns for partitioning the image area. From Publication IX.

comparison with individual hard histograms as it becomes beneficial to use larger histograms, and the generated histograms are also less sparse.

Spatial extensions to BoV

In its basic form, the BoV approach does not take into account the interest points' spatial distribution within images. However, many image categories are such that the spatial structure could be useful in their detection. A common extension to BoV has been to apply the "divide-and-disorder" principle by geometrically partitioning all the images with the same tiling pattern. Each part is described with a separate histogram and the image dissimilarity is formulated as the sum of the dissimilarities of corresponding tiles. Figure 3.9 shows some such tiling masks. There exist also other approaches for including spatial information in BoV. For example, [180] and [81] apply language modelling techniques—n-gram modelling in particular—to capture the spatial proximity of local features. The shape mask technique [112] integrates spatial cues in weaker, more statistical manner in object detection. There the relation of each local feature in training material to annotated object contours is recorded. Given a test image, each local feature of that image hypothesises a set of possible object contours. By averaging the hypotheses, one arrives in a probable object outline map, which can be used for weighting different local features accordingly.

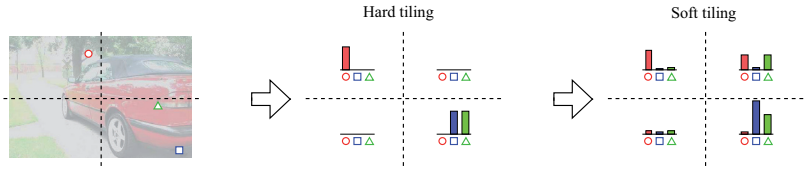


Figure 3.10. The principle of spatially soft tile assignment compared with conventional hard assignment. In hard tile assignment, each of the three indicated interest points falls to exactly one image quadrant and increments only the bin counters of the histogram associated with that quadrant. In contrast, when using soft tiling, each interest point increments the counters of all the four histograms. However, the magnitudes of the increments depend on the points' closeness to the corresponding image quadrants.

If spatial information is incorporated to BoV via geometrically subdividing the image, one is faced with a scale selection problem: how fine-grained spatial tiling patterns should be used? Here one can follow similar lines as in the case of histogram sizes, fusing together information from several spatial scales. An influential approach in this has been [99] that successfully applied the pyramid matching kernel of [66] into the spatial domain. In Publication IX, an alternative method is proposed that can be seen analogous to the soft histogram assignment. In the soft tiling technique, one assigns the interest points not only to one spatial image tile as in the traditional hard tiling but to several ones with varying degrees. Figure 3.10 illustrates this principle. The soft tiling can be presented with spatially varying tile membership masks. Figure 3.11 shows some such membership masks.

The experiments of Publication IX compare the combinations of spatial fusion techniques—kernel-level fusion (i.e. spatial pyramids), soft histograms and BBR post-classifier fusion—in the VOC Challenge 2007 setting. See Section 3.4 for more discussion on techniques for fusion. Besides the fusion algorithm, the experiments address the questions of the

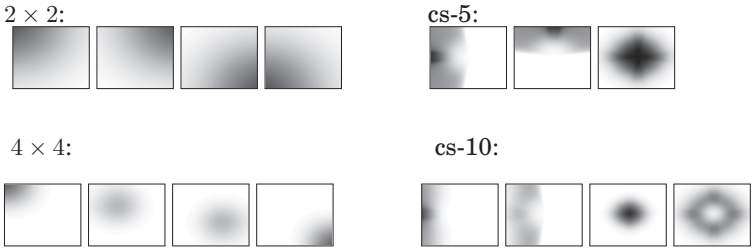


Figure 3.11. Membership masks of some spatially smooth tilings. Adapted from Publication IX.

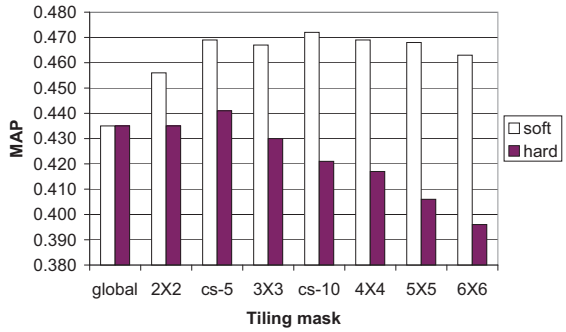


Figure 3.12. Comparison of hard and soft tiling techniques in the VOC 2007 object detection setting. From Publication IX.

selection of the set of tiling masks to fuse, and the selection of histogram granularity. The experiments show (Figure 3.12) that spatially soft tiling is a useful technique even in the case of a single tiling mask. The explicit kernel-level or BBR mechanisms can be combined with soft tiling and the performance slightly improved. This can be contrasted with the use of spatially hard tilings where fusion of several resolutions is an essential prerequisite for good performance.

Publication X compares two methods of combining the dissimilarities of the multiple histograms into a single image-level dissimilarity measure in the sub-image partitioning extension of BoV. In the traditional rigid matching method, the dissimilarity is obtained as a sum of dissimilarities of histograms of each sub-image. Each sub-image is compared with the sub-image of the other image that lies in exactly the same geometric location. The other considered alternative builds on the integrated region matching (IRM) scheme [203]. In this case the IRM matching corresponds to permuting the sub-images of one of the images before the dissimilarity calculation so that the dissimilarity is minimised. One can hope this to lead to better detection of categories where the exact locations of image contents are not relevant, images still belonging to the same category after slight transformations (Figure 3.13). Experiments show that on average, rigid matching works slightly better than IRM matching (Figure 3.14). However, for some categories IRM matching is significantly more accurate an alternative (Figure 3.15). The best results are obtained by combining the two schemes. Another result of that publication is that the soft tiling technique is very well compatible with the IRM matching,

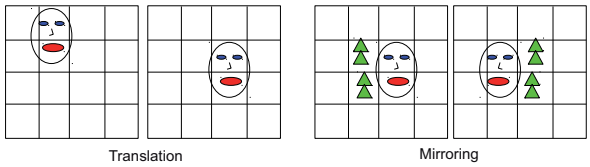
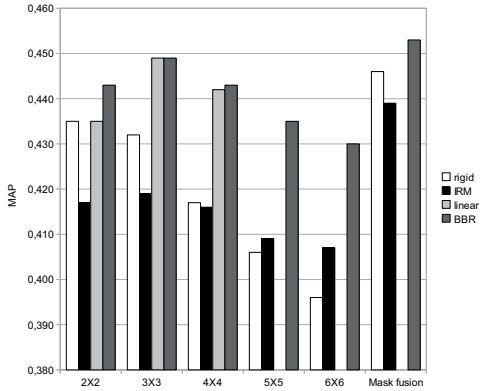
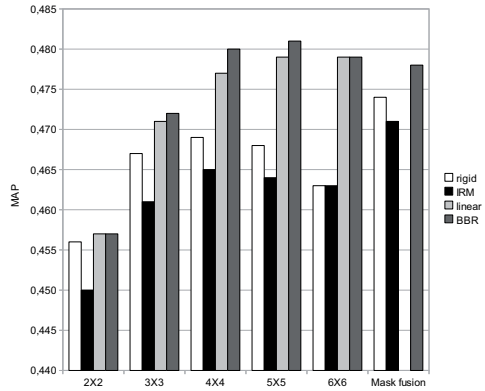


Figure 3.13. Examples of image transformations against which invariance might be useful in a category detection system.



(a)



(b)

Figure 3.14. Comparing the performance of rigid and IRM matching (white and black bars) to the fusion of the matching techniques with two different fusion algorithms (grey bars) in the case of either hard (a) or soft (b) spatial tiling. See Publication X for details.

providing substantially better performance than hard tiling. In the experimental results a situation is encountered where the best performance is provided by a single soft 5×5 tiling mask, outperforming the explicit fusion of several tiling masks.

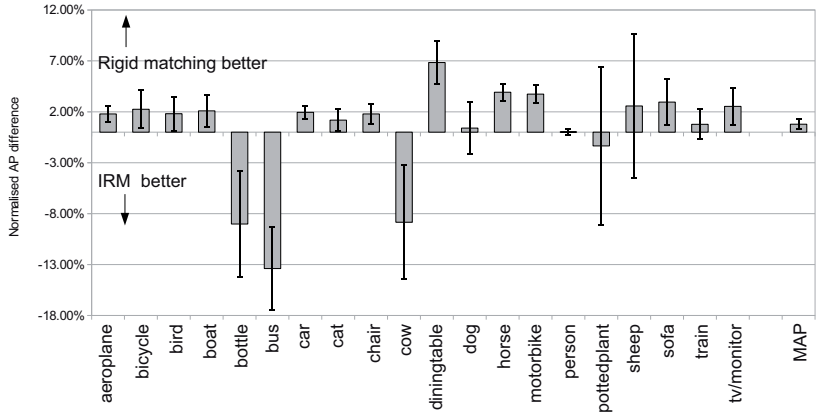


Figure 3.15. Category-wise comparison of rigid and IRM region matching when soft spatial tile assignment is employed. From Publication X.

It is somewhat surprising that the rigid region matching scheme works as well as it does. Certainly some of the image categories of the current experiments are invariant to some geometric image transforms. Apparently the complete geometric invariance of IRM is excessive. One would probably benefit from a region matching scheme that allows for some more transformations than the rigid matching, but would still not completely abandon rigidness as IRM matching does.

Metrics for comparing SIFT descriptors

The experiments in Publication XI address the issue of the metric that is used for comparing SIFT descriptors. Originally, in [108] Euclidean distance was proposed and it has been used ever since. However, since SIFT descriptors are essentially histograms, it is a natural idea to use the χ^2 distance

$$d_{\chi^2}(\mathbf{x}_i, \mathbf{x}_j)^2 = \sum_{l=1}^{k} \frac{(x_{il} - x_{jl})^2}{x_{il} + x_{jl}} \tag{3.1}$$

instead, as this distance measure has proven to work better for comparing histograms, in particular in BoV systems [215]. The idea is not specific to BoV use, it could be useful in point-wise matching applications as well. However, in Publication XI the idea was tested in the BoV context with the VOC Challenge 2007 images by replacing all the Euclidean comparisons between SIFT descriptors in our BoV system with comparisons based on the χ^2 distance. Figure 3.16 shows that the performance difference be-

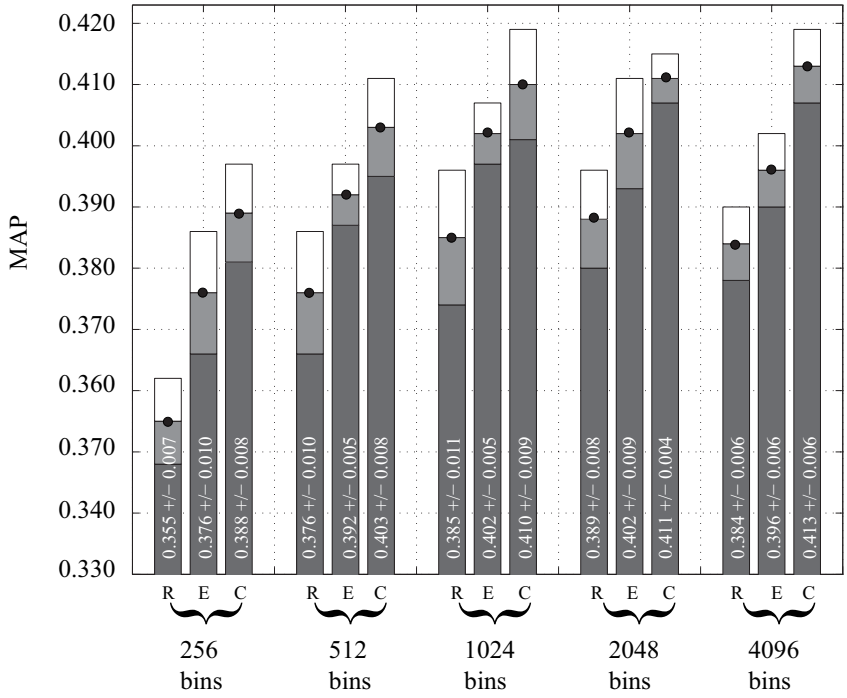


Figure 3.16. Mean average precision (MAP) in image category detection and its 95% confidence interval. Bars corresponding to different clustering algorithms are indicated by letters R (random codebook selection), E (Euclidean k-means) and C (χ^2 k-means). From Publication XI.

tween systems using Euclidean and χ^2 versions of k-means clustering and codeword assignment was almost as large as the difference between random codebooks and Euclidean k-means. Still, in absolute terms the difference is not very large. This actually tells that the conventionally used clustering techniques—such as k-means—produce codebooks that are not that much better than random selection.

3.2.3 Video features

Often a single well-chosen keyframe can compactly express the most central visual characteristics of a video shot. In practice, still-image features of keyframes form the backbone of the feature representations of current state-of-the art video retrieval systems [131].

Naturally, the single keyframe approach is not able to represent all the information that could be relevant to video category detection. A straight-

forward extension is multi-keyframe analysis, which still builds on the still-image features of the keyframes. This approach has proven to be useful in TRECVID evaluations of recent years (e.g. [172, 104]). One approach to multi-keyframe analysis is to perform the category detection separately on basis of each keyframe, and then combine the results e.g. by averaging or taking maximum of the results. Another approach, employed in the PicSOM system is to concatenate the feature vectors of the keyframes. Alternatively, a single feature vector can be aggregated over all the keyframes, or even all the frames of the video shot as is done in the SIFT-bag kernel technique of [217].

An additional source of features is the analysis of motion in a video shot. One can obtain motion-sensitive features by extending still-image features to take into account motion information. For example, the MoSIFT features of [30] extend the BoV methodology by including motion information in both the interest point selection operator and in the elementary image property that is collected into histograms over local image neighbourhoods in a SIFT-like fashion.

One may also aggregate more global characteristics of motion. An example of a global motion feature is the KLT Histogram feature that has been used with the PicSOM system. The feature is based on local feature points tracked using the Kanade-Lucas-Tomasi tracker [181]. For every frame, each feature point is classified either as missed, static, or moving. The moving feature points are then mapped into eight principal directions similarly as in [42]. Two relative directions are additionally used, one toward the centre of the image and one away from it, for detection of zoom-in and zoom-out. This results in a twelve-bin motion histogram, which can be used both as a statistical motion feature and a basis for detectors of different types of motion, such as camera pan, camera zoom, and object movement.

Videos are not limited to the visual modality, but often contain also audio, yet another source of features. Sound tracks of video shots might contain for example human speech, music or different environment sounds. The characteristics of the sound track can be described either globally or the track can be segmented into separately described parts. A popular approach for sound track characterisation is to calculate the mel-frequency cepstral coefficients feature (MFCC), which is given by the discrete cosine transform (DCT) applied to the logarithm of the mel-scaled filter bank energies [36]. This feature is quite generic and commonly used in speech

recognition, musical genre classification, and classification of the type of sound, e.g. whether it consists of speech, music or background noise.

Besides coarse general level description of audio, speech can often be automatically recognised, resulting in textual representation that can be processed with the very mature techniques of text retrieval, such as stemming and part-of-speech tagging, and existing tools for text-based indexing, such as the Lemur toolkit [135] or Apache Lucene [68]. Text-based information retrieval is not addressed in this thesis. The textual search module could easily be implemented as a separate component of the video analysis system whose output is fused afterwards with the other modalities.

3.2.4 Case example: the high-level feature extraction task of TRECVID 2009

This section gives a practical example of feature techniques employed in the systems used for tackling the high-level feature extraction task of the TRECVID 2009 video retrieval benchmark. First, the set of features employed in the PicSOM group's entry to the competition is described. In the benchmark, the PicSOM group showed the sixth best concept detection performance out of 40 groups. With slight system improvements afterwards, the performance is practically on-par with all but the top two participants. Thus the details of the PicSOM feature representation bear some practical significance. An empirical comparison of the individual PicSOM features in the task is therefore shown in the following. After that, the feature extraction approaches employed by the other top participants of the task are briefly reviewed.

The PicSOM feature set

The feature set employed in the PicSOM multimedia retrieval system for the TRECVID 2009 high-level extraction task consists of a diverse selection of image features extracted from the keyframes of the video shots, and their multi-keyframe extensions. Admittedly, the details of the feature set are not carefully optimised in terms of performance but dictated by the availability of software at hand and other convenience factors during time of the TRECVID 2009 participation.

The feature set includes eight different bag-of-visual-words (BoV) features. The eight features result from combining a number of indepen-

dent design choices. Firstly, either the SIFT [108] or the opponent colour space version of the Color SIFT descriptor [185] is used. Secondly, either the Harris-Laplace interest point detector [119] or dense sampling is employed for selection of image locations of interest. Third, some features use the soft-histogram refinement of the BoV codebooks [185]. Finally, some of the features use the spatial pyramid extension of the BoV model [99] that combines histograms of 1×1, 2×2 and 4×4 image partitionings. The codebooks for the histograms have been determined with k-means and self-organising map (SOM) clustering algorithms with either 1 000 (ColorSIFT) or 2 000 (SIFT) bins.

The non-BoV features divide into two groups: features closely resembling image descriptors specified in MPEG-7 standard [75], and custom features developed in the PicSOM group. Table 3.1 summarises these features. The MPEG-7 descriptors have been widely used in image retrieval research (e.g. [95, 142, 216]). This probably is not because the descriptor set would be a particularly effective feature representation, but because the feature set is a well-defined standard readily available to be used.

The MPEG-7 like features in PicSOM omit some details of the standard that are seemingly irrelevant to the image category detection system, such as the scaling of the features and issues related to transferring the features in real-time multimedia streams. Of the MPEG-7 descriptor types, the Color Layout descriptor describes the layout of colours in an image by a set of discrete cosine transform (DCT) coefficients of average colour in 20×20 rectangular grid. The Color Structure descriptor resembles a histogram in quantised HMMD colour space (see [75]), but encodes also information on the spatial distribution of colours. This is achieved by extending the histogram formation to take into account the colours present in rectangular neighbourhoods, instead of just individual pixels. The Dominant Color descriptor encodes two most prominent colours in an image's global colour distribution in the CIE L*a*b* colour space. The Edge Histogram descriptor concatenates the edge direction histograms of sub-images in a rectangular 4×4 grid. Also the counts of non-directional edges are included in the histograms. The edge directions are detected with Sobel operators. The Scalable Colour descriptor is the Haar transform of image-wide colour histogram in the HSV colour space.

For the remaining non-standard non-BoV features, the five-part centre-surround partitioning mask (Figure 3.2) is employed: the feature is first calculated for each of the sub-regions separately, after which the five fea-

Table 3.1. Non-BoV features employed in the PicSOM system for TRECVID 2009 HLFE task. The column "Mask" refers to the spatial partitioning mask of image area.Here cs-5 denotes the 5-part centre-surround mask. "Dim." denotes the dimensionality of feature vectors.

Feature		Mask	Multi-frame	Dim.
MPEG-7	Color Layout	global	no/yes	12/60
	Color Structure	global	no	256
	Dominant Color	global	no	6
	Edge Histogram	4×4	no/yes	80/400
	Scalable Color	global	no	256
non-standard	Average Colour	cs-5	no	15
	Colour Moments	cs-5	no	45
	Texture Neighbourhood	cs-5	no	40
	Edge Histogram	cs-5	no/yes	20/100
	Edge Co-occurrence	cs-5	no	80
	Edge Fourier	cs-5	no	128

ture vectors are concatenated. Of the features, the Average Colour feature is a three-element vector that contains the average RGB values of all the pixels within a zone. The Color Moments feature vector consists of the first three central moments (mean, variance and skewness) of the HSV colour distribution, separately for each channel. For the Texture Neighbourhood feature, relative values of the Y (luminance) component of the YIQ colour representation in all 8-neighbourhoods within the zone are characterised. The probabilities for neighbouring pixels being more luminous than the central pixel are estimated for all the eight surrounding pixel positions and collected as a feature vector. Edge Histogram is the histogram of four Sobel edge directions. The feature differs in details from the similarly named MPEG-7 descriptor. Edge Co-occurrence gives the co-occurrence matrix of four Sobel edge directions. The Edge fourier feature vector consists of the magnitude of the 16×16 discrete Fourier tranform (DFT) of the Sobel edge image.

In addition to single-keyframe features, some of the features have been extended to employ multi-keyframe analysis by concatenating the feature vectors of five temporally evenly spaced keyframes.

Table 3.2. Concept detection accuracy (MIAP) based on various BoV image features. From Publication VI.

Feature	sampling	histograms	spatial partitioning	MIAP
Color SIFT	dense	soft histograms	spatial pyramid	0.1166
Color SIFT	dense	soft histograms	global	0.1031
Color SIFT	interest points	soft histograms	spatial pyramid	0.1014
Color SIFT	interest points	soft histograms	global	0.0961
Color SIFT	dense	hard histograms	global	0.0988
SIFT	interest points	hard histograms	global	0.0832

Performance of PicSOM features

The performance of individual PicSOM features in the TRECVID task were measured as an intermediate result, even though the features are not intended to be used separately but as components of the final fusion-based category detection system. Tables 3.2 and 3.3 report the performances of a selection of best individual features averaged over all the 20 concepts that were detected in the TRECVID 2009 HLFE task. The performances are reported using the mean inferred average (MIAP) measure, the measure employed in the TRECVID 2009 evaluation (see Section 2.3.1). Figure 3.17 visualises the results. One sees that the best individual feature performances are obtained employing histograms of local image features collected according to the BoV paradigm, i.e. variants of SIFT and Color SIFT features.

Table 3.2 compares different BoV feature variants in terms of MIAP. As expected, dense sampling is a more effective approach than interest point detection. The soft histogram technique and spatial pyramids improve the performance of the BoV features as well. These results hold on average, but concept-wise differences are large. Table 3.3 lists the most accurate non-BoV features. From this table the observation can be made that the multi-keyframe ("video") features always perform better than their single-keyframe ("image") counterparts. The tabulated results are illustrated in Figure 3.17.

Successful feature representations in TRECVID 2009 HLFE task

Figure 3.18 shows the performance the best groups obtained in the high-level feature extraction task of the TRECVID 2009 evaluation. In the figure, the bar labelled "PICSOM2009" corresponds to the version of the system whose results were submitted to the official evaluation. The leftmost

Table 3.3. Selection of feature-wise concept detection accuracies (MIAP). From Publication VI.

Feature	type	MIAP
Edge Histogram	video	0.0625
Color Moments	image	0.0438
MPEG-7 Edge Histogram	image	0.0417
Edge Histogram	image	0.0403
Color Layout	video	0.0340
Color Layout	image	0.0309
Scalable Color	image	0.0330
Edge Fourier	image	0.0290
MPEG-7 Color Structure	image	0.0263

bar represents the performance the PicSOM system was able to achieve with slight adjustments made after the evaluation. Table 3.4 provides a key to the acronyms of the systems and Table 3.5 briefly summarises the feature presentations of the ten best-performing groups. From the table it can be seen that all the best performing groups use variants of BoV features. Another observation that can be made is that most of the groups employ a wide variety of features. In this light, the good performance of the fourth-ranked TIT group underlines the effectiveness of the SIFT-Bag video kernel BoV variant they use, as their performance results almost entirely from this feature alone, their audio features providing only a small help.

3.3 Supervised detection

The supervised detection stage of the discussed category detection architecture tries to re-produce the mapping between the feature representations and the categories to be detected. For each category, the detection is performed separately on basis of each feature. Naturally, several different detection algorithms or algorithm variants can be used for producing several alternative feature-category mappings, even for a single feature.

The input to the supervised detection stage consists of the feature vectors resulting from the feature extraction stage and category labels of the training part of the data objects. For concreteness, the example case of data objects being images is considered in this section from here on. However, they might be anything else as well, e.g. videos. The output of the

Table 3.4. The top groups in the TRECVID 2009 HLFE task. Each system is described in detail in the respective notebook paper in the TRECVID 2009 workshop proceedings [131].

Rank	System	MIAP	Affiliation
1	MediaMill	0.228	ISLA, University of Amsterdam INESC-ID, Lisboa CVSSP, University of Surrey
2	PKU-ICST	0.203	ICST, Peking University
3	FTRDBJ	0.170	France Telecom Orange Labs (Beijing) Beijing University of Posts and Telecommunications
4	TIT	0.168	Tokyo Institute of Technology Georgia Institute of Technology
5	CityUHK	0.163	Video Retrieval Group, City Univ. of Hong Kong Digital video and Multimedia Lab, Columbia Univ.
6	PicSOM2009	0.152	AIRC, Helsinki University of Technology
7	VITALAS	0.147	MUG, Aristotle University of Thessaloniki ITI, Centre for Research and Technology Hellas CWI, Amsterdam Fraunhofer IAIS French Audiovisual Institute Institut für Rundfunktechnik Robotiker-Tecnalia EADS Val de Reuil INRIA Rocquencourt
8	EURECOM	0.138	Institute Eurecom, France
9	Tsinghua	0.134	Intelligent Multimedia Group, TNList, Tsinghua Univ.
10	LIFM	0.132	Laboratoire dInformatique Fondamentale de Marseille, Université de la Méditerranée, Université de Provence

Table 3.5. The feature extraction techniques employed by the top groups in the TRECVID 2009 HLFE task.

Rank	System	MIAP	Features
1	MediaMill	0.228	Extensive sets of BoV features employing: interest point detectors and dense sampling multiple descriptor types: 1) SIFT, 2) several ColorSIFT variants, 3) SURF spatial pyramids multi-keyframe analysis fast approximation techniques 20 audio features
2	PKU-ICST	0.203	six SIFT BoV variants set of global image feature incl. grid colour moments, LBP, EH layout audio features
3	FTRDBJ	0.170	several SIFT BoV variants global colour, texture and edge features
4	TIT	0.168	SIFT-bag GMM of whole shots audio features (MFCC)
5	CityUHK	0.163	four BoV features several global image features incl. grid colour moments and wavelet texture
6	PicSOM2009	0.152	several SIFT-based BoV features several global grid and multi-keyframe visual features (cf. Section 3.2.4)
7	VITALAS	0.147	ColorSIFT BoV of keyframes, HOG/Optical flow of space time interest points of whole video shots, several low-level global image features incl. colour layout, Weibull edge statistics and face detectors
8	EURECOM	0.138	SIFT BoV set of global image features incl. grid colour moments, grid wavelet texture, LBP object detectors: face, person, bicycle audio features (MFCC)
9	Tsinghua	0.134	several BoV features (colour, SIFT, SURF) set of global grid features: colour, Gabor texture, edge statistics
10	LIFM	0.132	very extensive set of features (over 100) including BoV features (SIFT,SURF) global grid image features (colour, texture, edges) mid-level perceptual and semantic features audio features, and many others

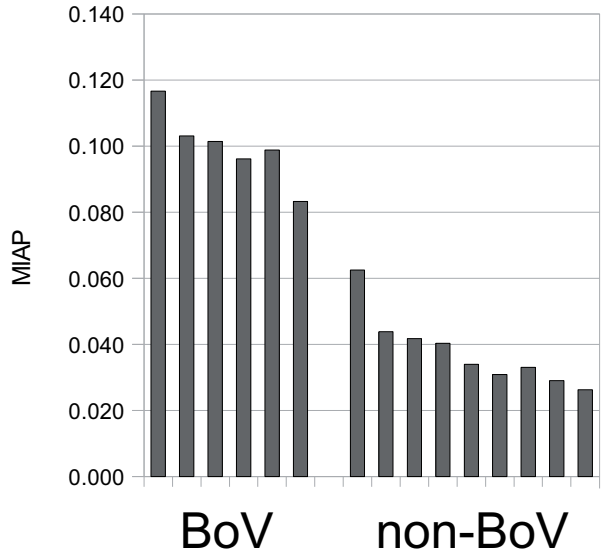


Figure 3.17. Comparison of the performance of individual BoV (Table 3.2) and non-BoV features (Table 3.3) in the TRECVID 2009 HLFE setting.

supervised detection stage is a detection score for each test set image, once again separately for each feature and category. The score values may directly be estimates of the probability of the category membership, or they can be values more inherent to the learning algorithm.

In addition to the test image detection scores, one may also need to estimate the scores for the training images if the subsequent stages of the category detection system employ supervised algorithms. In practice, the estimation can be carried out by using cross-validation techniques [39]. There the training set is partitioned into multiple folds. For each fold, the category detection scores are obtained from a feature–category mapping that is learnt using the other training images apart from that particular fold. Special measures for ensuring unbiased selection of the folds may need to be taken if the training set is ordered in a way that all training examples can not be considered independent [187]. This is the case with video shots, for example, where temporally adjacent shots often belong to the same program and have very similar contents.

Given the extracted features, the learning problem is no longer specific to any particular application domain—be it images, videos or text

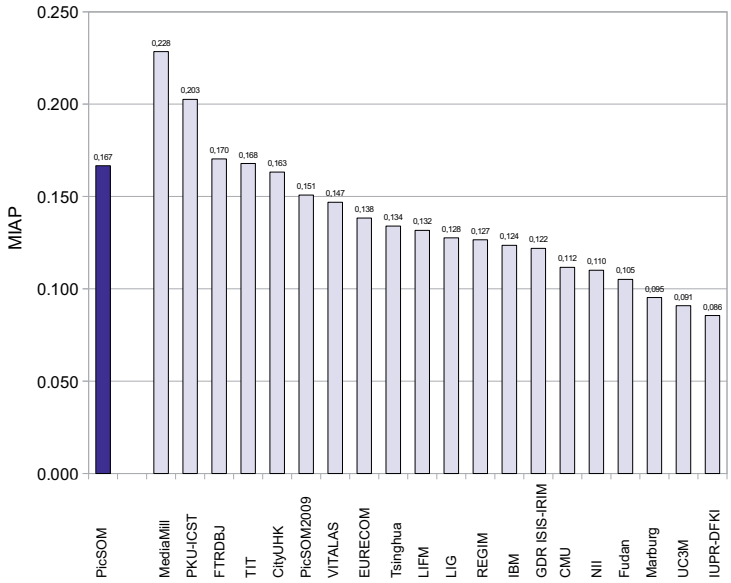


Figure 3.18. The performance of top groups in the TRECVID 2009 high-level feature extraction task. Adapted from Publication VI.

documents—but any supervised learning algorithm can be employed. This core problem in machine learning has been intensively studied over the decades. It is out of scope of this thesis to try to review it. Rather, a more empirical view is taken. In the following, two learning algorithms used in the PicSOM category detection framework are first discussed in Section 3.3.1, namely the tree-structured self-organising maps (TS-SOM) and support vector machines (SVM). After that, Section 3.3.2 lists learning approaches that have been successfully applied in other category detection systems.

3.3.1 Learning algorithms in the PicSOM system

The PicSOM category detection system implements two types of learning algorithms. The construction of the SOM-based classifiers begins with quantising the feature spaces using the TS-SOM [91] algorithm, a tree-structured variant of the SOM [89]. In the subsequent learning algorithm, the bottom levels of TS-SOMs define the quantisation and the upper levels act as an index structure for rapid search. Typically, TS-SOMs from two to

four stacked levels have been used, the bottom levels measuring from 16×16 to 256×256 map units, respectively. Figure 3.19 shows an example of a TS-SOM quantisation of a feature space based on the colour and texture distribution of image segments.

The TS-SOM preparation step needs to be performed only once for each feature type in an image collection. After that, generating a classifier for any binary partitioning of the training images is very fast. Any partitioning is characterised by the division of the training images into positive and negative examples. The classifier for the partitioning is created by subtracting the proportion of negative examples that fall into each bottom-level TS-SOM unit, i.e. quantisation bin, from the corresponding proportion of positive examples. This way a classification score is assigned to each quantisation bin. After this initial scoring, the scores are low-pass filtered on the two-dimensional TS-SOM grid surface, taking advantage of the topology-preserving characteristic of the SOM clustering and efficiently emphasising the differences between the feature space regions where positive and negative examples are well separated, or occur mixed with each other. Figure 3.20 illustrates the filtering step. Now the preparation step is complete and a detection score is associated with each quan-

Figure 3.19. A TS-SOM partitioning of the feature space defined by colour and texture distribution of image segments. From [194].

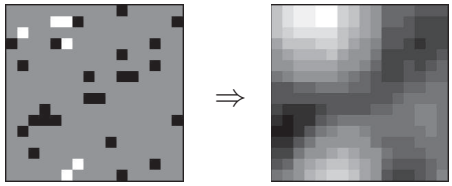


Figure 3.20. An example of low-pass filtering a SOM surface. On the left, positive and negative training images are shown with white and black marks, respectively. On the right, the filtered result is shown where relevance information is spread around the training examples. From [193].

tisation bin of the feature space. Assigning a feature-wise detection score to an independent test image is then simple: the extracted feature vector of the image is quantised using the same quantisation scheme and the image receives the detection score of the quantisation bin into which its feature vector is mapped.

Support vector machine (SVM) [33] variants are machine learning algorithms that provide very competitive accuracies in diverse generic classification and regression tasks. Consequently, using SVMs is very popular nowadays. The basic idea in the SVM learning algorithm is to formulate a cost function for optimisation that seeks to find a hyperplane—i.e. a linear decision surface—that separates the positive training examples from the negative ones with the widest possible margin. Figure 3.21 illustrates this idea in two dimensions. The maximisation of the margin can be shown to approximately correspond to structural risk minimisation [191, 190], a learning paradigm that guarantees the best possible generalising ability of the trained classifier. In practice the optimisation of the cost function leads to the parameters of the hyperplane being determined by a small subset of the training examples. These examples are called support vectors, hence the name of the method. SVM was first introduced in its linear form, i.e. the decision surface in the feature space was given by a linear function. In the 90s, the SVM was extended to non-linear classification via the kernel trick. This means replacing the linear inner products of feature vectors appearing in the optimisation cost function with a non-linear kernel function. This results in non-linear decision surfaces. Alternatively, the kernel trick may be interpreted to map the feature vectors to a transformed space, which can be of very high dimensionality. In the transformed space, the decision surfaces are linear.

The SVM implementation used in the PicSOM system is an adaptation of the C-SVC classifier of the LIBSVM software library [29]. The library

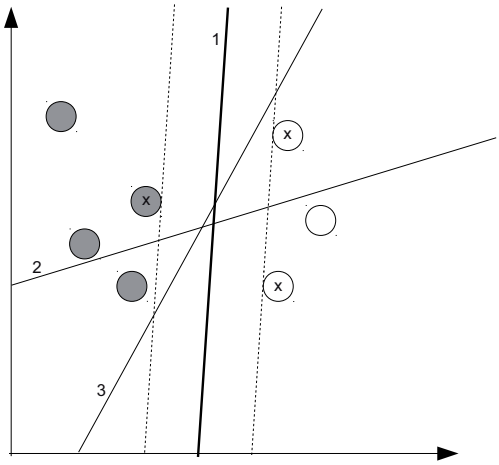


Figure 3.21. Three hyperplanes (1-3) in the 2-dimensional data space. Planes 1 and 3 both separate the training examples, symbolised with the grey and white balls, whereas plane 2 does not. Plane 1 separates the training examples with maximum margin (illustrated by the dashed lines). Support vectors are denoted with 'x'.

solves the C-SVC optimisation problem

$$\min_{\mathbf{w},b,\xi} J = \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} + C\sum_{i=1}^{N_{\text{train}}} \xi_i \tag{3.2}$$

subject to constraints

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \tag{3.3}$$

$$\xi_i \geq 0 \tag{3.4}$$

given the pairs $(\mathbf{x}_i, y_i)$ of feature vectors and class labels for the N_{train} examples in the training set. Here $\mathbf{x}_i \in \mathbb{R}^k$ and $y_i \in \{-1, +1\}$. In the equation, $\mathbf{w}$ and b are the parameters of the separating hyperplane. The function $\phi(\cdot)$ specifies the kernel function g via $g(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. The penalty coefficient C is a free parameter of the method.

The PicSOM system employs kernels that are exponential functions of a distance function $d(\mathbf{x}_i, \mathbf{x}_j)$ between feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$g(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma d(\mathbf{x}_i, \mathbf{x}_j)^2\right). \tag{3.5}$$

In the kernels, γ is a free parameter. For the distance function, the Euclidean distance is the basic alternative:

$$d_{\text{E}}(\mathbf{x}_i, \mathbf{x}_j)^2 = \|\mathbf{x} - \mathbf{x}'\|^2 = \sum_{l=1}^{k}(x_{il} - x_{jl})^2, \tag{3.6}$$

resulting in radial basis function (RBF) kernels. In the above, m denotes the dimensionality of feature vectors. In addition, the χ^2 distance is used for histogram-like features, the BoV features in particular:

$$d_{\chi^2}(\mathbf{x}_i, \mathbf{x}_j)^2 = \sum_{l=1}^{k} \frac{(x_{il} - x_{jl})^2}{x_{il} + x_{jl}}. \tag{3.7}$$

The motivation for this is the well-known empirical observation that the χ^2 distance is well-suited for comparing histograms ([215] and Publication VII). Also other distance functions are sometimes used (cf. Section 3.4.1 and Publication X).

The free parameters C and γ of the SVMs are selected with a cross-validation search procedure. The details of the search procedure have been developed over the time. In recent work, a scheme has been used where the average precision (AP) performance measure is optimised in the 10-fold cross-validated training set using a procedure where—inspired by [210]—line search is first used to identify a promising parameter region, followed by a grid search in that region. Earlier, the ROC AUC performance metric, 6-fold cross-validation and an importance sampling search procedure were used, initialised with grid search. To speed up the computation, the data set has been radically downsampled for the parameter search phase. Further speed-up is gained by optimising the C-SVC cost function only very approximately during the search.

For the final SVM detectors, downsampling the data set has also been employed, but less radically than in the parameter search phase. Often there have been much fewer positive examples of a category than negative examples. Consequently, for such categories, the sampling has been able to retain all the positive examples and just limit the number of negative examples. The exact degree of sampling applied has varied according to the computation resources available and the required accuracy of the outputs. Generally, the downsampling has always resulted in degraded detection accuracy in SVM experiments with PicSOM.

Comparing the SOM and SVM based detection algorithms, the advantage of SVM-based detectors is their significantly greater accuracy. This is demonstrated by the Figure 3.22, adopted from Publication IV. This publication concerns the means by which the system performance was improved in detecting categories of Corel images. The figure shows that in that application, SVM-based detectors (the solid line with small dots labelled “Improved”) reach much higher accuracy than SOM-based detectors (the solid line with open circles, “Baseline”). SOM-based classifiers,

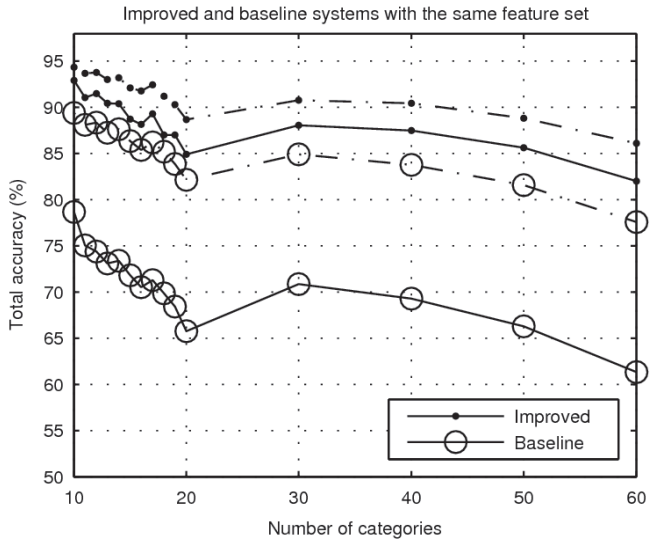


Figure 3.22. Solid lines: accuracy of the SOM-based baseline and improved SVM-based categorisation systems with a basic set of image features. Dashed lines: the performances with augmented set of features. From Publication IV.

in turn, are much faster to train. In particular, a new SOM-based classifier is very fast to generate to discriminate any pair of two semantic classes once the data set specific preparation step has been performed. This makes the SOM-based approach scalable to extremely large concept ontologies. For example, it has enabled the modelling of a total of 294 concepts from the LSCOM ontology [129] without excessive computational requirements [165].

Historically, the SOM-based detectors have always been part of the PicSOM throughout the system's existence. The early emphasis was on interactive CBIR, where the rapidness of the SOM approach in learning new category definitions is essential for a satisfactory user experience. Also much of the early work where the PicSOM system has been used in off-line category detection tasks used SOM-based detectors, including the work on automatic image annotation and participation to the VOC evaluations in 2005 and 2006. Later, the PicSOM system has been migrated towards the use of SVM detectors in all off-line tasks.

3.3.2 Learning algorithms commonly used for category detection

One can get an overview of learning algorithms employed in the successful category detection systems of the day by looking at the systems of the recent TRECVID evaluations. One can, for example, consider the top ten groups of the high-level feature extraction task of the TRECVID 2009 evaluation (cf. Section 2.4.2). All the systems are described in the TRECVID 2009 workshop proceedings [131]. A variant of the SVM learning algorithm (cf. Section 3.3.1) is employed in supervised learning subsystems of all ten groups. Interesting and somewhat useful, but apparently still not very essential variations to the basic SVM algorithm include the use of over/undersampling of the training set and boosting for SVM training, and the multiple kernel learning (MKL) SVM extension [149]. The sampling approach is employed in the second best ranked PKU system to combat the unbalanced numbers of positive and negative training examples. In their OnUm algorithm, multiple parallel SVMs are trained using versions of the training set with duplicated positive examples and undersampled negative examples. In the ninth ranked TSINGHUA system, a boosting algorithm [199] is used for training five parallel SVMs that complement each other for each feature. The third ranked FTRDBJ group tried to employ MKL for selecting weights of different components of feature vectors concatenated in the early fusion fashion (cf. also Section 3.4). Other alternative fusion algorithms worked somewhat better in their system, however.

Some of the top groups supplemented SVMs with other learning algorithms. The most accurate MediaMill system employs two other advanced kernel learning approaches in addition to SVMs: kernel discriminant analysis using spectral regression (SR-KDA) [23] and non-sparse multiple kernel Fisher discriminant analysis (MK-FDA) [208]. These additions can be considered to be significant for the system's performance. The same can not be said in the cases of the TIGT system, where hidden Markov models (HMM) are employed for classifying audio features, nor of the LIF system, where the simple k-nearest neighbour classifier (KNN) [156] is used in parallel with SVMs.

Outside the top ten of TRECVID 2009, examples of other interesting and decently well working learning approaches used in recent TRECVID HLFE tasks include IBM Research's random subspace SVM bagging (RS-Bag) and principle component semi-supervised SVM algorithms [24, 132],

random forest classifiers (e.g. [6, 143]) and Fisher discriminant methods [56, 170]. One may also mention the rather successful investigations of Microsoft Research Asia at TRECVID 2007 and 2008 [115, 116] on various supervised and semi-supervised learning algorithms. The supervised methods include SVM variants, multi-layer multi-instance kernel [214] and correlative multi-label learning [147], whereas the semi-supervised methods consist of optimising multi-graph learning [204], transductive multi-label learning [202], manifold ranking, linear neighbourhood propagation [201], and the use of temporally consistent Gaussian random field [179] models.

3.4 Fusion: early, late and intermediate

In fusion approaches to learning, the data is first processed in several different ways in parallel and then the results are combined. Over the years, feature fusion approaches have proven to be very useful in image analysis tasks. In those approaches, several alternative visual features are extracted from the images. The combination of the features usually performs better in a given image analysis task than any single one of the features that were combined. Feature fusion is especially useful when one detects a wide variety of categories from a generic data collection. Then it might simply be impossible to devote the effort to consider the usefulness of different features separately for each category. Consequently, one must resort to automatic methods for combining the information the different features provide. In narrower domains, it is more likely that one can come up with a single feature that rather well reveals whether an image belongs to categories of interest, or at least the difference between the performances of fusion and the best single feature could be less pronounced.

In the empirical spirit of this thesis, the practical observations of beneficiality of fusion give sufficient justification for applying feature fusion in a category detection system. However, it can be pointed out that also in biological visual systems colour, shape, motion and depth information seem to be processed using separate pathways [106]. In the machine learning research community, it is a long-known empirical fact that an ensemble of classifiers usually is much more accurate than any single classifier of the ensemble (see e.g. [5, 87]). Quite concretely this principle was brought to

the attention of computer vision researchers by the Viola-Jones face detection algorithm [198] that builds on the principle of combining a number of very rudimentary pixel pattern classifiers via a boosting algorithm. An overview of boosting in general can be found e.g. in [157].

In a category detection system, feature fusion needs not to be a separate system stage, but fusion techniques can be utilised in almost any phase of the category detection pipeline. There are two basic approaches to feature fusion: early fusion and late fusion, the latter being known also as post-classifier fusion. In early fusion, different features are combined in the earliest possibility after their extraction from images. The combination can then be treated just as a single feature in subsequent processing. In contrast, in the late fusion approach the features—often called feature channels—are processed separately as long as possible and only as the last step in the processing the outcomes of the feature channels are combined. Early and late fusion approaches represent opposite extremes. Feature fusion can also occur in the middle of a category detection system. The multi-level feature extraction hierarchies (cf. Section 3.2.2), for example, offer many possible intermediate points for fusion. Early and late fusion approaches are not exclusive: both can be simultaneously put to use in a category detection system.

As a demonstration of the early/late fusion continuum, Figure 3.23 shows a hypothetical BoV type system where interest points are first detected. Histograms of local colour (C) and texture (T) at these points are then used as features. The colour and texture feature vectors of each point could be concatenated to form a single feature vector (1). Alternatively, a separate histogram could be generated on basis of both channels and the histograms then concatenated (2). Third option (3) would be to feed both histograms as an input to a supervised detection algorithm that would produce a decision based on both (e.g. kernel fusion via multiple kernel learning, cf. Section 3.4.1). In the last alternative (4), the two histograms would be fed separately to SVM detectors. The decision scores of the SVMs could then be averaged or combined in some other manner. The first approach is clearly early fusion in its purest form, whereas the last qualifies as late fusion. Often approaches like the second and third one are also regarded as early fusion.

Another dimension to the variety of feature fusion approaches is incorporated by hierarchical fusion. In hierarchical fusion, the features are divided into groups. Feature fusion is first performed within the groups,

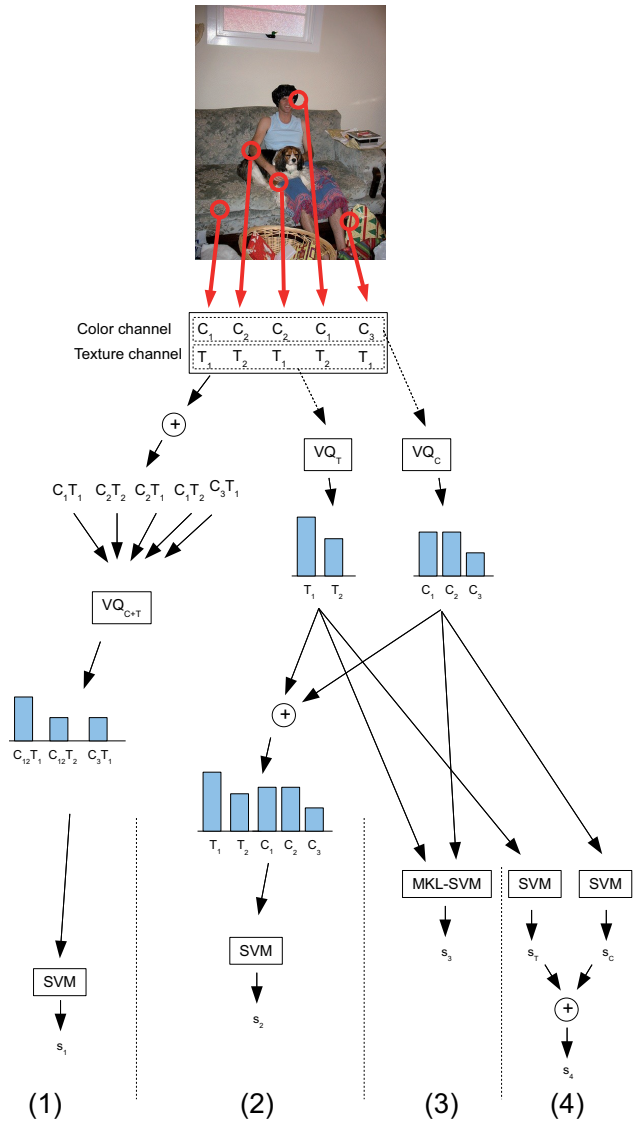


Figure 3.23. Fusion alternatives in a BoV system employing colour histograms (C) and texture histograms (T) as local image features.

and after that the outcomes of each feature group are combined. Quite often the features are grouped by heuristic grounds, for example visual, audio and textual features of videos. This can have practical reasons: some learning algorithms, for example, are conventional to apply to textual features, but are not well applicable to visual features.

It is not only the features that can be combined. Similar fusion techniques can be applied, e.g. in late fusion scheme where one trains several SVM detectors on basis of each feature with different training parameters. One can also train the detectors with different subsets of the training data and fuse the results, as is done for example in bagging type learning algorithms [21].

With feature fusion, one has to decide which features to fuse. One approach is to treat all the features symmetrically and fuse them together with equal weights. This gives arise to unsupervised fusion methods. However, one can decide to fuse just a subset of the available features or weight the features unequally. Naturally, one can use a priori information to do the selection, in which case also this kind of fusion is unsupervised. However, if one estimates the optimal combination strategy from the training data, the fusion becomes supervised. This links feature fusion to feature (or variable) selection and extraction methods, research problems that have been devoted a lot of attention over the years (see e.g. [67]).

Many feature extraction and supervised detection techniques can be interpreted to contain fusion-like elements. It is sometimes difficult to decide whether they should be addressed separately as fusion techniques or to be left outside the discussion as inherent details of the respective techniques. For example, the colour of image pixels—their most elementary property—is a combination of three more-or-less independent channels. Same applies to many multi-component elementary features, such as colour histograms. Speaking of an early fusion composite feature here, however, probably just confuses matters. Another example is the common feature extraction technique, where one geometrically partitions the image area and concatenates the feature vectors of the parts (cf. Section 3.2.1). Kernel fusion and bagging/boosting learning techniques further blur the boundary between fusion and learning methods. It should also be pointed out that fusion techniques and other parts of a category detection system are not isolated. For example, the particular way the outputs of supervised learners are scaled implicitly contributes to the weight-

ing of features in late fusion in the traditional PicSOM fusion approach.

There has been an enormous amount of research in applying various fusion techniques to many kinds of learning problems. A comprehensive review of the techniques or their applications is outside the scope of the thesis. The following section presents some fusion-related techniques that are used in the category detection systems of the day, the PicSOM system in particular.

3.4.1 Fusion techniques

For early fusion, the prototypical approach is to concatenate the feature vectors to be fused. In addition, one can perform some normalisation or weighting of the features. In PicSOM, for instance, a normalisation technique has been used where all the feature vector components have been made to have unit variance and zero mean.

Kernel fusion can also be categorised as an early fusion technique. In kernel fusion, the kernel-based learning algorithm (usually SVM) is provided with all the extracted feature vectors, and the learning algorithm then combines the feature channels. The combination weights can be either set a priori or learnt from the data. In multiple kernel learning (MKL) [149] the kernel of an SVM (cf. Section 3.3.1) is expressed as a linear combination of feature-wise kernels g_l:

$$g_{\mathrm{MKL}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_l \lambda_l g_l(\mathbf{x}_i, \mathbf{x}_j). \tag{3.8}$$

The combination weights λ_l are learnt simultaneously with the support vectors in the optimisation process of the SVM cost function. Another kernel fusion technique is the combination of feature-wise distances, resulting in the kernel

$$g_{\mathrm{dist}}(\mathbf{x}, \mathbf{x}') = e^{-\sum_l \gamma_l d_l(\mathbf{x}_i, \mathbf{x}_j)} \tag{3.9}$$

corresponding to multiplication of kernels. The distance combination technique is used with a priori weighting e.g. in pyramid matching kernels [66, 99] and their relatives to combine features extracted in different scales. This technique is employed also in Publications VIII, IX and XI. In the distance combination scheme of [113], the feature weights are separately optimised with a genetic algorithm. For many distance metrics d, the distance combination kernel fusion is equivalent to weighted concatenation type of early feature fusion.

In late fusion, the feature-wise detector outputs are mapped into a single detection score s through some fusion function:

$$s = f_{\mathrm{F}}(s_1, \cdots, s_{N_{\mathrm{D}}}) = f_{\mathrm{F}}(\mathbf{s}). \tag{3.10}$$

Here s denotes the vector that collects the outcomes of the feature-wise detectors:

$$\mathbf{s} = [s_1 \; s_2 \; \cdots \; s_{N_{\mathrm{D}}}]^T. \tag{3.11}$$

Some examples of widely used simple mapping functions are given below. A very elementary late fusion mechanism is forming the sum or equivalently the average of the detector outputs:

$$f_{\mathrm{avg}}(\mathbf{s}) = \frac{1}{N_{\mathrm{D}}} f_{\mathrm{sum}}(\mathbf{s}) = \frac{1}{N_{\mathrm{D}}} \sum_{i=1}^{N_{\mathrm{D}}} s_i. \tag{3.12}$$

In the summation, one may weight the detectors individually

$$f_{\mathrm{linear}}(\mathbf{s}) = \sum_{i=1}^{N_{\mathrm{D}}} \lambda_i s_i \tag{3.13}$$

resulting in weighted linear fusion. Product fusion or equivalently the geometric mean

$$f_{\mathrm{geom}}(\mathbf{s}) = f_{\mathrm{prod}}(\mathbf{s})^{1/N_{\mathrm{D}}} = \prod_{i=1}^{N_{\mathrm{D}}} s_i^{1/N_{\mathrm{D}}} \tag{3.14}$$

often performs somewhat better than the sum fusion. Also the product formula can be modified to include detector-wise weights.

As long as the mapping function does not contain free parameters, the late fusion can be performed in unsupervised, feed-forward mode. All of the above mentioned fusion mechanisms except the weighted sum are examples of such parameter-free fusion techniques. Mapping functions containing free parameters can also be used for unsupervised fusion if the parameter values can be set using a priori knowledge. Despite their simplicity, unsupervised sum and product fusion can provide surprisingly good fusion results, as demonstrated in Publication IV.

However, if the optimal values of the fusion mapping parameters are learnt from training data and associated feature-wise cross-validated detector outcomes, one obtains a supervised late fusion method. Basically any generic vector space supervised learning algorithm can be used for this purpose by interpreting the collection s of detector outcomes as a feature vector. Then the vector space learning algorithm can be used for selecting the optimal mapping function $f_{\mathrm{F}}(\mathbf{s})$ among the ones the learner

can form. Of the methods falling to this category, use of SVMs and neural networks has been popular and rather successful. In the PicSOM system, the SVM approach with RBF kernels has been used (cf. Section 3.3.1), as well as the Bayesian logistic regression [61] method. In the two-class case (category present or not), the method is called Bayesian binary regression (BBR). The mapping function in BBR is a weighted linear sum that is converted to the scale $[0, 1]$ via the logistic sigmoid:

$$f_{\mathrm{BBR}}(\mathbf{s}) = \varphi(\sum_{i=1}^{N_{\mathrm{D}}} \lambda_i s_i) \qquad \varphi(x) = \frac{e^x}{1+e^x}. \tag{3.15}$$

Because of the sigmoidal mapping, the values of f_{BBR} can be interpreted as probabilities. This facilitates the determination of the parameters λ_i as the MAP estimates if the prior distribution for λ_i is specified. This, employing Laplacian priors, is accomplished by the BBR software package [60]. In practice, BBR fusion often gives results comparable to SVMs, but is computationally considerably lighter (cf. e.g. Publication IX).

Another route to supervised fusion is to combine a late fusion mechanism with a wrapper method for variable selection. In wrapper methods, the conceptual idea is to train the actual category detection system with different subsets of available variables and evaluate which of the combinations works best. This evaluation requires the feature-wise detector outcomes for the training set (e.g. via cross validation), and thus the fusion procedure is supervised. The use of the wrapper approach needs not to be limited to variable selection. Also other parameters of the mapping such as the linear combination weights could be selected using this principle, as long as a search scheme is devised that produces candidate values for the parameters.

Sequential floating search methods are classical well-performing wrapper methods for variable selection [76, 146]. In floating search, the active set of variables is either augmented with one variable (sequential forward selection, SFS) or one variable is removed (sequential backward selection, SBS), whichever choice results in most improvement in the search cost function. The cost function can, for example, measure classification performance in the training set. Two often used initialisation strategies of the active variable set are starting from empty set, and starting from the full set of all variables. These alternatives have been denoted sequential floating forward (SFFS) and backward (SFBS) search.

The wrapper-based fusion approach used in PicSOM system for TRECVID 2009 builds on the geometric mean fusion mechanism. Floating

search is used for selecting such set of variables that maximises the AP ranking performance measure in the training set when the feature-wise detection scores are combined with the geometric fusion mechanism. The geometric mean fusion is then performed for the test set using the same set of features. The acronym SFBS (from sequential forward-backward search) has been used for the search procedure for indicating two directions of active set update (SFS, SBS), both in the forthcoming discussion in this thesis and in Publication VI. This use of the acronym is somewhat ambiguous, as in the above taxonomy of floating search methods, the same acronym has been used to refer to the search starting from the full set of variables.

In a modified version of the SFBS procedure, the search is made more robust by partitioning the training data into N_P folds. A set of overlapping training subsets is obtained by forming all possible N_P combinations of folds where one of the folds is left out. In the multifold-SFBS (mSFBS) fusion procedure, the search is repeated for all the training subsets and the results are used for weighting the geometric fusion:

$$f_{\mathrm{mSFBS}}(\{s_i\}) = \prod_{i=1}^{N_\mathrm{D}} s_i^{\frac{q_i}{Q}} \qquad Q = \sum_i q_i. \tag{3.16}$$

Here $q_i \in [0, \ldots, N_\mathrm{P}]$ is the number of training subsets in which the ith variable is included in the SFBS search result.

Besides the floating search methods, one may consider also the following procedure as a rudimentary manual wrapper method: the researcher selects, for one reason or another, some fusion parameter values and tests the system's cross-validation performance using all selected parameter combinations in turn. This kind of manual wrapper procedure is probably very commonly used among researchers.

3.4.2 Observations from experiments

In this section some fusion experiments that have been performed with the PicSOM system are first described. These experimental results are then supplemented with a brief review of fusion techniques employed by the systems that participated in the TRECVID 2009 high-level feature extraction task.

Table 3.6. Annotation quality obtained by the PicSOM system with different feature sets in Corel image annotation task.

System	DTMI / bits	F_1
Elementary	4.71	0.244
Binary	6.38	0.329
N-ary	6.51	0.350

Fusion results experiments with PicSOM

The annotation of the Corel images (cf. Section 2.4.2) on basis of a set of global image features was studied in Publication III. Table 3.6 illustrates the usefulness of early fusion in that setting. The annotation quality is shown in terms of DTMI and F_1 performance measures (cf. Section 2.3). For both measures, larger numerical value corresponds to better performance. In the experiments, the set of elementary visual features is that of Table 4.2 on page 110. For the system denoted "Elementary", a separate SOM-based detector is trained for each feature, and the detector outputs are then fused in a late fusion manner. The other systems are otherwise similar but the set of features consists of early fusion combinations of two ("Binary") or one, two, three and four elementary variables ("N-ary").

In Publication IV, methods of improving performance in detecting the categories of the Corel images were studied. Figure 3.24 adapted from that publication compares the performance of different sets of variables in the Corel category detection task when SVM-based late fusion is employed. The feature sets form a hierarchy of subsets of the feature set ALL, the full feature set of the system. ALL decomposes into feature sets FC and ELEM+IP. ELEM+IP is the union of ELEM, the ten elementary image features of Table 4.2 and IP, the set of six SIFT-based BoV features. FC denotes the set of twelve early-fusion combinations of features in ELEM. One observes that here the accuracies generally increase as more features are added to the feature set.

The last example comes from the experiments performed with the PicSOM system for TRECVID 2009 HLFE task [166] and is adapted from Publication VI. For TRECVID, a preliminary evaluation of the various post-classifier fusion algorithms was performed in a setting where the annotated training part of the video corpus was further partitioned to training and validation parts in 2:1 proportions. In this prelimi-

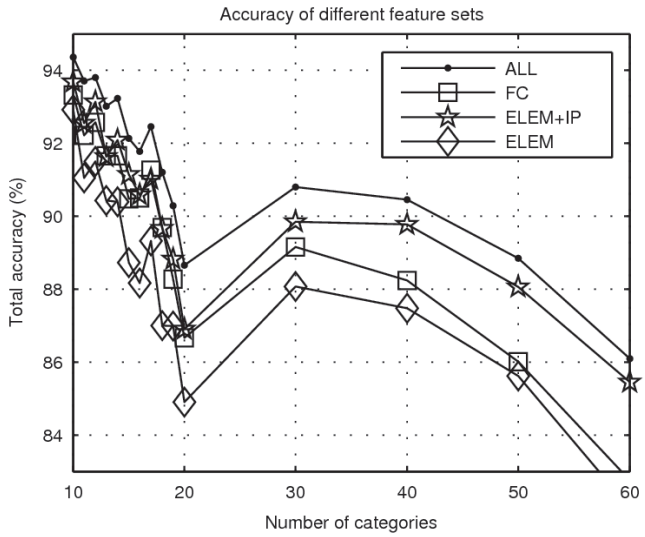


Figure 3.24. The accuracies of feature sets in detection of categories of the Corel images. From Publication IV.

nary experiment SVM-based fusion was significantly and consistently outperformed by geometric-mean-based fusion algorithms, both by the unsupervised basic version and by the supervised SFBS variants. Moreover, the SVM fusion mechanism is computationally much more costly. Consequently, the evaluation with the full data set was constrained to the variants of geometric mean fusion.

Figure 3.25 compares different geometric mean based fusion algorithms with the whole video corpus and four different sets D1–D4 of detectors to be fused. These sets result from different sets of shot-wise features and different training parameters. The number of fused detectors ranges between 77 (D1) and 26 (D4). One can see that the geometric mean of all detectors (the leftmost bar) is always inferior to methods where the set of detectors is selected with sequential forward-backward search (SFBS). The figure also shows that multifold-SFBS performs better than the basic SFBS.

The results of this section—when compared with the MIAP values of the best individual features in Tables 3.2 and 3.3—can be used for confirming the observation that fusion of features usually outperforms individual features, even if the best individual features are better than some other

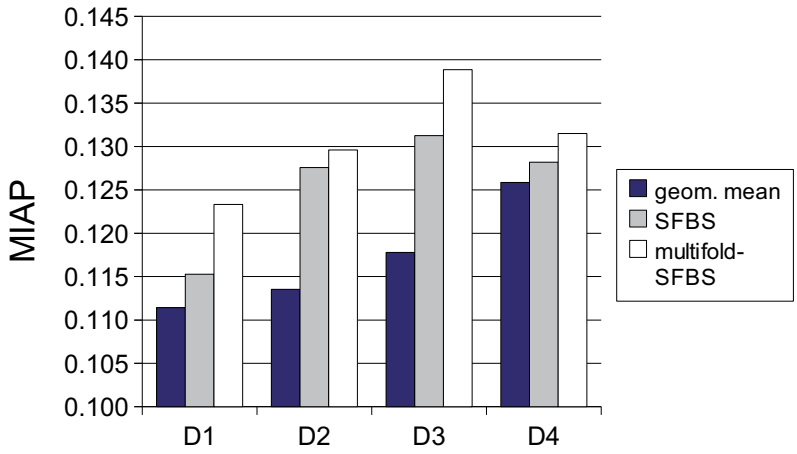


Figure 3.25. Comparison of algorithms for selecting detectors for geometric mean fusion from four different sets of detectors D1–D4. Adapted from Publication VI.

features being fused by quite large a margin. With a good fusion algorithm, benefit can be obtained from individually rather badly-performing features. In one experiment approximately 75% of the best features were chosen for fusion, thus leaving just the worst performing 25% of the features outside. Still, with the multifold-SFBS fusion algorithm the fusion accuracy improved when the worst 25% were returned to the feature set. With a less-developed fusion algorithm, the saturation point is reached earlier where further addition of features no longer improves the fusion result. An example of this behaviour can be seen in Figure 3.25 when comparing the sets of detectors labelled as D3 and D4. Here set D3 is a superset of D4 having almost three times as many detectors. When the geometric mean fusion is used, better performance is obtained by using the smaller set D4, whereas with the more advanced SFBS fusion algorithms the situation reverses: benefit can be obtained from the extra detectors in D3.

Fusion in TRECVID 2009 category detection systems

Systems that were the most successful in the high-level feature extraction task of TRECVID 2009 evaluation can be taken to represent the current state-of-the-art in category detection. Systems of all the top ten groups employ fusion strategies, as do most other participants of the evaluation. Many systems employ large and diverse sets of features (cf. Table 3.5), and generally this seems to be beneficial for the performance. Table 3.7

Table 3.7. Fusion techniques in the TRECVID 2009 HLFE top systems. See Table 3.4 for the key of the system abbreviations.

Rank	System	MIAP	Fusion techniques
1	MediaMill	0.228	late fusion with SFS feature selection
2	PKU-ICST	0.203	early and late
3	FTRDBJ	0.170	late fusion with hierarchical linear combination
4	TIT	0.168	late fusion via combined likelihood ratio
5	CityUHK	0.163	late fusion with the average rule
6	PicSOM	0.152	late fusion with multifold-SFBS
7	VITALAS	0.147	early fusion
8	EURECOM	0.138	hierarchical weighted linear fusion
9	Tsinghua	0.134	late fusion with the average rule
10	LIFM	0.132	hierarchical weighted linear fusion

summarises the used fusion techniques.

If one considers all the systems participating in the TRECVID 2009 HLFE task, the late fusion approach seems to be more common than early fusion. Some researchers are even of the opinion that the superiority of late fusion over early fusion is already a proven matter. However, a significant portion of the systems rely on early fusion, some with a degree of success. In the second ranked PKU-ICST system, features are divided—apparently heuristically—into two groups. Within these groups, early fusion is employed. The detection outcomes based on the two feature groups are combined in a late fusion manner. The third-ranked FTRDBJ group obtained their best result with hierarchical late fusion mechanism. The first layer of the hierarchy employs the average fusion mechanism within feature groups while in the second layer a weighted linear combination is formed. However, the group obtained almost equally good results by employing kernel fusion based on multiple kernel learning (MKL), which can be put into category of early fusion methods.

4 Application examples

This chapter briefly reviews some successful applications of the generic PicSOM category detection architecture to diverse visual analysis tasks. The covered tasks include automatic image annotation, object detection, localisation and search, semantic multimedia search, and robot navigation. The material in the following sections has been duplicated for the most part from Publications I–VI of the thesis and can be considered to be a contribution of the respective publications.

4.1 Automatic image annotation

Publications III and IV of the thesis investigate the automatic image annotation task. In that task, images are automatically labelled with keywords. During the last decade, a wealth of automatic annotation approaches has been proposed in the literature, starting with the pioneering work of [126]. Many of the proposed methods have been specifically designed for the auto-annotation application. The leading thought behind the experiments discussed in this section has been to try how good an annotation performance one is able to achieve by following a somewhat different approach. The generic PicSOM visual category detection architecture has been used as a back-end. This generic image content analysis is followed with a front-end that has been tailored particularly to the annotation task. The approach has been tested by comparing the resulting annotation performance against methods reported in literature, replicating the experimental setups the methods have been demonstrated with. In practice, this means mostly annotation tasks defined in the Corel image collection (cf. Section 2.4.2). The performance measures mentioned already in Section 2.3.2 and discussed below in greater length are used for evaluating the performance in the tasks. The experiments give insight into the properties of these performance measures, in addition to helping

in understanding the system that is used for generating the annotations.

Image annotation experiments were among the first attempts to widen the application area of the PicSOM framework from the original interactive image retrieval to include also off-line image analysis tasks. Therefore, the first experiments have been performed with early versions of the category detection system whose components were not so well-developed as in the more recent experiments. In practice, this translates to a limited set of elementary visual features and SOM-based classifiers. The later annotation experiments demonstrate how the annotation performance can be improved by improving the components of the category detection system. Another development has been the widening of the range of tasks addressed. After initial experiments, the same back-end was used in tasks that required optimising different measures of annotation quality by modifying the front-end, and in different image collections. This has encouraged widening the application area of the PicSOM back-end even further, beyond image annotation and interactive image retrieval.

Here the different variants of image annotation tasks are defined as follows. In a one-to-one annotation task, the keywords are exclusive, i.e. each image is annotated with exactly one keyword. One-to-many annotation is a generalisation of this: any number of keywords up to an upper limit $w_{\max}$ can be associated with an image. For use in the forthcoming discussion, the number of keywords associated with an image is denoted as the annotation length.

Image categorisation is a term closely related to image annotation. The usage of the terms in the literature has been somewhat inconsistent as they have been used for either similar or slightly different variants of image content analysis tasks. Quite often the term “annotation” has been reserved for one-to-many annotation, and the term “categorisation” has been used for one-to-one annotation.

4.1.1 Performance measures of automatic image annotation

For given training and test sets of images and keywords, the annotation problem can be reworded as “maximising the quality of the predicted annotations.” The solution is thus inherently determined by the used quality measure. In the following, the cursory treatment of Section 2.3.2 on the quality measures of binary decision matrices is expanded to discuss the measures in greater detail, as the practical examples in this chapter

demonstrate the properties of the measures and thus support the discussion.

In the case of one-to-one annotation, performance measurement is a simple matter. Usually it is sensible to just calculate the fraction of the predicted category labels that match the ground truth. This is the case in the reported one-to-one experiments, where one tries to detect the thematic categories of Corel images.

In one-to-many annotation, the situation is more subtle and several different measures have been proposed in the literature. Common to all of them is that the measures attain their maximum value for a perfectly predicted annotation. The difference between the measures is in the way they punish different kinds of perturbations from the perfect predictions.

The discussion below follows that in Publication III. Two most widely used performance measures for binary annotation—the normalised score (NS) and precision/recall (PR) statistics–are first discussed. Then also a measure proposed in that publication—the de-symmetrised termwise mutual information (DTMI)—is described and motivated as a principled compromise between these two extremes. In addition to the discussion below, the experiments described in Sections 4.1.2 and 4.1.3 give some insight to these performance measures.

Normalised score

The average normalised score (NS) [11, 193]

$$\mathrm{NS} = E_i\left[\frac{c(i)}{w(i)} - \frac{n(i)}{W - w(i)}\right] \tag{4.1}$$

is a performance measure widely used in literature. Here $E_i[\cdot]$ denotes the average over the test set images i, $w(i)$ is the ground truth annotation length, $c(i)$ is the number of correctly predicted keywords and $n(i)$ that of incorrect predictions. W is the size of the vocabulary. As is desirable, the measure rewards correct predictions (sensitivity) and punishes incorrect predictions (specificity). The balance between these components is somewhat arbitrary. In practical cases with a large vocabulary, the sensitivity component largely dominates over the specificity because of the small value $\frac{1}{W-w(i)} \approx \frac{1}{W}$ of the balancing constant in the formula.

The NS-measure treats all the keywords equally in the sense that it only determines whether a prediction is correct or not. No attention is paid on the keyword being frequent or rare. Usually, but not always, the NS-measure is calculated for predicted annotations of unlimited maximum

length $w_{\max}$.

Precision and recall

Measuring annotation quality by average precision (P) and recall (R) statistics [110] (see Equations 2.1 and 2.2) is also common in literature. The PR-statistics are motivated by the image retrieval setting where annotations are first predicted in a binary manner for a set of test images and the image set is then queried by one keyword at time, based on the predicted annotations. For an individual keyword w the precision P_w and recall R_w are defined as

$$P_w = |w_c|/|w_{\mathrm{pred}}|, \quad R_w = |w_c|/|w_{\mathrm{gt}}|, \tag{4.2}$$

where $|w_{\mathrm{gt}}|$ and $|w_{\mathrm{pred}}|$ are the number of occurrences of keyword w in the ground truth and the predicted annotations, respectively, and $|w_c|$ of the predicted annotations are correct. To arrive at statistics that describe all the keywords simultaneously, the precision and recall values are averaged uniformly over the keywords:

$$P = E_w[P_w], \quad R = E_w[R_w]. \tag{4.3}$$

When different information retrieval systems are compared in terms of precision and recall, a single performance measure is often desirable. One possibility is the F-measure [189]

$$F_\alpha = \frac{(1+\alpha)PR}{\alpha P + R}. \tag{4.4}$$

By setting the parameter $\alpha = 1$ one obtains the commonly used balanced F-measure, the harmonic mean of precision and recall.

A technical shortcoming of the PR-statistics is that the average precision and recall are normally evaluated separately. In the sense of this measure, it may then be optimal to annotate images so that some keywords have high values of precision and different keywords high values of recall. Another deficiency is the characteristic of the PR-statistics that results from the uniform averaging scheme of Eq. (4.3): All the keywords are treated equally in a certain sense: for the measure, it is equivalent, for instance, to predict correctly all, say, 50 occurrences of keyword “car” or just the one occurrence of keyword “saguaro” in the set of test images. In the image retrieval setting, this would correspond to the assumption that “saguaro” queries are made as frequently as “car” queries.

When calculating PR-measures, the length of annotations is often limited to some maximum value $w_{\max}$. Five keywords has been a common

limit in the experiments discussed in literature, but also other values have been used.

De-symmetrised termwise mutual information

The NS- and PR-measures represent two extremes in the way different keywords are weighted. The NS-measure essentially considers each individual prediction to be equally important, regardless of the predicted keyword. The PR-statistics in turn consider the fraction of the occurrences of a keyword that can be predicted. A single prediction may have either very large or small contribution to that fraction, depending on the prevalence of the predicted keyword.

In information theory, the information content of a random variable is commonly measured with its entropy [161]. For a discrete random variable X with support $\mathcal{X}$, entropy $H(X)$ is given by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{4.5}$$

Entropy represents a well-grounded compromise that balances the importance assigned to rare and common values of the random variable. Using this definition of information content, mutual information $I(X, Y)$ measures the amount of information that is revealed of the value of a random variable X by the value of another random variable Y. Mutual information can be expressed using entropies

$$I(X; Y) = H(X) - H(X|Y). \tag{4.6}$$

Here $H(X|Y)$ is the conditional entropy of X given Y:

$$H(X|Y) = \sum_{y \in \mathcal{X}} p(y) H(X|Y = y) = -\sum_{x \in \mathcal{X},\ y \in \mathcal{Y}} p(x, y) \log p(x|y), \tag{4.7}$$

where $\mathcal{Y}$ is the support of Y. The de-symmetrised termwise mutual information (DTMI) annotation quality measure, introduced in [193] and Publication III, essentially measures the mutual information between the ground truth and predicted annotations. However, two adaptations appear necessary. On one hand, the estimation of the measure from a limited number of samples must be made practical. On the other hand, the undesirable symmetry of mutual information to the presence/absence of the keywords in the annotation must be removed. For example, the mutual information of the following two cases would be equal: 1) the desirable case of a certain keyword appearing in the predicting annotation exactly when the same keyword appears in the ground truth, and 2)

the complement of the previous case, the keyword being predicted exactly when it does not appear in the ground truth.

The two requirements are fulfilled by the measure

$$\mathbf{DTMI}(\mathbf{a}^{\mathrm{gt}}, \mathbf{a}^{\mathrm{pred}}) = \sum_{w=1}^{W} \left[\mathrm{H}(a_w^{\mathrm{gt}}) - \hat{\mathrm{H}}(a_w^{\mathrm{gt}} | a_w^{\mathrm{pred}}) \right], \tag{4.8}$$

where a_w^{gt} and a_w^{pred} are the binary random variables indicating the presence of the keyword w in the predicted and ground truth annotations,

$$\mathbf{a}^{\mathrm{gt}} = [a_1^{\mathrm{gt}}\ a_2^{\mathrm{gt}}\ \cdots a_W^{\mathrm{gt}}]^T \qquad \mathbf{a}^{\mathrm{pred}} = [a_1^{\mathrm{pred}}\ a_2^{\mathrm{pred}}\ \cdots a_W^{\mathrm{pred}}]^T, \tag{4.9}$$

and

$$\hat{\mathrm{H}}(X|Y) = \begin{cases} \mathrm{H}(X|Y) & \text{if } p(X=1|Y=1) \geq \\ & \quad p(X=1|Y=0) \\ 2\mathrm{H}(X) - \mathrm{H}(X|Y) & \text{otherwise} \end{cases}. \tag{4.10}$$

For a reasonably well performing annotation system, the DTMI-measure essentially coincides with mutual information. Therefore, it enjoys the same interpretation in terms of the length of an optimal code as does the mutual information. Mutual information measures how much the optimal code describing the ground truth annotations can be shortened if the predicted annotations are revealed to both the coder and the decoder. The entropy in the ground truth annotations, i.e. the original code length, gives a natural scale to the DTMI measure. One can, for example consider the ratio DTMI/H. The DTMI measure can be evaluated for annotations of limited as well as unlimited length.

4.1.2 Application of PicSOM system to image annotation

The application of the PicSOM image category detection system to the annotation problem consists of two phases. For the first phase, each keyword appearing in the training annotations is used for defining a separate image category. The generic PicSOM image category detection architecture is employed for providing a detections score of each image category for each of the N_{C} test images that are to be annotated. This results in a $N_{\mathrm{C}} \times W$ matrix $\mathbf{S}$ of detector outcomes.

After the generic first phase, the second phase is specific to the automatic annotation application. In that phase, such annotations are assigned to the test images that the expected value of the chosen performance metric is maximised, given the matrix $\mathbf{S}$. The approach employed

in Publication III converts the SOM detector outputs to probability estimates. To this end, a logistic sigmoid model has been used, optimised with the Newton-Raphson iteration technique. Given the probability estimates, the expected values of the performance measures are easy to evaluate for any choice of annotating keywords. What remains to be done is to select promising candidate annotations to be evaluated. The task is straightforward for one-to-one annotation: just the category with maximum estimated probability is chosen for each image. Also the NS measure can be optimised separately for each image. In practice, this almost always leads to predicting the $w_{\max}$ keywords with largest probability estimates, if the estimates are not exceptionally low.

If $w_{\max}$ is unlimited, optimising the PR and DTMI measures is also simple as the annotations can be performed separately for each keyword. In [193] this case of unlimited $w_{\max}$ is considered. There a prediction threshold can be set to each keyword without converting S to probability estimates. With a limit set on $w_{\max}$, however, the annotations of different keywords and images become fully coupled. In Publication III, one resorts to an approximate iterative search algorithm when searching for a reasonably good solution.

4.1.3 Results of image annotation experiments

Table 4.1 extracted from [193] summarises the results of the first attempt in using PicSOM for automatic image annotation. In those experiments, the annotation was performed for subsets of the Corel image collection defined in [11]. The PicSOM annotations were performed for two of the ten subsets that were originally defined. The two sets of PicSOM results are for the two data sets 006 and 008. For [11, 125] the results are averages of ten and nine data sets, respectively. The result of Monay et al. is for their method LSA 2. The method of Barnard et al. has identifier binary-D-2-region-cluster in [11]. In the table, $\mathrm{NS}_{\mathrm{prior}}$ denotes the maximum of the NS score that can be obtained by assigning a fixed annotation for every image regardless of its visual contents. ΔNS denotes the NS increase over this number. The DTMI values in the table can be put into a natural scale by comparing them with the unconditional entropy values in the rightmost column. It can thus be said that the annotations produced by the PicSOM system are able to capture approximately 25 % of the information contained in the ground truth annotations of the images.

Method	ΔNS	$\mathrm{NS}_{\mathrm{prior}}$	DTMI	$\sum_i H(\mathbf{a}_i^{gt})$
$\mathrm{PicSOM}_{\mathrm{NS,006}}$	**0.262**	0.406	2.95	14.66
$\mathrm{PicSOM}_{\mathrm{DTMI,006}}$	0.115	0.406	**3.78**	14.66
$\mathrm{PicSOM}_{\mathrm{NS,008}}$	**0.213**	0.448	2.56	14.27
$\mathrm{PicSOM}_{\mathrm{DTMI,008}}$	0.070	0.448	**3.26**	14.27
Monay et al. [125]	**0.153**	0.383	-	-
Barnard et al. [11]	**0.179**	0.425	-	-

Table 4.1. Results of annotating subsets of Corel images defined in [11]. From [193].

The two sets of PicSOM results are for the two data sets 006 and 008. For [11, 125] the results are averages of ten and nine data sets, respectively. The result of Monay et al. is for their method LSA 2. The method of Barnard et al. has identifier binary-D-2-region-cluster in [11]. The results in the table show that competitive annotation performance could be achieved by a straightforward application of the PicSOM system of that time.

In Publication III, a one-to-many annotation problem with 5000 Corel images was considered. The task was defined by Duygulu et al. [43], and has been subsequently used by many others, e.g. [27, 78, 98], for demonstrating their systems. The data set consists of 5000 images from 50 commercial Corel stock photograph CDs. The images are partitioned into 4500 training and 500 test images. In this one-to-many annotation task, a 1–5 word ground truth annotation is given for each image. The vocabulary of the annotations consists of 374 keywords. 371 of the keywords appear in the training set and 263 in the test set.

Annotations were determined using the PicSOM system by optimising the three different performance measures: normalised score (NS), the F_1-measure and DTMI. The category detection part of the system employs early-fusion combinations of the visual features of Table 4.2 and SOM-based classifiers. Figure 4.1 shows selected examples of the annotations. Table 4.3 summarises the obtained performance values in the test set. Table 4.4 and Figure 4.2 compare the accuracy of the PicSOM annotations to the state-of-the-art methods of the day in terms of the F_1-measure. The PicSOM system is based on a set of low-level global visual features, hence the abbreviation PicSOM-g.

The annotation examples in Figure 4.1 show that even though the quan-

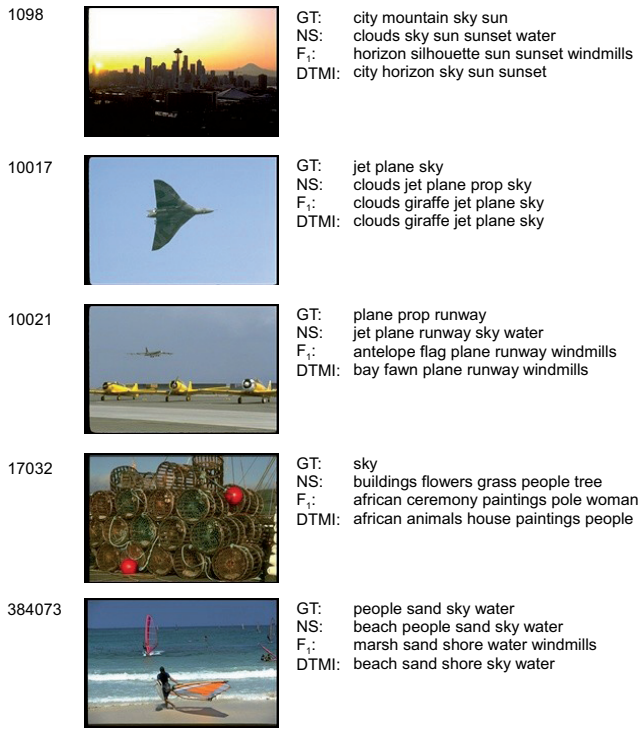


Figure 4.1. Examples of Corel images, their ground truth annotations (GT), and predicted annotations selected on basis of maximising the expected values of the NS, F_1 and DTMI performance metrics.

titative performance of the PicSOM annotation system is very good in comparison to many competing systems, there still is room for many inaccurate predictions on this performance level. Completely unsuccessful annotations (image 17032) and less extreme cases (like the image 10021) are not uncommon amongst the predicted annotations. This puts the positive example cases used for demonstrating almost all the annotation systems into perspective. If the quantitatively measured performance is not much better than that of the the PicSOM system, there simply must be also plenty of examples where the annotations are incorrect.

The example figures also show that one should not expect total agreement of the predictions and the ground truth keywords, even if the annotation system was working very well. The manual annotations are subjective, and therefore the ground truth does not include all the keywords

Feature	Tiling	Dim.
DCT coefficients of average colour in rectangular grid	global	12
CIE L*a*b* colour of two dominant colour clusters	global	6
MPEG-7 EdgeHistogram descriptor	4×4	80
Haar transform of quantised HSV colour histogram	global	256
Average CIE L*a*b* colour	5	15
Three central moments of CIE L*a*b* colour distribution	5	45
Co-occurence matrix of four Sobel edge directions	5	80
Magnitude of the 16×16 DFT of Sobel edge image	global	128
Histogram of four Sobel edge directions	5	20
Histogram of relative brightness of neighbouring pixels	5	40

Table 4.2. The elementary visual features extracted from the images. All the features are encoded as feature vectors whose dimensionality is given in the rightmost column. The first four rows correspond to features that more or less closely resemble the ColorLayout, DominantColor, EdgeHistogram and ScalableColor features of the MPEG-7 standard [75]. The features calculated for five tiles employ the tiling mask of Figure 3.2. From Publication III.

		Measure evaluated		
		NS	F_1	DTMI
Optimised for	NS	**0.548**	0.293	6.51
	F_1	0.359	**0.352**	5.97
	DTMI	0.454	0.318	**6.65**

Table 4.3. Annotation accuracy in the 5000 Corel image collection. Rows correspond to different systems versions of the PicSOM system optimised for each one of the three performance measures. From Publication III.

that would be valid annotations for an image. An example is seen in image 384073. The predicted annotations “beach” and “shore” make perfect sense even though they do not match the ground truth.

Image 100017 illustrates the fact that the image analysis happens on quite coarse a level. The system might be able to deduce that the scene probably is about an aircraft flying against the sky. In such scenes, there often are also clouds visible, although the resolution of the image analysis is not sufficient to confirm this with certainty. The keyword “clouds” is predicted anyway, since it often is a correct prediction in a similar situation.

Table 4.3 and Figure 4.1 confirm that the same annotations are not optimal in terms of all the three performance measures. However, in each case the optimal annotations can be determined from the same probability estimates, and thus using the same category-detection back-end. By

Method	P	R	F_1	Reference
Co-occ	0.03	0.02	0.02	[126]
Trans	0.06	0.04	0.05	[43]
CMRM	0.10	0.09	0.10	[78]
TSIS	0.10	0.09	0.10	[28]
MaxEnt	0.09	0.12	0.10	[79]
CRM	0.16	0.19	0.17	[98]
CT-3×3	0.18	0.21	0.19	[211]
CRM-rect	0.22	0.23	0.23	[52]
InfNet	0.17	0.24	0.23	[117]
MBRM	0.24	0.25	0.25	[52]
MixHier	0.23	0.29	0.26	[27]
PicSOM-g	0.35	0.35	0.35	Publication III

Table 4.4. PR annotation results reported for the data set of Duygulu et al. [43]. Key to the method abbreviations: Co-occurence Model (Co-occ), Translation Model (Trans), Cross-Media Relevance Model (CMRM), Text Space to Images Space (TSIS), Maximum Entropy model (MaxEnt), Continuous Relevance Model (CRM), 3×3 grid of colour and texture moments (CT-3×3), CRM with rectangular grid (CRM-rect), Inference Network (InfNet), Multiple Bernoulli Relevance Models (MBRM), Mixture Hierarchies model (MixHier), and the PicSOM system of Publication III employing global image features (PicSOM-g). Adapted from Publication III.

looking at the example annotations, one can notice some specific characteristics of the three measures. The NS annotations are the most conservative, always trying to maximise the probability of each prediction being a correct one. The F_1 and DTMI measures balance the correctness with the informativeness of keywords, so that sometimes a bit less certain rare and informative keyword is favoured over the safer choice of a common and easily detectable one. Image 1098 in Figure 4.1 shows examples of this. All the keywords of the NS-optimal annotations are very common, while “horizon” and “silhouette” are much more specific and informative. Of course, this approach leads also to increased probability of false alarms, such as “windmills”. From the same example image, one can also note that all the suggested annotations include the keyword pair “sun”–”sunset”. Actually, in the ground truth annotations these words never co-occur. This demonstrates how the annotation system works by treating each keyword separately, without the capability to utilise such co-occurrences known a priori.

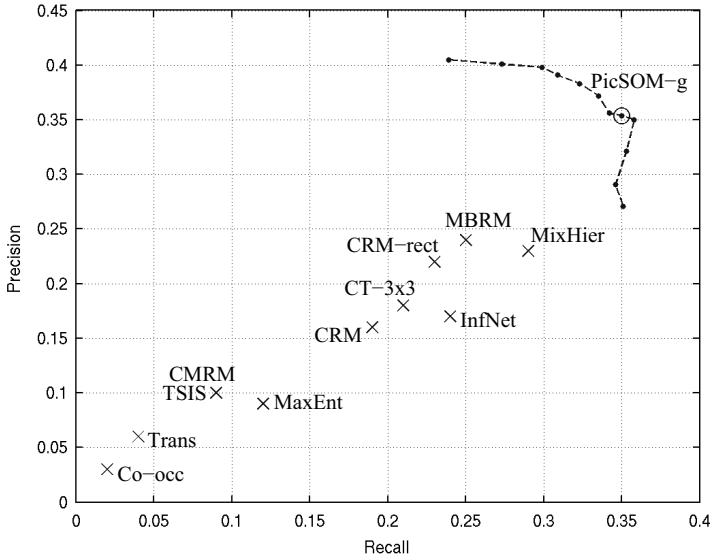


Figure 4.2. PR annotation results reported for the data set of Duygulu et al. [43]. See Table 4.4 for the key to the method abbreviations. From Publication III.

Image 17032 in Figure 4.1 illustrates another property of the optimal annotations. The ground truth annotation of this image consists of just one keyword. Since the detectors for keywords are not very reliable, the optimal keyword selection strategy usually predicts the full set of five allowed keywords for the image, even though some of the predictions probably go wrong. Given the far-from-perfect general prediction accuracy, the metrics just do not penalise the wrong predictions heavily enough to counterbalance the occasional right guesses of uncertain keywords. In this case, all the predicted keywords happen to be wrong.

Figure 4.3 that is taken from Publication IV displays the performance obtained with the PicSOM category detection system in another annotation task. The PicSOM performance is compared it with some systems reported in the literature in the task of one-to-one annotation of Corel images according to their thematic categories (see Section 2.4.2). In Figure 4.3, the baseline version of the PicSOM system is the same as the PicSOM-g system discussed above. The improved version of the system

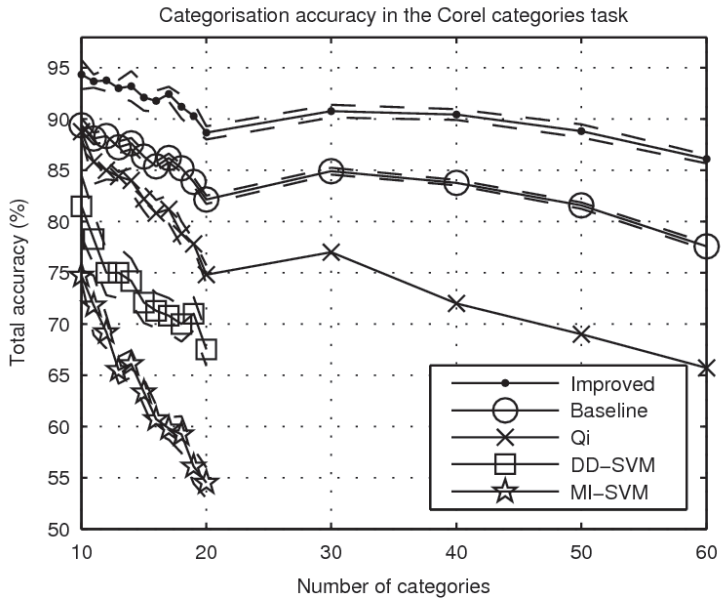


Figure 4.3. Total image categorisation accuracies of different image categorisation systems in the Corel categories task. From Publication IV.

adds SIFT-based BoV features to the feature set and replaces SOM-based classifiers with support vector machines. In the figure, the abbreviation Qi refers to method proposed in [148]. DD-SVM method was introduced in [31] and the MI-SVM in [8]. These are different SVM-based multiple instance learning methods. The results in the figure confirm PicSOM's competitive performance level also in this variant of image annotation.

4.2 Object detection, localisation and segmentation

In recent years, the PASCAL Visual Object Classes (VOC) Challenge has become a standard benchmark in the area of object detection and localisation from images (cf. Section 2.4.2). The PicSOM category detection system has been used for participating in the challenge in in years 2005–2008. The main focus has been in the classification task of the challenge, where the success has been fair. Never has PicSOM been the best one of the classification systems, but neither has its performance ever been too far behind the best systems.

Figure 4.4. Examples of VOC Challenge 2007 images and their annotations. From Publication II.

The example included here has been adapted from Publication II and originates from the VOC Challenge of year 2007. In that year, the PicSOM group participated in three of the challenge competitions: the main competitions in image classification and object detection, and a smaller-scale taster competition in pixel-level object segmentation. The image collection for the 2007 challenge consisted of 9963 photographic images of natural scenes. Each of the images contained at least one occurrence of the pre-defined 20 object classes. The images of the training half of the collection had been manually annotated with the bounding boxes of all the occurrences of the 20 object classes. Additionally, pixel-wise segmentation masks of approximately 8% of the training images were provided as training material for the segmentation taster competition. Figure 4.4 shows some examples of the images and objects.

In the classification task the test images were ranked according to their likelihood to contain objects from each one of the 20 classes. This is the kind of category detection task for which the PicSOM category detection architecture is directly applicable by taking the presence or absence of an object of a certain class to define a category. The classification system used ten global or tiled image descriptors—such as colour and edge histograms—along with several variants of SIFT-based BoV features. The elementary features were combined in an early fusion stage to give 141 composite features, which were fed into 141 SVM classifiers. The late fusion stage combined greedy forward/backward feature selection procedure with a SVM-based fusion mechanism.

In the detection task the goal was to find and rank bounding boxes of objects of the 20 classes. The task was addressed with an approach outlined in Publication I. The system first segments the images with a rather rudimentary hierarchical region merging technique and then assesses the likelihood of each segment to correspond to the targeted object class. This problem is factored into a product of two parts: 1) the probability of the

image containing the segment to contain the object somewhere, and 2) the conditional probability of just the considered segment to correspond to the targeted object, given that the object is known to appear somewhere in the image. More formally,

$$p(I_\mathrm{s} = 1|s_\mathrm{i}, s_\mathrm{s}) \approx p(I_\mathrm{s} = 1|I_\mathrm{i} = 1, s_\mathrm{s})p(I_\mathrm{i} = 1|s_\mathrm{i}). \tag{4.11}$$

Here I_s and I_i are binary indicator variables for an image segment and the image containing that segment, respectively, belonging to a certain object class. s_i and s_s are the corresponding category detector outputs.

The first probability in the product was readily available from the solution to the classification task. The latter probability was obtained by applying the category detection framework to image segments instead of whole images. For this purpose, a somewhat reduced feature set, but otherwise similar to that of the whole images, was used for describing the image segments. It was augmented with two types of segment shape features: Zernike moments and Fourier descriptors of segment contour.

Figure 4.5 shows the median APs of the classification competition participants over all the 20 object classes. The challenge participants are identified by abbreviations, the methods described here by "TKK". In the classification contest, the PicSOM group placed 4th out of 19 participants. The achieved median AP was 0.506, the best being 0.575. The detection performance of PicSOM was rather modest, 7th best of 9 participants, when measured by counting class-wise placements in top-three of detection accuracy. The PicSOM method was in the top-three for five object classes. The object segmentation accuracy was determined for each object class by calculating the percentage of actual pixels of the class that have received the correct label. Mean accuracy over all the 20 classes and background class was used to rank the participants (Figure 4.6). In this ranking, the PicSOM method was the best among 7 participants, with mean accuracy of 30.4%. All the segmentation contest entries except one were automatically derived from the bounding box detections.

4.3 Semantic multimedia search

In this section, it is described how the outlined category detection architecture is used as a part of a semantic multimedia retrieval system. The example is based on the PicSOM group's participation in the search task

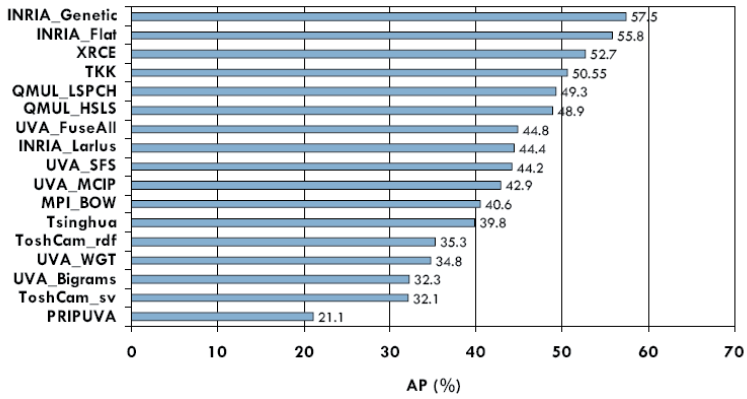


Figure 4.5. Results of the VOC Challenge 2007 classification competition: the median AP of all the 20 object classes [44].

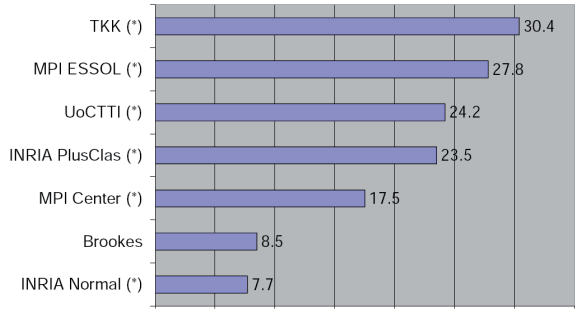


Figure 4.6. Mean Segmentation accuracies in the VOC Challenge 2007 segmentation taster task. Asterisk (*) denotes the entry being automatically generated from detection task results [44].

of TRECVID 2009 multimedia retrieval evaluation campaign [166] that is described in Publication VI. An overview of TRECVID was given in Section 2.4.2. For video search, TRECVID specifies three modes of operation: fully-automatic, manual, and interactive search. Manual search refers to the situation where the user specifies the query and optionally sets some retrieval parameters based on the search topic before submitting the query to the retrieval system. In this section, only the case of fully automatic video search is considered.

In 2009, the search task was performed on a video corpus that consisted of approximately 380 hours of Dutch television programmes: documentaries, news reports and educational programming. To enforce a uniform framework of retrieval among the search task participants and thus fa-

Table 4.5. Examples of the search topics of TRECVID 2009 search task

1.	Find shots of a road taken from a moving vehicle through the front window.
2.	Find shots of a crowd of people, outdoors, filling more than half of the frame area.
3.	Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible.
4.	Find shots of a person talking on a telephone.
5.	Find shots of a closeup of a hand, writing, drawing, coloring, or painting.
6.	Find shots of exactly two people sitting at a table.

cilitate the meaningful comparison of results, the TRECVID video material used in the search tasks was divided in advance into approximately 133 000 shots. These reference shots were then used as the unit of retrieval [141].

Given the shot partitioning, the search task participants were given 24 search topics, some of which are listed in Table 4.5. For each topic, the participants were required to return a ranked list of 1000 video shots most likely to correspond to the topic. The search topics consisted of a textual description along with a small number of both image and video examples of the information need. In addition to the raw video material in the corpus, the participants could use these mid-level concept detections of the same year's high-level feature extraction task as source of information. Additionally, some groups contributed concept detection scores for larger concept lexicons. Furthermore, the output from an automatic speech recognition (ASR) software was provided to all participants together with automatic machine translation of all non-English material. Due to the multi-step process, however, the quality of this textual data was remarkably poor.

In the following, the particulars of the PicSOM system for TRECVID 2009 are first outlined, after which the performance of the system in the search task is reported.

Figure 4.7. Processing stages of a video retrieval system. Adapted from Publication VI.

4.3.1 Parts of the PicSOM video retrieval system

Following the presentation in Publication VI, Figure 4.7 schematically shows processing stages of a generic video retrieval system —the PicSOM system being one example of such a system. The operation of the system generally consists of two phases. In the first phase, the system is prepared for a video corpus. The corpus is divided into an annotated training part and an unannotated testing part, on which video retrieval is going to be performed in the following search phase. The preparation phase is allowed to be time-consuming as it is intended to be performed off-line prior to the actual on-line use of the retrieval system.

Preparation phase

In the preparation phase the whole video corpus is first segmented into shots and the annotations are associated with the shots. For the present experiments with TRECVID video material, the openly available master definition of shots [141] is employed. The shot-wise training annotations originate from collaborative annotation efforts [10] organised in both years among the TRECVID participants.

Another preparation phase task is the extraction of one or more keyframes from each video shot. The keyframes are needed both for extracting visual features to describe the content of the shot and for presenting them to the users of the system as still replacements for the dynamic video content. In the PicSOM system, the keyframes are selected on the basis of a score that is increased for frames close to the temporal centre of the shot. The scoring scheme penalises frames different from the average image calculated over the shot, as well as frames with large temporal changes in the content.

A central task in the preparation phase is the extraction of shot-wise low-level features. The features employed in the PicSOM system, including still-image keyframe features, genuine visual video features and audio

features have been discussed in Section 3.2.

Concept detection is the final task in the preparation phase. Each one of the concepts defined in the training annotations is directly used as a category that is detected in a straightforward manner by employing the framework outlined in Chapter 3. After this, a post-processing step is employed in order to exploit the concept correlations of temporally nearby shots. The post-processing combines concept-wise N-gram modelling with clustering of temporal neighbourhoods [196]. In the TRECVID 2009 search experiments, the concept lexicons of the high-level feature extraction tasks of TRECVID 2008 and 2009 have been used, consisting of 30 concepts in total. In addition to the detections of these 30 concepts with the PicSOM category detection system, two external sets of concept detectors were used in the search experiments: the 64 detector set of the MediaMill [172] group and the CU-VIREO374 [80] detector set for 374 concepts. These detectors had been made available to all TRECVID participants.

Search phase

After the preparation phase, the retrieval system is ready to be used in the search phase. In this phase, the system is queried with a textual phrase, combined with image and video examples of the desired query topic. The result of a query is a list of video shots, ranked in the order of decreasing predicted likelihood to match the query. The system operation in the search phase is intended to be sufficiently fast to enable the retrieval needs of a real user to be satisfied while she is waiting, typically in a couple of seconds. The example images and video shots will require pre-processing, feature extraction and classification that cannot be performed during the preparing phase, but will inevitably need to be taken care of while the user is waiting for the output.

The conventional approach to video search has been to rely on textual descriptions, keywords, and other meta-data, but this requires manual annotation and does not usually scale well to large and dynamic video collections. Content-based video retrieval can overcome these obstacles as much of the manual labour can be replaced with automatic methods. A traditional paradigm in content-based retrieval has been query-by-example, where the queries are based on a small number of provided examples. The material of a video collection is ranked based on its similarity to the examples according to the material's low-level features [40,

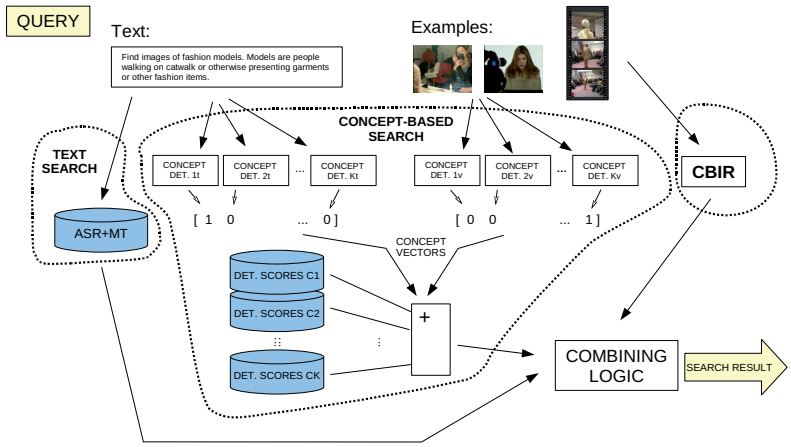


Figure 4.8. General architecture of the PicSOM automatic search subsystem. Concept-based search is supplemented with textual and content-based (CBIR) search. The text is extracted from the video soundtracks using a combination of automatic speech recognition (ASR) and machine translation (MT). From Publication VI.

168, 169].

In recent works, the example-based techniques are commonly combined with separately pre-trained detectors for various semantic concepts (query-by-concepts) [69, 174]. By now visual concept lexicons or ontologies have been shown to be integral parts of effective content-based video retrieval systems. In automatic concept-based video retrieval, the fundamental problem is how to map the user's information need into the space of available concepts in the used concept ontology [133]. The basic approach is to select a small number of concept detectors as active and weight them based on either the performance of the detectors or their estimated suitability for the current query. Another question is how to fuse the output of the concept detectors with the other modalities such as text search and example-based retrieval.

Figure 4.8 schematically shows the architecture of the automatic search subsystem the PicSOM system for TRECVID 2009. There concept-based search is supplemented with the outputs of the example-based CBIR and textual search modules. In the concept-based module, the text query is used for selecting the set of concept detectors that is activated. The tex-

tual query is compared to the words associated with the concepts. The concept “animal”, for instance, is triggered for names of common animals such as “dog”, “cat”, “horse” etc. This triggering is based on regular text expressions and word lists that are generated by expanding the words in concept names with WordNet [51] synonyms.

Also the visual examples can be used for selection of concepts. The examples are of the same modality as the database objects, i.e. videos or keyframe images, and can thus be scored using the concept detectors that were already trained in the preparing phase. A particular concept is activated if the sum of detection scores over all examples exceeds a heuristically determined threshold.

4.3.2 TRECVID 2009 search results

Figure 4.9 shows the performance of the PicSOM system in the TRECVID 2009 search setting in comparison with the other participants of the evaluation [131]. The performance is measured using the MAP metric officially used in the evaluation, i.e. the mean average precision over all the 48 query topics (cf. Section2.3). The leftmost bar represents the PicSOM performance after minor changes in the triggering of concepts made after the official TRECVID evaluation. We see that the performance of the PicSOM system compares very well with the other top systems. This result is strengthened by the fact that the top system (BUPT) exploited manual annotation of the training data according to 2008 search topics and thereby detectors tailored for these very specific concepts. This was not forbidden by the evaluation rules, but was unfortunate as over half of the search topics were re-used in 2009 either in unchanged form or with only small modifications.

4.4 Robot navigation

Visual indoor localisation is a fundamental task that autonomous robots equipped with cameras have to perform [18, 58]. A number of different approaches have been proposed, but arguably the prevailing method is to combine camera-based visual information to some additional input modalities [145, 160], such as laser range sensors, sonar, stereo vision, temporal continuity, odometry, and the floor plan of the environment. In addition

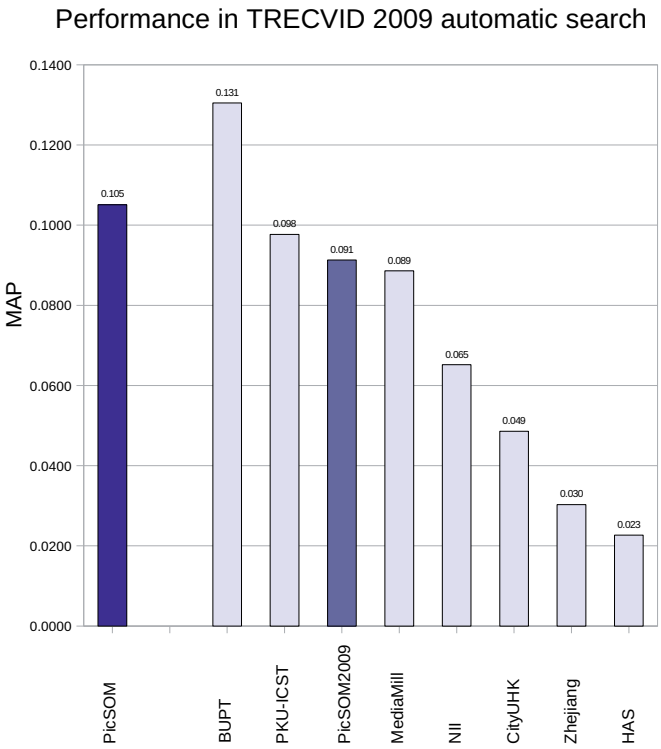


Figure 4.9. The MAP performance in automatic search task of TRECVID 2009. The PicSOM system is compared with the best systems of the groups that submitted their runs of the evaluation. The comparison is made within the most popular class of systems: the systems that use the training data provided by the TRECVID organisers. From Publication VI.

to autonomous robots, need for localisation arises, for example, in many applications of mobile augmented reality [50]. In fact, indoor localisation constitutes also one of the sub-tasks in the development of Aalto University's platform for accessing abstract information in real-world environments through augmented reality displays [3, 4].

This section of the thesis summarises the application of the PicSOM category detection system to a task in mobile robot navigation that has been reported in Publication V. There a supervised formulation of the localisation problem is considered. The system is provided with a sequence of training images that have been annotated with location labels. The task of the system would be to predict the locations in a previously unseen image sequence. A straightforward vision-based approach to the task is to match the query image directly to the images in the training set, using e.g. pairwise matching of interest points [53]. Here an alternative

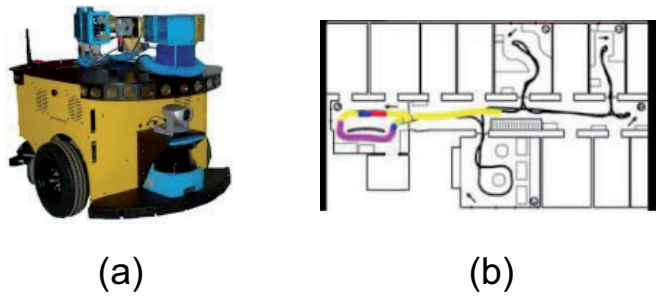


Figure 4.10. a) The mobile robot equipped with stereo cameras, b) the floor plan of the office with some routes taken by the robot marked. The images are from website the of the ImageCLEF campaign (www.imageclef.org).

approach is employed that is based on the category detection framework described in this thesis. The different locations in the training images are interpreted as categories, and the generic PicSOM category detection system is used with a standard collection of visual features. The approach is compared against state-of-the-art localisation methods in the setting of ImageCLEF@ICPR2010 RobotVision benchmark competition[1]. The following sections detail the experimental task setting, the way that the PicSOM system is applied to the task, and the result of the competition.

4.4.1 The ImageCLEF@ICPR2010 RobotVision task

In the ImageCLEF RobotVision competition setting, a mobile robot equipped with stereo cameras moves in an office environment, capturing images at the rate of 5 fps. Given the captured images, the task of the participants of the competition was to determine the room where the robot was in. Figure 4.10 shows the robot and the floor plan of the office.

Two training data sets ("easy" and "hard") and a validation set, all from the COLD-Stockholm database [144], were released in connection with the competition. All these three sets depict a total of nine locations shown with example images in Figure 4.11. The easy training set (4074 frame pairs) differs from the hard one (2267 frame pairs) by showing each location from multiple points and angles, thus giving more training data for each location. Furthermore, the hard set was acquired by driving the robot in the opposite direction from the other two data sets.

[1] http://www.imageclef.org/2010/ICPR/RobotVision/

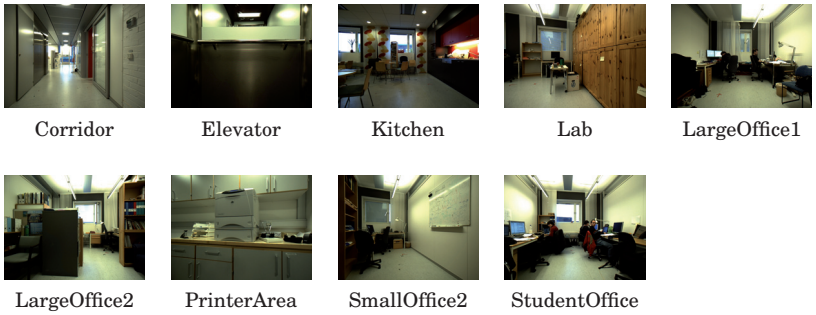


Figure 4.11. The nine known rooms in the training set. Adapted from Publication V.

Both the easy and hard training sets were acquired in daytime, with cloudy weather outside. The validation set was created under similar conditions as the easy set, but in nighttime. In addition to changing illumination, variations in the visual scene were also caused by people or various objects being variably present or absent. The test sequence has 2551 frame pairs. During the capturing of the test sequence the robot moved in the same rooms as in the training sequences, and additionally in four previously unseen rooms.

There were two obligatory tasks in the competition. In both tasks, the location label of each frame pair in the test sequence was to be predicted independent of the other pairs. Given a frame pair, the system was to output either the name of one of the nine previously seen locations or flag the location as "unseen", corresponding to the novel rooms in the test sequence. The system could also refrain from making a classification. The difference between the two tasks was that in one, the predictions should be based on learning of the "easy" training sequence, while the other used the "hard" training sequence. When evaluating the prediction accuracy, a score of $+1.0$ was awarded for each correctly classified frame pair and the score -0.5 for a misprediction. Zero score was given for every frame pair for which the system declined a prediction. The final result in the competition was obtained by summing the scores for both "easy" and "hard" training sequences.

In addition to the obligatory part of the competition where the recognition system was allowed to use only one pair of frames at the time of making the prediction, the competition had also the option to submit results using the whole sequence seen until the frame in question. In that way the temporal continuity of the sequence, for example, could be utilised. However, here only the instantaneous case is considered.

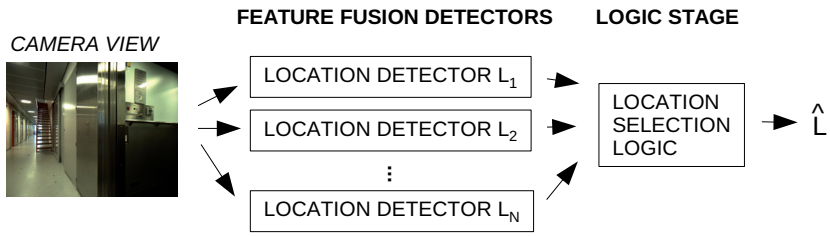


Figure 4.12. General architecture for predicting location $\hat{L}$ based on a camera view. From Publication V.

4.4.2 Applying the PicSOM system to robot navigation

Figure 4.12 illustrates the PicSOM approach to location prediction. Given training images with location labels, first a separate detector for each location L_i is trained using the PicSOM category detection framework. Here each of the nine known locations defines a category. The probabilistic outcomes of the single-location detectors are then used as inputs to a multi-class classification logic module that selects one of the known locations L_i as the final location label $\hat{L}$ for a test image. Alternatively, the logic module can predict that the image is taken in a novel unknown location, or declare the location to be uncertain.

The global visual features employed in the single-location detectors (Table 4.6) are a subset of the features used in the TRECVID 2009 PicSOM system (Section 3.2.4). The BoV features (Table 4.7) use the TRECVID 2009 techniques in slightly different combinations. The detectors use the C-SVC algorithm for supervised learning (Section 3.3.1) and multifold-SFBS for late fusion (Section 3.4.1).

The operation principle of the room selection logic module is simple: in the case of confident and consistent single-room predictions, the room label with maximum confidence is chosen. Unseen location is predicted if all estimates fall below a threshold. Otherwise the system declines to predict the location.

4.4.3 Robot navigation results

A total of eight runs was submitted to the RobotVision contest with the group name “PicSOM TKK”. The best submitted PicSOM result for the “easy” set received a score of 2176.0, which is 85% of the best possible

Table 4.6. The global visual features employed in the single-location detectors for robot navigation. See Section 3.2 for details.

MPEG-7	Color Layout
	Dominant Color
	Edge Histogram
	Scalable Color
non-standard	Average Colour
	Colour Moments
	Texture Neighbourhood
	Edge Histogram
	Edge Co-occurrence
	Edge Fourier

score. This result was based on detectors trained on the left camera data only, and it obtained the overall highest score in the competition in the obligatory task (i.e. the instantaneous case). The same setup achieved the best PicSOM result (1117.0) for the "hard" set as well. The PicSOM result was slightly better than the median of the submitted results that were based on the "hard" training data. The overall best submitted run trained with the "hard" set was 1777.0. The best submitted results for all participating groups are shown in Figure 4.13 for the two training sets, and also the overall score that combines these two.

Some of the submitted PicSOM runs and also some additional runs are summarised in Table 4.8, with "•" denoting that the run was submitted to the competition. The additional runs could be performed as the participants were given access to the class labels for the testing dataset after the competition. The first column in the table specifies how the training data was selected with regard to the two cameras. The word "separate" indicates that separate models were trained for each camera and then averaged, while "both" uses all images to train a single model. The second column states whether fusion of single-feature classifiers or just the single best performing feature (ColorSIFT with dense sampling, soft clustering with a spatial pyramid) is used. One can see that feature fusion is highly beneficial: with a single well-performing feature the results are significantly weaker.

The obtained results indicate that a general-purpose algorithm for vi-

Table 4.7. The BoV features employed in the single-location detectors for robot navigation. See Section 3.2 for details.

Feature	sampling	histograms	spatial partitioning
Color SIFT	dense	soft histograms	spatial pyramid
Color SIFT	interest point	soft histograms	spatial pyramid
Color SIFT	dense	hard histograms	global
Color SIFT	interest point	hard histograms	global
SIFT	dense	soft histograms	spatial pyramid
SIFT	interest point	soft histograms	spatial pyramid
SIFT	dense	hard histograms	global
SIFT	interest point	hard histograms	global

Table 4.8. RobotVision recognition scores. From Publication V.

cameras	features	easy	hard	total
• left only	fusion	2176.0	1117.0	3293.0
right only	fusion	2210.5	1072.0	3282.5
separate	fusion	2207.5	1057.0	3264.5
• both	fusion	2065.0	665.5	2730.5
• both	single	964.0	554.5	1518.5

sual category recognition can perform well in indoor location recognition, given that enough training data is available. However, with limited training data the performance of the presented method is less competitive. There seem to be some implementation issues in the PicSOM system as the after-the-competition analysis of the competing methods that were successful in the hard task shows that even simpler appearance-only based approaches produced good performance. Overlearning is one plausible explanation. With the larger training set, just memorising all the camera views that appear in the training material might be a viable strategy, whereas the smaller training set calls for generalising between views.

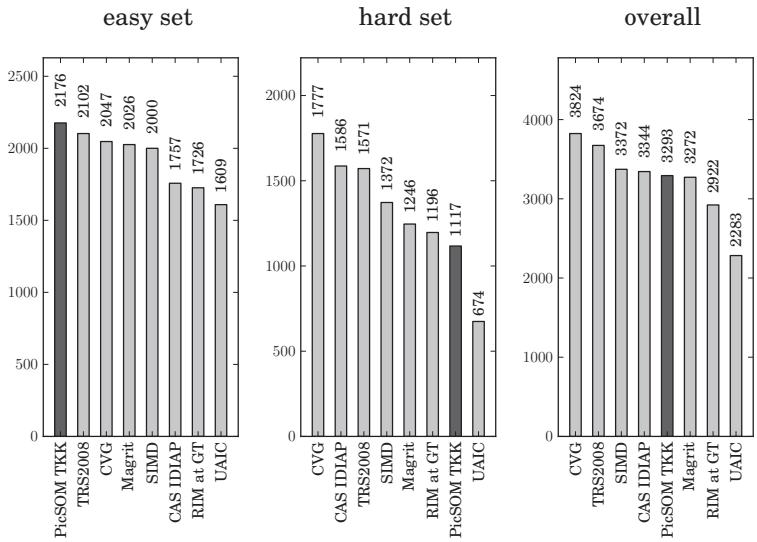


Figure 4.13. Scores of participating groups for the easy and hard sets, and the overall scores. From Publication V.

5 Summary and conclusions

Giving computers the ability to automatically recognise and analyse the content of visual material—images and videos—still remains one of the open questions in computer science. In spite of research efforts extending over several decades, a completely satisfactory solution is not even foreseeable in the near future. In the world with increasing amounts of visual material produced continuously, automatic visual analysis methods would be highly desirable, however.

In this thesis, the general visual analysis problem has been approached by partitioning the problem into small binary category detection problems. A practical architecture for perfoming category detection has emerged over the years from the experiments of researchers in the field. This architecture, consisting of fusion of multiple feature-wise detectors, has been described in this text and the techniques for implementing its components have been discussed. The validity of the approach has been experimentally demonstrated in diverse applications of visual analysis.

The experiments have enabled the comparison of different component technologies within the discussed category detection architecture. Here one has to remember that modern category detection systems are complex multi-component systems. It may be difficult to isolate the effects of distinct interacting system components by studying the published results of various category detection benchmark evaluations. Therefore, results of a single experiment should not be taken as a convincing proof of the superiority of a given technique. Instead, it may be wise to look at the body of experiments in a more probabilistic way, each experiment either strengthening or weakening the belief about the relative performances of the techniques used in category detection.

The publications of the thesis include a wide variety of experiments. Combined with an analysis of a large set of experimental results of others, a conclusion has been drawn that feature fusion is an essential component in modern category detection systems. The prevailing supervised learning

technology is based on kernel methods, usually support vector machines. Other learning methods have not yet essentially outperformed SVMs, at least what comes to accuracy, although some promising results have been demonstrated.

The bag of visual words (BoV) features form the backbone of visual feature extraction in the current category detection systems. At the moment, the BoV techniques do not seem to have been perfected yet, as indicated by experiments in Publications VII–XI. Even though the experiments have not been able to exhaustively and systematically address all the aspects of BoV systems, notable gains in performance can still be observed. In particular, combining information on several levels of granularity—both the granularity of visual vocabulary and granularity of spatial subdivisioning of the image area—has been demonstrated to be a useful approach. Also the issues of selecting the metric in the descriptor space and the spatial matching strategy of sub-images have shown some potential for performance improvement. Since these alterations of the BoV techniques have turned out to be useful, it is reasonable to believe that also other equally useful alterations could be found with further research.

The usefulness of a supervised category detection system is determined largely by the availability and quality of labelled training examples. It is plausible to argue that any visual category could be—in principle—reliably detected, given a large and diverse enough set of training examples. One could even argue that, given sufficiently large amounts of training data, the learning problem might become much easier, if only one could cope with the practical problem of huge data volumes. Ability to exploit large volumes of training data might open a whole new set of application areas where automatic category detection would become practically useful. For example, the large number of images and videos distributed around the Internet could provide such huge training sets, if only the data could be somehow collected and exploited in system training.

However, the above arguments are just conjectures without a sound experimental or theoretical backing. One could equally well argue that a well-functioning general purpose vision system can not be synthesised by merely learning from a set of example images, even from a very large set. A more profound model of how the world works would be needed in the background. However, an argument speaking for the possibility of learning very useful and sophisticated models from examples alone is the

following: the information flow to human brain from the world also goes through the bottleneck of eyes (and ears and other sense organs). Also for humans the learning is example-based.

Experimental demonstrations would be a way to settle the dispute when the arguments in either direction are not ultimately compelling. Until recently, attempts of such demonstrations have not been even thinkable because of lack of large data sets, and more importantly, lack of computational resources. The ever-increasing power of computers combined with the explosion-like growth of the Internet and its collections of visual information provides some hope, however. Still, the demands set on the computational power by category detection systems in this scenario would be taxing. The matter is made very challenging by the huge number of categories required to cover the content types of generic images in a decent semantic resolution. Current category detection systems are not yet up to such challenges. They are able to deal with relatively limited data volumes and a few hundred categories at most. Addressing these issues, alongside with improving the category detection accuracy with a given set of training images, seems to provide both enough challenges and viable research directions in which to proceed in the future.

Bibliography

[1] K. Aas and L. Eikvil. Text categorisation: a survey. Technical Report 941, Norwegian Computing Center, 1999.

[2] G. Agin. *Representation and description of curved objects*. PhD thesis, Stanford University, 1972.

[3] A. Ajanki, M. Billinghurst, H. Gamper, T. Järvenpää, M. Kandemir, S. Kaski, M. Koskela, M. Kurimo, J. Laaksonen, K. Puolamäki, T. Ruokolainen, and T. Tossavainen. An augmented reality interface to contextual information. *Virtual Reality*, 15(2-3):161–173, 2011.

[4] A. Ajanki, M. Billinghurst, T. Järvenpää, M. Kandemir, S. Kaski, M. Koskela, M. Kurimo, J. Laaksonen, K. Puolamäki, T. Ruokolainen, and T. Tossavainen. Contextual information access with augmented reality. In *Proceedings of 2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 95–100, Kittilä, Finland, August-September 2010.

[5] M. Aksela. *Adaptive combinations of classifiers with application to on-line handwritten character regognition*. PhD thesis, Laboratory of Computer and Information Science, Helsinki University of Technology, 2007.

[6] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

[7] C.-N. E. Anagnostopoulos, I. E. Anagnostopoulos, I. D. Psoroulas, V. Loumos, and E. Kayafas. License plate recognition from still images and video sequences: a survey. *IEEE Transactions on Intelligent Transportation Systems*, 9(3):377–391, 2008.

[8] S. Andrews, I. Tsochantaridis, and T. Hoffman. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, pages 561–568. MIT Press, 2003.

[9] K. Arbter. Affine-invariant Fourier descriptors. In J. C. Simon, editor, *From Pixels to Features*, pages 153–164. Elsevier Science Publishers B.V.(North-Holland), 1989.

[10] S. Ayache and G. Quénot. Video corpus annotation using active learning. In *Proceedings of 30th European Conference on Information Retrieval (ECIR'08)*, pages 187–198, Glasgow, UK, March-April 2008.

[11] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research,*

Special Issue on Machine Learning Methods for Text and Images, 3:1107–1135, February 2003.

[12] T. Barnard and M. A. Fischler. Computational stereo. *ACM Computing Surveys*, 14(4):553–572, 1982.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. SURF: speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[14] I. Biederman. Aspects and extension of a theory of human image understanding. In *Computational Processes in Human Vision: An Interdisciplinary Perspective*. Ablex Publishing Corporation, Norwood, New Jersey, 1988.

[15] I. Biederman. *An invitation to cognitive science, Vol 2: Visual cognition*, chapter Visual Object Recognition, pages 121–165. MIT Press, 1995.

[16] T. Binford. Visual perception by computer. In *Proceedings of the IEEE Conference on Systems and Control*, 1971.

[17] M. Blachnik and J. Laaksonen. Image classification by histogram features created with learning vector quantization. In *Proceedings of International Conference on Artificial Neural Networks (ICANN'08)*, pages 827–836, Sept. 2008.

[18] J. Borenstein, H. R. Everett, and L. Fen. *Navigating Mobile Robots: Sensors and Techniques*. A. K. Peters, Ltd., 1996.

[19] A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):55–73, 1990.

[20] J. M. Brady, J. Ponce, A. Yuille, and H. Asada. Describing surfaces. In *International Symposium on Robotics Research*, pages 5–16. MIT Press, 1985.

[21] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[22] M. Burl, M. Weber, and P. Perona. A probabilistic approach to to object recognition using local photometry and global geometry. In *Proceedings of European Conference on Computer Vision*, pages 628–641, 1998.

[23] D. Cai, X. He, and J. Han. Efficient kernel discriminant analysis with spectral regression. In *Proceeding of International Conference on Data Mining*, pages 427–432, 2007.

[24] M. Campbell, A. Haubold, M. Liu, A. P. Natsev, J. R. Smith, J. Tešić, R. Yan, , and J. Yang. IBM Research TRECVID-2008 video retrieval system. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2007. TRECVID.

[25] J. F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[26] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proceedings of the Eight European Conference on Computer Vision*, pages 350–362, Prague, May 2004.

[27] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as supervised learning problem. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 163–168, 2005.

[28] E. Celebi and A. Alpkocak. Combining textual and visual clusters for semantic image retrieval and auto-annotation. In *Proceedings of 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (EWIMT 2005)*, pages 219–225, London, UK, November 2005.

[29] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[30] M.-Y. Chen and A. Hauptmann. MoSIFT: recognizing human actions in surveillance videos. Technical Report CMU-CS-09-161, Carnegie Mellon University, 2009.

[31] Y. Chen and J. Z. Zwang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

[32] CIE. Supplement No. 2 to CIE publication No. 15 Colorimetry (E-1.3.1) 1971: Official recommendations on uniform color spaces, color-difference equations, and metric color terms, 1976.

[33] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[34] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision at European Conference on Computer Vision*, pages 1–22, 2004.

[35] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Procedings of Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[36] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In A. Waibel and K. Lee, editors, *Readings in speech recognition*, pages 65–74. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[37] G. V. de Wouwer, P. Scheunders, and D. V. Dyck. Statistical texture characterization from discrete wavelet representations. *IEEE Transactions on Image Processing*, 8:592–598, 1999.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B-39(1):1–38, 1977.

[39] P. A. Devijver and J. Kittler. *Pattern recognition: a statistical approach*. Prentice-Hall, London, 1982.

[40] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor. Applications of video-content analysis and retrieval. *IEEE MultiMedia*, 9(3):42–55, July-September 2002.

[41] Y. Dong, X. Zhao, C. Dong, J. Liu, L. Lu, Z. Wei, G. Xiao, S. Lian, R. Wang, and K. Tao. The France Telecom Orange Labs (Beijing) video high-level feature extraction systems—TrecVid 2009 notebook paper. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2009. TRECVID.

[42] C. Dorai and V. Kobla. Perceived visual motion descriptors from MPEG-2 for content-based HDTV annotation and retrieval. In *Proceedings of IEEE 3rd Workshop on Multimedia Signal Processing*, pages 147–152, Copenhagen, Denmark, September 1999.

[43] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of Seventh European Conference on Computer Vision*, pages IV:97–112, 2002.

[44] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[45] 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*, 88:303–338, 2010.

[46] F. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving 'bag-of-keypoints image categorisation'. Technical report, University of Southampton, 2005.

[47] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[48] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of the 9th International Conference on Computer Vision*, pages 1134–1141, 2003.

[49] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005.

[50] S. Feiner, B. MacIntyre, T. Höllerer, and A. Webster. A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. *Personal and Ubiquitous Computing*, 1(4):208–217, December 1997.

[51] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[52] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1002–1009, 2004.

[53] Y. Feng, M. Halvey, and J. M. Jose. University of Glasgow at ImageCLEF 2009 Robot Vision task. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, September-October 2009.

[54] R. Fergus. *Visual Object Category Recognition*. PhD thesis, University of Oxford, Oxford, UK, December 2005.

[55] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67–92, 1973.

[56] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[57] M. Flickner, H. Sawhney, W. Niblack, et al. Query by image and video content: The QBIC system. *IEEE Computer*, 28:23–31, September 1995.

[58] D. Fox, W. Burgard, and S. Thrun. Markov localization for mobile robots in dynamic environments. *Journal of Artificial Intelligence Research*, 11:391–427, 1999.

[59] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, September 1991.

[60] A. Genkin, D. D. Lewis, and D. Madigan. BBR: Bayesian logistic regression software, 2005. Software available at http://www.stat.rutgers.edu/ madigan/BBR/.

[61] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, August 2007.

[62] J.-M. Geusebroek. Compact object descriptors from local colour invariant histograms. In *Proceedings of British Machine Vision Conference*, pages 1029–1038, Edinburgh, UK, 2006.

[63] J.-M. Geusebroek and A. W. M. Smeulders. A six-stimulus theory for stochastic texture. *International Journal of Computer Vision*, 62(1/2):2005, 2005.

[64] J.-M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.

[65] M. M. Gorkani and R. W. Picard. Texture orientation for sorting photos "at a glance". In *12th IAPR International Conference on Pattern Recognition*, pages 459–464, 1994.

[66] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760, April 2007.

[67] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1957–1182, 2003.

[68] E. Hatcher and O. Gospodnetic. *Lucene in Action*. Manning Publications Co., 2004.

[69] A. G. Hauptmann, M. G. Christel, and R. Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, April 2008.

[70] D. Hearn and M. P. Baker. *Computer graphics*. Prentice Hall, 2 edition, 1994.

[71] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Spatial color indexing and applications. *International Journal of Computer Vision*, 35(3):245–268, 1999.

[72] T. S. Huang, S. Mehrotra, and K. Ramchandran. Multimedia analysis and retrieval system (MARS) project. In *Proceedings of 33rd Annual Clinic on Library Application on Data Processing - Digital Image Access and Retrieval*, Urbana-Champaign, IL, USA, March 1996.

[73] D. P. Huttenlocher and S. Ullman. Object recognition using alignment. In *Proceedings of the 1st International Conference on Computer Vision*, pages 102–111, London, UK, 1987.

[74] J. Iivarinen, R. Rautkorpi, J. Pakkanen, and J. Rauhamaa. Content-based retrieval of surface defect images with PicSOM. *International Journal of Fuzzy Systems*, 6(3):160–166, 2004.

[75] ISO/IEC. Information technology - Multimedia content description interface - Part 3: Visual, 2002. 15938-3:2002(E).

[76] A. K. Jain and D. Zongker. Feature-selection: evaluation, application and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[77] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the European Conference on Computer Vision*, volume 1, pages 304–317, 2008.

[78] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 119–126, Toronto, Canada, July-August 2003.

[79] J. Jeon and R. Manmatha. Using maximum entropy for automatic image annotation. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, pages 24–32, 2004.

[80] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for large scale semantic concept detection. Technical Report ADVENT #223-2008-1, Columbia University, August 2008.

[81] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia*, 12(1):42–53, 2010.

[82] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pages 604–610, 2005.

[83] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[84] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. WEBSOM—self-organizing maps of document collections. *Neurocomputing*, 21(1-3):101–117, 1998.

[85] A. Khotanzad and Y. H. Hong. Invariant image recognition by Zernike moments. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(5):489–497, 1990.

[86] S. Kim, I. S. Kweon, and C.-W. Lee. Visual categorization robust to large intra-class variations using entropy-guided codebook. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 3793–3798, Rome, Italy, 2007.

[87] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):110–115, 1998.

[88] J. Koenderink and A. V. Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31(2/3):158–168, 1999.

[89] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer-Verlag, Berlin, third edition, 2001.

[90] T. Kohonen, E. Oja, A. Kortekangas, K. Mäkisara, and P. Lehtiö. Demonstration of pattern processing properties of the optimal associative mappings. In *Proceedings of the International Conference on Cybernetics and Society*, pages 581–585, Washington D.C., USA, 1977.

[91] P. Koikkalainen and E. Oja. Self-organizing hierarchical feature maps. In *Proceedings of International Joint Conference on Neural Networks*, volume II, pages 279–284, San Diego, CA, USA, 1990.

[92] M. Koskela and J. Laaksonen. Semantic concept detection from news videos with self-organizing maps. In I. Maglogiannis, K. Karpouzis, and M. Bramer, editors, *Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations*, pages 591–599, Athens, Greece, June 2006. IFIP, Springer.

[93] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:19–86, 1951.

[94] J. Laaksonen, M. Koskela, and E. Oja. PicSOM - a framework for content-based image database retrieval using self-organizing maps. In *11th Scandinavian Conference on Image Analysis (SCIA99)*, pages 151–156, Kangerlussuaq, Greenland, June 1999.

[95] J. Laaksonen, M. Koskela, and E. Oja. PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. *IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing*, 13(4):841–853, July 2002.

[96] Y. Lamdan, J. T. Schwartz, and H. J. Wolfson. Object recognition by affine invariant matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 335–344, 1988.

[97] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorisation. In *Proceedings of the British Machine Vision Conference*, pages 959–968, 2006.

[98] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceeding of the Seventeenth Annual Conference on Neural Information Processing Systems*, volume 16, pages 553–560, 2003.

[99] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006.

[100] S. Lazebnik, C. Schmid, and J. Ponce. *Object Categorization: Computer and Human Vision Perspectives*, chapter Spatial Pyramid Matching. Cambridge University Press, 2009.

[101] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[102] T. S. Lee. Image representation using 2D Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):1–13, 1996.

[103] J. Li, W. Wu, T. Wang, and Y. Zhang. One step beyond histograms: Markov stationary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[104] Y. Liang, X. Liu, Z. Wang, J. L. B. Cao, Z. Cao, Z. Dai, Z. Guo, W. Li, L. Luo, Z. Meng, Y. Qin, S. Qiu, A. Tian, D. Wang, Q. Wang, C. Zhu, X. Hu, J. Yuan, P. Yuan, B. Zhang, S. Chen, J. Li, T. Wang, and Y. Zhang. THU and ICRC at TRECVID 2008. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2008. TRECVID.

[105] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, January 1980.

[106] M. Livingstone and D. Hubel. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240:740–749, 1988.

[107] D. G. Lowe. The viewpoint consistency constraint. *International Journal of Computer Vision*, 1(1):52, 72 1987.

[108] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[109] B. S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons Ltd., 2002.

[110] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[111] R. Maree, P. Geurts, J. Piater, and L. Wehenkel. Random subwindows for robust image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 34–40, 2005.

[112] M. Marszałek and C. Schmid. Accurate object localization with shape masks. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 1–8, June 2007.

[113] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition, October 2007. Visual Recognition Challenge workshop, in conjunction with ICCV.

[114] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[115] T. Mei, X.-S. Hua, W. Lai, L. Yang, Z.-J. Zha, Y. Liu, , Z. Gu, G.-J. Qi, M. Wang, J. Tang, X. Yuan, Z. Lu, and J. Liu. MSRA-USTC-SJTU at TRECVID 2007: high-level feature extraction and search. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2007. TRECVID.

[116] T. Mei, Z.-J. Zha, Y. Liu, M. Wang, G.-J. Qi, X. Tian, J. Wang, L. Yang, and X.-S. Hua. MSRA at TRECVID 2008: high-level feature extraction and automatic search. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2008. TRECVID.

[117] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval (CIVR 2004)*, pages 42–50, 2004.

[118] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *Proceedings of IEEE International Conference on Computer Vision*, volume 2, pages 1972–1799, 2005.

[119] K. Mikolajczyk and C. Schmid. Scale and affine point invariant interest point detectors. *International Journal of Computer Vision*, 60(1):68–86, 2004.

[120] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, October 2005.

[121] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

[122] M. Miyahara and Y. Yoshida. Mathematical transform of (R,G,B) color data to Munsell (H,V,C) color data. In *SPIE Visual Communications and Image Processing*, volume 1001, pages 650–657, 1988.

[123] M. Molinier, J. Laaksonen, and T. Häme. Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):861–874, April 2007.

[124] M. Molinier, V. Viitaniemi, M. Koskela, J. Laaksonen, Y. Rauste, A. Lönnqvist, and T. Häme. Improving content-based target and change detection in Alos Palsar images with efficient feature selection. In *Proceedings of ESA-EUSC 2008: Image Information Mining*. ESA, March 2008.

[125] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 275–278, Berkeley, CA, 2003.

[126] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management*. ACM Press, 1999.

[127] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about Corel – evaluation in image retrieval. In *Proceedings of The Challenge of Image and Video Retrieval (CIVR 2002)*, pages 38–49, London, UK, July 2002.

[128] H. Murase and S. K. Nayar. Visual learning and recognition of 3D objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.

[129] M. Naphade, J. R. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[130] M. R. Naphade and T. S. Huang. Extracting semantics from audiovisual content: The final frontier in multimedia retrieval. *IEEE Transactions on Neural Networks*, 13(4):793–810, July 2002.

[131] National Institute of Standards and Technology (NIST). *Notebook papers of TRECVID 2009 Workshop*, Gaithersburg, Maryland, USA, 2009. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[132] A. Natsev, W. Jiang, M. Merler, J. R. Smith, J. Tešić, L. Xie, and R. Yan. IBM Research TRECVID-2008 video retrieval system. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2008. TRECVID.

[133] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of ACM Multimedia (ACM MM'07)*, pages 991–1000, Augsburg, Germany, September 2007.

[134] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *Proceedings of ECCV06*, volume 3954 of *LNCS*, pages 490–503. Springer Verlag, 2006.

[135] P. Ogilvie and J. Callan. Experiments using the Lemur toolkit. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001)*. National Institute of Standards and Technology, special publication 500-250, 2002.

[136] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[137] K. Peng, X. Cheng, D. Zhou, and Y. Liu. 3D reconstruction based on SIFT and Harris feature points. In *Proceedings of IEEE International Conference on Robotics and Biomimetics*, pages 960–964, 2009.

[138] Y. Peng, Z. Yang, L. Cao, J. Yi, N. Wan, Y. Feng, X. Zhai, E. Shi, and H. Li. PKU-ICST at TRECVID 2009: high level feature extraction and search. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2009. TRECVID.

[139] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. In *Storage and Retrieval for Image and Video Databases II*, volume 2185 of *Proceedings of SPIE*, pages 34–47, San Jose, CA, USA, 1994.

[140] F. Perronnin. Universal and adapted vocabularies for generic visual categorisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1243–1256, 2008.

[141] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2004.

[142] K. Petridis, S. Bloehdorn, C. Saathoff, N. Simou, S. Dasiopoulou, V. Tzouvaras, S. Handschuh, Y. Avrithis, I. Kompatsiaris, and S. Staab. Knowledge representation and semantic annotation of multimedia content. *IEE Proceedings on Vision Image and Signal Processing*, 153(3):255–262, 2006.

[143] J. Philbin, M. Marin-Jimenez, S. Srinivasan, A. Zisserman, M. Jain, S. Vempati, P. Sankar, and C. V. Jawahar. Oxford/IIIT TRECVID 2008 notebook paper. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2008. TRECVID.

[144] A. Pronobis and B. Caputo. COLD: COsy Localization Database. *The International Journal of Robotics Research (IJRR)*, 28(5), May 2009.

[145] A. Pronobis, O. Martínez Mozos, and B. Caputo. SVM-based discriminative accumulation scheme for place recognition. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA08)*, Pasadena, CA, USA, May 2008.

[146] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, November 1994.

[147] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proceedings of ACM Multimedia*, pages 17–26, 2007.

[148] X. Qi and Y. Han. Incorporating multiple SVMs for automatic image annotation. *Pattern Recognition*, 40:728–741, 2007.

[149] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. More efficiency in multiple kernel learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 775–782, 2007.

[150] I. Rigoutsos and R. Hummel. A Bayesian approach to model matching with geometric hashing. *Computer Vision and Image Understanding*, 62:11–26, 1995.

[151] C. Rothwell, D. Forsyth, A. Zisserman, and J. Mundy. Extracting projective structure from single perspective views of 3D point sets. In *Proceedings of*

4th International Conference on Computer Vision, pages 573–582, Berlin, Germany, 1993.

[152] C. Rothwell, A. Zisserman, D. Forsyth, and J. Mundy. Planar object recognition using projective shape presentation. *Interational Journal of Computer Vision*, 16(1):57–99, 1995.

[153] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 59–66, India, January 1998. Code available on-line at `http://www.cs.duke.edu/~tomasi/emd.htm`.

[154] M. S. Sarfraz and O. Hellwich. Head pose estimation in face recognition across pose scenarios. In *Proceedings of VISAPP2008*, pages 235–242, Madeira, Portugal, January 2008.

[155] T. Sato, Y. Nishiumi, M. Susuki, T. Nakagawa, and N. Yokoya. Camera position and posture estimation from still image using feature landmark database. In *Proceedings of SICE annual conference*, pages 1514–1519, 2008.

[156] R. J. Schalkoff. *Pattern Recognition: Statistical, Structural and Neural Approaches*. John Wiley & Sons, Ltd., 1992.

[157] R. E. Schapire. The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear Estimation and Classification*, pages 149–171. Springer, 2003.

[158] B. Schiele and J. L. Crowley. Object recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[159] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, 1997.

[160] S. Se, D. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *The International Journal of Robotics Research*, 21(8):735–758, 2002.

[161] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[162] T. Sirithinaphong and K. Chamnongthai. The recognition of car license plate for automatic parking system. In *Proceedings of 5th International Symposium on Signal Processing and its Applications*, pages 455–457, 1998.

[163] M. Sjöberg. Content-based retrieval of hierarchical objects with PicSOM. Master's thesis, Laboratory of Computer and Information Science, Helsinki University of Technology, 2006.

[164] M. Sjöberg, M. Koskela, M. Chechev, and J. Laaksonen. PicSOM experiments in TRECVID 2010. In *Proceedings of the TRECVID 2010 Workshop*, Gaithersburg, MD, USA, November 2010. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[165] M. Sjöberg, H. Muurinen, J. Laaksonen, and M. Koskela. PicSOM experiments in TRECVID 2006. In *Proceedings of the TRECVID 2006 Workshop*, Gaithersburg, MD, USA, November 2006. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[166] M. Sjöberg, V. Viitaniemi, M. Koskela, and J. Laaksonen. PicSOM experiments in TRECVID 2009. In *Proceedings of the TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009. Available online at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[167] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[168] A. F. Smeaton, P. Wilkins, M. Worring, O. de Rooij, T.-S. Chua, and H. Lua. Content-based video retrieval: Three example systems from TRECVid. *International Journal of Imaging Systems and Technology*, 18(2-3):195–201, 2008.

[169] S. W. Smoliar and H. Zhang. Content-based video indexing and retrieval. *IEEE MultiMedia*, 1(2):62–72, 1994.

[170] C. Snoek, I. Everts, J. van Gemert, J. Geusebroek, B. Huurnink, D. Koelma, M. van Liempt, O. de Rooij, K. van de Sande, A. Smeulders, J. Uijlings, and M. Worring. The MediaMill TRECVID 2007 semantic video search engine. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2007. TRECVID.

[171] C. Snoek, K. van de Sande, O. de Rooij, B. Huurnink, E. Gavves, D. Odijk, M. de Rijke, T. Gevers, M. Worring, and A. Smeulders. The MediaMill TRECVID 2010 semantic video search engine. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2010. TRECVID.

[172] C. Snoek, K. van de Sande, O. de Rooij, B. Huurnink, J. van Gemertand J.R.R. Uijlings, J. He, X. Li, I. Everts, V. Nedovi, M. van Liempt, R. van Balen, F. Yan, M. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J. Geusebroek, T. Gevers, M. Worring, A. Smeulders, and D. Koelma. The MediaMill TRECVID 2008 semantic video search engine. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2008. TRECVID.

[173] C. G. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of ACM Multimedia*, pages 421–430, Santa Barbara, USA, 2006.

[174] C. G. M. Snoek and M. Worring. Are concept detector lexicons effective for video search? In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME 2007)*, pages 1966–1969, Beijing, China, July 2007.

[175] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[176] M. Stricker and M. Orengo. Similarity of color images. In *Storage and Retrieval for Image and Video Databases III (SPIE)*, volume 2420 of *SPIE Proceedings Series*, pages 381–392, San Jose, CA, USA, February 1995.

[177] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[178] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Database*, pages 42–51, Bombay, India, January 1998.

[179] J. Tang, X.-S. Hua, T. Mei, G.-J. Qi, and X. Wu. Video annotation based on temporally consistent Gaussian random field. *Electronics Letters*, 43(8), 2007.

[180] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorisation. In *Proceedings of ACM International Conference on Image and Video Retrieval*, pages 249–258, Niagara Falls, Canada, 2008.

[181] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.

[182] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context based vision system for place and object recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 273–280, 2003.

[183] M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, Maui, HI USA, June 1991.

[184] S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.

[185] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[186] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *Proceedings of the European Conference on Computer Vision*, pages 696–709, 2008.

[187] J. C. van Gemert, C. J. Veenman, and J. M. Geusebroek. Episode-constrained cross-validation in video concept retrieval. *IEEE Transactions on Multimedia*, 11(4):780–785, 2009.

[188] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[189] C. V. van Rijsbergen. *Information Retrieval*. Butterworth, 2nd edition, 1979.

[190] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[191] V. N. Vapnik and A. Y. Chervonenkis. *Theory of Pattern Recognition (in Russian)*. Nauka, Moscow, 1974.

[192] V. Viitaniemi and J. Laaksonen. Content-based browsing of mail-order catalogue with PicSOM system. In *Proceedings of the Ninth International Conference on Distributed Multimedia Systems / The 2003 Conference on Visual Information Systems (VIS'2003)*, pages 381–386, Miami, FL, USA, September 2003.

[193] V. Viitaniemi and J. Laaksonen. Keyword-detection approach to automatic image annotation. In *Proceedings of 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (EWIMT 2005)*, pages 15–22, London, UK, November 2005.

[194] V. Viitaniemi and J. Laaksonen. Use of image regions in context-adaptive image classification. In Y. Avrithis, S. Staab, and N. O'Connor, editors, *Proceedings of the 1st International Conference on Semantic and Digital Media Technologies (SAMT 2006)*, Lecture Notes in Computer Science, pages 169–183, Athens, Greece, December 2006. Springer.

[195] V. Viitaniemi and J. Laaksonen. Empirical investigations on benchmark tasks for automatic image annotation. In *Proceedings of the 9th International Conference on Visual Information Systems (VISUAL 2007)*, volume 4781 of *Lecture Notes in Computer Science*, pages 93–104. Springer, June 2007.

[196] V. Viitaniemi, M. Sjöberg, M. Koskela, and J. Laaksonen. Exploiting temporal and inter-concept co-occurrence structure to detect high-level features in broadcast videos. In *Proceedings of 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)*, pages 12–15, Klagenfurt, Austria, May 2008.

[197] V. Viitaniemi, M. Sjöberg, M. Koskela, and J. Laaksonen. Automatic video search using semantic concepts. In *Proceedings of 8th European Conference on Interactive TV and Video (EuroITV 2010)*, Tampere, Finland, June 2010.

[198] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages I:511–518, 2001.

[199] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang. Video Diver: generic video indexing with diverse features. In *Proceedings of the MIR workshop at ACM Multimedia*, pages 61–70, 2007.

[200] F. Wang and B. Merialdo. Eurocom at TRECVID 2009 high-level feature extraction. In *TRECVID Online Proceedings*, Gaithersburg, USA, November 2009. TRECVID.

[201] F. Wang and C. Zhang. Label propagation through linear neighborhoods. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 985–992, 2006.

[202] J. Wang, Y. Zhao, and X. S. Hua. Transductive multi-label learning for video concept detection. In *Proceedings of ACM Multimedia*, pages 298–304, Vancouver, Canada, 2008.

[203] J. Z. Wang, J. Liu, and G. Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, September 2001.

[204] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai. Optimizing multi-graph learning: towards a unified video annotation. In *Proceedings of ACM Multimedia*, pages 862–871, 2007.

[205] M. Weber. *Unsupervised learning of models for object recognition*. PhD thesis, California Institute of Technology, Pasadena, CA, USA, 2000.

[206] S. A. J. Winder and M. Brown. Learning local image descriptors. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[207] J. Winn, A. Criminisi, and T. Minka. Object categoristation by learned universal visual dictionary. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pages 1800–1807, 2005.

[208] F. Yan, J. Kittler, K. Mikolajczyk, and A. Tahir. Non-sparse multiple kernel learning for Fisher discriminant analysis. In *Proceedings of IEEE International Conference on Data Mining*, pages 1064–1069, 2009.

[209] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the ACM Multimedia Information Retrieval Workshop (MIR 2007) at ACM Multimedia*, pages 197–206, 2007.

[210] Z. Yang and J. Laaksonen. Approximated classification in interactive facial image retrieval. In *Proceedings of 14th Scandinavian Conference on Image Analysis (SCIA 2005)*, pages 770–779, Joensuu, Finland, June 2005.

[211] A. Yavlinsky, E. Schofield, and S. Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *Proceedings of International Conference on Image and Video Retrieval*, pages 507–517, Singapore, July 2005.

[212] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of 15th International Conference on Information and Knowledge Management (CIKM 2006)*, Arlington, VA, USA, November 2006.

[213] N. H. C. Yung, K. H. Au, and A. H. S. Lai. Recognition of vehicle registration mark on moving vehicles in an outdoor environment. In *Proceedings of IEEE International Conference on Intelligent Transportation Systems*, pages 418–422, 1999.

[214] T. M. Z. Gu, X.-S. Hua, J. Tang, and X. Wu. Multi-layer multi-instance kernel for video concept detection. In *Proceedings of ACM Multimedia*, pages 349–352, 2007.

[215] J. Zhang, M. Marszałek, 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.

[216] Q. Zhang and E. Izquierdo. Adaptive salient block-based image retrieval in multi-feature space. *Image Communications*, 22:591–603, July 2007.

[217] X. Zhou, X. Zhuang, S. Yan, S.-F. Chang, M. Hasegawa-Johnson, and T. S. Huang. Sift-bag kernel for video event analysis. In *Proceeding of the 16th ACM international conference on Multimedia*, pages 229–238, October 2008.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

TKK-ICS-D20 Miche, Yoan.
Developing Fast Machine Learning Techniques with Applications to Steganalysis Problems. 2010.

TKK-ICS-D21 Sorjamaa, Antti.
Methodologies for Time Series Prediction and Missing Value Imputation. 2010.

TKK-ICS-D22 Schumacher, André
Distributed Optimization Algorithms for Multihop Wireless Networks. 2010.

Aalto-DD99/2011 Ojala, Markus
Randomization Algorithms for Assessing the Significance of Data Mining Results. 2011.

Aalto-DD111/2011 Dubrovin, Jori
Efficient Symbolic Model Checking of Concurrent Systems. 2011.

Aalto-DD118/2011 Hyvärinen, Antti
Grid Based Propositional Satisfiability Solving. 2011.

Aalto-DD136/2011 Brumley, Billy Bob
Covert Timing Channels, Caching, and Cryptography. 2011.

Aalto-DD11/2012 Vuokko, Niko
Testing the Significance of Patterns with Complex Null Hypotheses. 2012.

Aalto-DD19/2012 Reunanen, Juha
Overfitting in Feature Selection: Pitfalls and Solutions. 2012.

Aalto-DD33/2012 Caldas, José
Graphical Models for Biclustering and Information Retrieval in Gene Expression Data. 2012.

In the modern world, huge collections of digital visual material are accessible to an individual. However, automatically distilling useful information from such collections requires one to somehow answer the grand long-standing question of computer vision: how to make computers understand images. Here a partial solution is considered: the general question about the image contents is decomposed into smaller yes/no questions about whether the images belong to certain visual categories, such as images captured during nighttime or depicting a person. In this thesis the details of the traditional way of implementing visual category detection are extensively discussed and compared in the light of performance in actual category detection experiments. The thesis also demonstrates how a set of diverse visual analysis problems can be addressed using the same visual category detection system as a backbone component by equipping the system with a task-specific front-end.

Aalto-DD 45/2012

ISBN 978-952-60-4585-6
ISBN 978-952-60-4586-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS