

ACTA

UNIVERSITATIS OULUENSIS

*Mika Rautiainen*

CONTENT-BASED SEARCH  
AND BROWSING IN  
SEMANTIC MULTIMEDIA  
RETRIEVAL

FACULTY OF TECHNOLOGY,  
DEPARTMENT OF ELECTRICAL AND INFORMATION ENGINEERING,  
UNIVERSITY OF OULU

C  
TECHNICA





ACTA UNIVERSITATIS OULUENSIS  
C Technica 262

*MIKA RAUTIAINEN*

**CONTENT-BASED SEARCH AND  
BROWSING IN SEMANTIC  
MULTIMEDIA RETRIEVAL**

Academic dissertation to be presented, with the assent of  
the Faculty of Technology of the University of Oulu, for  
public defence in Raahensali (Auditorium L10), Linnanmaa,  
on December 13th, 2006, at 12 noon

OULUN YLIOPISTO, OULU 2006

Copyright © 2006  
Acta Univ. Oul. C 262, 2006

Supervised by  
Professor Tapio Seppänen

Reviewed by  
Professor Samuel Kaski  
Professor Alan Smeaton

ISBN 951-42-8299-X (Paperback)  
ISBN 951-42-8300-7 (PDF) <http://herkules.oulu.fi/isbn9514283007/>  
ISSN 0355-3213 (Printed)  
ISSN 1796-2226 (Online) <http://herkules.oulu.fi/issn03553213/>

Cover design  
Raimo Ahonen

OULU UNIVERSITY PRESS  
OULU 2006

## **Rautiainen, Mika, Content-based search and browsing in semantic multimedia retrieval**

Faculty of Technology, University of Oulu, P.O.Box 4000, FI-90014 University of Oulu, Finland,  
Department of Electrical and Information Engineering, University of Oulu, P.O.Box 4500, FI-90014 University of Oulu, Finland

*Acta Univ. Oul. C 262, 2006*

Oulu, Finland

### ***Abstract***

Growth in storage capacity has led to large digital video repositories and complicated the discovery of specific information without the laborious manual annotation of data. The research focuses on creating a retrieval system that is ultimately independent of manual work. To retrieve relevant content, the semantic gap between the searcher's information need and the content data has to be overcome using content-based technology. Semantic gap constitutes of two distinct elements: the ambiguity of the true information need and the equivocalness of digital video data.

The research problem of this thesis is: what computational content-based models for retrieval increase the effectiveness of the semantic retrieval of digital video? The hypothesis is that semantic search performance can be improved using pattern recognition, data abstraction and clustering techniques jointly with human interaction through manually created queries and visual browsing.

The results of this thesis are composed of: an evaluation of two perceptually oriented colour spaces with details on the applicability of the HSV and CIE Lab spaces for low-level feature extraction; the development and evaluation of low-level visual features in example-based retrieval for image and video databases; the development and evaluation of a generic model for simple and efficient concept detection from video sequences with good detection performance on large video corpuses; the development of combination techniques for multi-modal visual, concept and lexical retrieval; the development of a cluster-temporal browsing model as a data navigation tool and its evaluation in several large and heterogeneous collections containing an assortment of video from educational and historical recordings to contemporary broadcast news, commercials and a multilingual television broadcast.

The methods introduced here have been found to facilitate semantic queries for novice users without laborious manual annotation. Cluster-temporal browsing was found to outperform the conventional approach, which constitutes of sequential queries and relevance feedback, in semantic video retrieval by a statistically significant proportion.

*Keywords:* content-based, interactive browsing, multi-modal search, semantic concept, video retrieval, visual feature



## Acknowledgements

The work reported in this thesis was carried out in the MediaTeam Oulu research group of the Department of Electrical and Information Engineering at the University of Oulu, Finland, and in the Language and Media Processing Laboratory at the University of Maryland, United States. A large part of this work is the result of discourse and collaboration with many people in the image and video systems development team, to whom I wish to express my gratitude. I would like to acknowledge Professor Tapio Seppänen for supervising the thesis and for the professional support I have received from him. I would also like to thank Professor Timo Ojala for teaching me indefatigably the art of science. I would also like to thank all my colleagues in MediaTeam for creating an enthusiastic and encouraging atmosphere to work in.

I am grateful to Doctor David Doermann and Doctor Daniel DeMenthon for supervising me during my stay at the University of Maryland.

For the financial support obtained for this thesis, I would like to thank the Graduate School in Electronics, Telecommunications, and Automation, the Tauno Tönning foundation, the Foundation for the Promotion of Technology (TES) and the Nokia Foundation. Finally, I would like to deeply thank my fiancée Mari for providing invaluable support and encouragement in the domestic front and my parents Hilikka and Kalevi for encouraging me to study in my youth.

Oulu, November 2006

Mika Rautiainen



## List of Original Publications

- I Ojala T, Rautiainen M, Matinmikko E & Aittola M (2001) Semantic image retrieval with HSV correlograms. Proc. 12th Scandinavian Conference on Image Analysis, Bergen, Norway, 621-627.
- II Rautiainen M & Doermann D (2002) Temporal color correlograms for video retrieval. Proc. 16th International Conference on Pattern Recognition, Quebec, Canada, 1: 267-270.
- III Rautiainen M, Ojala T & Kauniskangas H (2001) Detecting perceptual color changes from sequential images for scene surveillance. IEICE Transactions on Information and Systems, Special Issue on Machine Vision Applications, E84-D: 1676-1683.
- IV Rautiainen M, Seppänen T, Penttilä J & Peltola J (2003) Detecting semantic concepts from video using temporal gradients and audio classification. International Conference on Image and Video Retrieval, Urbana, IL, 260-270.
- V Rautiainen M & Seppänen T (2005) Comparison of visual features and fusion techniques in automatic detection of concepts from news video. Proc. IEEE International Conference on Multimedia & Expo, Amsterdam, Netherlands, 932-935.
- VI Rautiainen M, Ojala T & Seppänen T (2004) Analysing the performance of visual, concept and text features in content-based video retrieval. Proc. 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York, NY, 197-205.
- VII Rautiainen M, Ojala T & Seppänen T (2003) Cluster-temporal video browsing with semantic filtering. Proc. 5th International Conference on Advanced Concepts for Intelligent Vision Systems CD-ROM, Ghent, Belgium, 116-123.
- VIII Rautiainen M, Ojala T & Seppänen T (2004) Cluster-temporal browsing of large news video databases. Proc. 2004 IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, 2: 751-754.

- IX Rautiainen M, Ojala T & Seppänen T (2005) Content-based browsing in large news video databases. 5th IASTED International Conference on Visualization, Imaging and Image Processing, Benidorm, Spain, 731-736.
- X Rautiainen M, Seppänen T & Ojala T (2006) Advancing content-based retrieval effectiveness with cluster-temporal browsing in multilingual video databases, Proc. 2006 IEEE International Conference on Multimedia and Expo, Toronto, Canada, 377-380.
- XI Rautiainen M, Seppänen T & Ojala T (2006) On the significance of cluster-temporal browsing for generic video retrieval – a statistical analysis. Proc. ACM Multimedia Conference, Santa Barbara, CA, United States, 125-128.

## List of Symbols and Abbreviations

2-D	Two-dimensional
ABC,BBC,CNN	Television broadcast networks
AP	Average Precision, a retrieval performance measure
CBIR	Content-based image retrieval
CBMR	Content-based multimedia retrieval
CBR	Content-based retrieval
CBVIR	Content-based visual information retrieval
CBVR	Content-based video retrieval
CBAR	Content-based audio retrieval
CIE	Commission Internationale de l'Eclairage, International Commission on Illumination
CIF	Common Intermediate Format
C-SPAN	Cable-Satellite Public Affairs Network
HSV	Hue, saturation, value, a colour model
IxQ, IxB	Identifiers for retrieval system test configurations
IxV, IxT, IxVT	Identifiers for cluster-temporal browser test configurations
IR	Information retrieval
$L^*a^*b^*$ , Lab	Perceptually uniform colour space from CIE
NN <sup>k</sup>	Nearest neighbour network for $k$ features
MAP	Mean Average Precision, a retrieval performance measure
MIR	Multimedia information retrieval
MPEG	Moving Pictures Expert Group
MPEG-7	Multimedia Content Description Interface (ISO 15938)
NIST	National Institute of Standards and Technology
NTSC	National Television Systems Committee, a television broadcast standard
PAL	Phase Alternating Line, a television broadcast standard
QBIC	Query By Image Content, image retrieval system from IBM research
RGB	Red, green, blue, a colour model

Sn	Test user identifier
SOM	Self organising map
SUNET	Swedish University Computer Network
TCC, TGC	Temporal colour correlogram, temporal gradient correlogram
TF-IDF	Term Frequency Inverse Document Frequency
TGn	Test topic group identifier
TREC	Text Retrieval Conference
TRECVID	TREC Video Retrieval Evaluation
VCD	Video Compact Disk
$A$	Weight matrix
$c_i$	Quantised colour $i$
$C$	Set of quantised colours
$\mathbf{d}_j$	Feature vector of weights
$d_j$	$j^{\text{th}}$ text document
$d$	Spatial pixel distance
$D; D$	Computational dissimilarity; set of spatial pixel distances
$D_l(k)$	List of ranked dissimilarities for the example $k$ in a feature space $l$
$d_l(k, n)$	Dissimilarity or rank of a database item $n$ in a feature space $l$
$DB$	Set of documents in the database
Det	Set of detected pixels
$Diff$	A difference image
$Dir; Dir^S$	A discrete directional image; set of $N$ directional images from shot $S$
$F_P$	Set of false positive pixels
$F$	Set of semantic concept features
$F$	Number of concepts in a set
$f$	A semantic concept
GT	Set of ground truth pixels
$Hist(c_i, I)$	Histogram of an image for the colour $c_i$
$H; S; V$	Values of hue; saturation; and brightness, HSV colour
$I$	An image
$I_c; I^S$	Set of pixels with colour $c$ in an image $I$ ; set of $N$ images from shot $S$
$i; j$	Indices
$k_i$	Lexical index term $i$
$k$	An example image, an example shot
$K$	Set of lexical index terms
$K$	Number of examples in a set
$L^t; l^t(n)$	Ranked list of database items for lexical query; rank for item $n$
$L$	Number of feature spaces
$L^*; a^*; b^*$	Values of luminance ( $L^*$ ); and chrominance ( $a^*, b^*$ ), CIE Lab colour
$L_1; L_2; L_\infty$	City-block distance; Euclidean distance; Chebyshev distance
$l$	Feature space index
$MAX(A, B)$	Fuzzy maximum operator for sets $A, B$

$MIN$	Minimum rank combination operator
$MIN(A, B)$	Fuzzy minimum operator for sets $A, B$
$MINDIST$	Minimum distance combination operator
$m$	Number of term weights
$N; n_{i,j}$	Number of image samples; Number of index terms in document $j$
$p; p_n$	Extended Boolean model $p$ -norm, Minkowski norm, a pixel, probability of null hypothesis; normalised test probability of null hypothesis
$p(x, y)$	A pixel associated with the coordinate $x, y$
$Q; Q_{\text{and}}; Q_{\text{or}}$	Query image; conjunctive set of query terms; disjunctive set of q. terms
$\mathbf{q}$	Feature vector of query weights
$R$	Reference image
$R^f(k)$	List of database items ordered by similarity to example $k$ in concept $f$
$r^f(k, n)$	Confidence rank for the database item $n$
$S^t; s^t(n)$	Ranked list of database items for concept query; rank for item $n$
$S^f; s^f(n)$	Ranked list of database items for the concept $f$ ; confidence for item $n$
$S$	A video shot
$s^2$	Sample variance
$sim(q, d_j)$	Similarity by the $p$ -norm for the extended Boolean model
$SUM$	Additive combination operator
$T_P$	Set of true positive pixels
$t$	Query definition
$t_{\text{lex}}$	Lexical query
$U$	Universe of fuzzy sets
$u$	An element of the fuzzy universe $U$
$V^t; v^t(n)$	Ranked list of database items for visual query; rank for item $n$
$var_R$	Relative variance
$w_{i,j}; w_s; w_v; w_t$	Text feature; weights for concept; visual; and lexical search modalities
$W; W_n$	Wilcoxon value; Wilcoxon value for normalised test
$z; z_n$	$z$ -ratio for the Wilcoxon test; $z$ -ratio for a normalised Wilcoxon test
$\bar{x}$	Sample mean
$\alpha(I, j)$	Feature vector value $j$ of an image $I$
$\alpha_c^{(d)}(I)$	Autocorrelogram algorithm
$\gamma_{c_i, c_j}^{(d)}(I)$	Correlogram algorithm
$\bar{\gamma}_{c_i, c_j}^{(d)}(I)$	Temporal correlogram algorithm
$\mu_A(u)$	Fuzzy membership function for an element $u$ in set $A$
$\Theta$	Combination operator of low-level feature fusion
$\Omega$	Combination operator of example fusion or search modality fusion
$\Psi^t; \psi^t(n)$	Final ordered result set for the query definition $t$ , final rank for item $n$



# Contents

Abstract	
Acknowledgements	
List of Original Publications	
List of Symbols and Abbreviations	
Contents	
1 Introduction .....	15
1.1 Background.....	15
1.2 Research Problem and Hypothesis .....	18
1.3 Research Scope and Approach.....	18
1.4 Contributions and Summary of Original Papers.....	19
2 Literature Review .....	23
2.1 Information Retrieval .....	23
2.2 Content-based Retrieval of Video.....	28
2.3 Low-level Visual Features .....	32
2.4 Semantic Concept Features .....	34
2.5 Combining Features and Modalities for Search.....	35
2.6 User Interaction and Relevance Feedback.....	37
2.7 Similarity Measures.....	38
2.8 Video Retrieval Evaluation.....	39
2.8.1 Evaluation Measures .....	40
3 Manual Video Retrieval.....	41
3.1 Low-level Visual Feature Descriptors .....	41
3.1.1 Detecting Temporal Colour Changes.....	42
3.1.2 Colour Histograms and Colour Correlograms .....	44
3.1.3 Temporal Colour and Gradient Correlograms .....	46
3.1.4 Experiments in Visual Similarity Retrieval .....	49
3.1.5 Summary .....	51
3.2 Semantic Concept Descriptors.....	52
3.2.1 Visual Concept Detectors .....	52
3.2.2 Experiments in Semantic Concept Detection .....	54
3.2.3 Summary .....	56

3.3 Experiments in Manual Video Retrieval.....	57
3.3.1 Experiments with Semantic Filtering .....	57
3.3.2 Experiments with Visual, Semantic and Lexical Feature Combinations .....	59
3.3.3 Summary .....	63
4 Interactive Video Retrieval .....	64
4.1 Sequential Queries and Relevance Feedback .....	64
4.2 Cluster-temporal Browsing.....	65
4.3 Interactive Retrieval Experiments .....	68
4.3.1 Experiments with Cluster-temporal Browsing .....	69
4.3.2 Experiments with Browser Configurations .....	71
4.3.3 Statistical Validation of Cluster-temporal Browsing Experiments .....	73
4.3.4 Summary .....	74
4.4 Overview of the Results in TRECVID Video Retrieval Benchmarks.....	75
5 Summary .....	77
References	
Original Publications	

# 1 Introduction

## 1.1 Background

“Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, ‘memex’ will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.”

Vannevar Bush, *As We May Think*, 1945

Over 60 years ago, Vannevar Bush envisioned a theoretical analogue device that stores large amounts of human-generated information (Bush 1945). Bush introduced a tool that would support the human thought processes by allowing the creation and display of ‘trails’ of information through an unordered mass of stored media. His ideas were an inspiration to the creators of the contemporary hypertext and hyperlink concepts. The ‘memex’ was in essence created from the need for a machine that would extend human biological memory and “give man access to and command over the inherited knowledge of the ages” (Bush 1945). It was a conceptual design of the futuristic information technology that did not consider one important problem present with the contemporary design of a ‘memex’ like multimedia systems, the indexing of the stored media. In his design, Bush assumed that user would manually create a descriptive index of a new document into a personal code book. Today, after the revolutionary deployment of Bush-inspired hyperlink technology on the World Wide Web, one of the key research issues of information systems deals with the development of technology that would facilitate the indexing and creation of metadata for the ever-growing surge of digital media.

Content-based retrieval (CBR) research strives to create a retrieval system that utilises digital content in the indexing process in a way that is ultimately independent of manual work. CBR is an umbrella term for content-based multimedia retrieval (CBMR), content-based visual information retrieval (CBVIR), content-based image retrieval (CBIR),

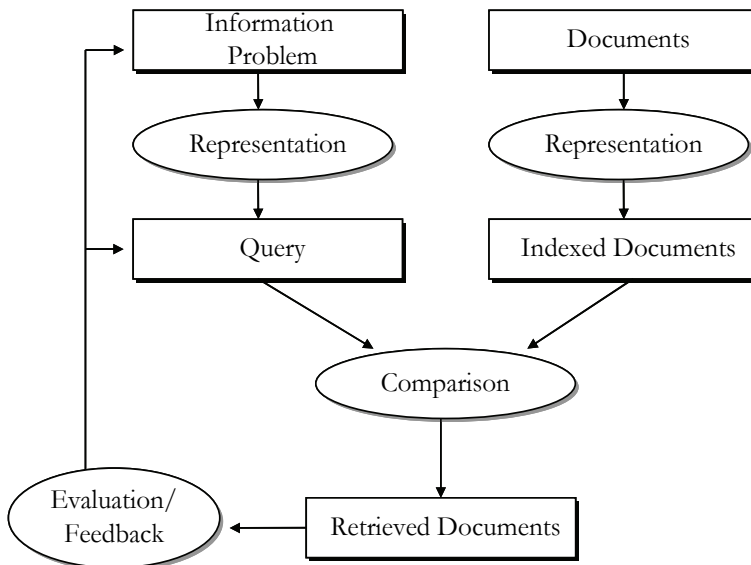
content-based video retrieval (CBVR) and content-based audio retrieval (CBAR). CBR may also be referred to as multimedia information retrieval (MIR). CBR originates from the fields of pictorial databases and visual information management, (Blaser 1979, Tamura & Yokoya 1984, Jain 1992) where the first systems were textual database management systems with manually-created metadata and a query language for pictures.

In addition to the early pictorial database development, another branch of research emerged with the focus on time-dependent information. The first studies that have been influential in the content-based analysis of temporal data were from the fields of knowledge representation and artificial intelligence. The articles from McDermott (1982) and Allen (1983) introduced models for describing relations between temporal events. Based on Allen's temporal logic, Abe *et al.* (1989) built a scene retrieval method for video database applications using temporal condition changes.

During the 1990's CBR research rapidly gained more momentum in automatic content analysis and understanding for indexing audio, images and video (Wold *et al.* 1996, Flickner *et al.* 1995, Tonomura *et al.* 1994). In order to accelerate the use of content-based technology, Moving Pictures Expert Group MPEG initiated the standardisation of the multimedia content description interface MPEG-7 in October 1996. MPEG-7 was approved as a standard in 2002. (Manjunath *et al.* 2002.)

CBR communicates closely with information retrieval (IR) that studies models for text document storage and retrieval. IR distinguishes itself from simple exact match techniques and data retrieval as it focuses on information that is relevant to the search task instead of data itself. It also operates with incomplete queries, partial relevance and natural query language. (van Rijsbergen 1979)

Information retrieval presents a commonly adopted model for the search processes, illustrated in Fig. 1.



**Fig. 1. Basic information retrieval processes (Croft 1993).**

The figure shows that the retrieved document set is the result of a computational comparison between a human-generated query and indexed document representation. Since the results are susceptible to having varying degrees of relevance, a retrieval system attempts to maximise the relevance of all retrieved documents using ordered lists.

The challenge for retrieving maximal relevance lies in the uncertainty of representations generated by both the human and the system. The information problem of a person may not be well articulated when the search process is started; it can simply be a vague idea that is associated with some retrieval need. Even if the person would have a concise definition of her information need, she might find it difficult to formulate with the system's query syntax. Due to this difference between user intentions and query formulation, retrieval becomes inaccurate with mere exact matching of parameters. Therefore there has been a large focus on creating ranking algorithms that sort results by their calculated relevance. Another problem identified in IR is document representation, which may have altering syntax for the same semantic information (Oard & Dorr 1996). This challenges the association of relevant documents to the query definition. These uncertainties are also known as the semantic gap in retrieval research. In order to reduce the uncertainty, IR models have included feedback routines in the system to converge towards the maximal relevance iteratively. This is also evident in Fig. 1.

CBR adopts the previously described basic information retrieval process and applies it to the domain of digital multimedia. The problem of the semantic gap escalates as the data extends beyond linguistic description: pictures, music, sound and video. Another problem arises with the visual data in particular. The sensory gap makes the description of the world objects an ill-posed problem and yields to uncertainty in what is known about the state of the object. For example, when recording a two-dimensional image from a three-dimensional field, part of the information is lost, which leads to ambiguity of the two-dimensional representation. (Smeulders *et al.* 2000.)

Successful content-based systems have many application areas and targeted user groups. Dimitrova *et al.* (2002) divide users broadly into two extremes that require different functionality from a CBR system. Non-technical consumers require simplicity and robustness, whereas technical professional corporate users who regularly use the system can be trained for more sophisticated use. According to Dimitrova *et al.* (2002), both user groups have a need for content-based technology, but professional users are better motivated to learn new practices even before the technology has fully matured. They mention some example applications in professional and educational areas such as automated authoring of content for the World Wide Web, searching and browsing for large video archives, easy access to educational material, indexing and archiving multimedia presentations, and indexing and archiving multimedia collaborative sessions. In the consumer domain they describe applications for video overview and access, video content filtering and enhanced access to broadcast video.

## 1.2 Research Problem and Hypothesis

In this thesis, the research problem is: What computational content-based models for retrieval reduce the semantic gap in order to accomplish effective semantic retrieval of digital video? Moreover, this work concentrates on the following sub-problems: What is the significance of the low-level computational descriptors in retrieval effectiveness? How do you connect the computational content descriptions to higher level, i.e. human understandable, concepts? What types of user interface techniques can aid the user to converge towards semantically meaningful search results more efficiently? Application-wise, the goal is to create a prototype of a video retrieval system that provides access to the task-related relevant content without laborious human involvement in the annotation and categorisation of the media.

The main hypothesis of this study is that the semantic gap can be significantly narrowed using pattern recognition, data abstraction and clustering techniques jointly with human interaction through manually created queries and visual browsing.

## 1.3 Research Scope and Approach

Smeulders *et al.* (2000) describe two domains for image collections. A narrow domain has limited and predictable variability in all relevant aspects of its appearance. A broad domain has unlimited and unpredictable variability in its appearance even for the same semantic meaning. This division can be expanded to every type of multimedia, including video collections. This division has a profound influence on the development of content-based methods. Whereas the narrow domain allows for the development of specialised methods, such as frontal face recognition for biometric recognition, methods for broad domain result in partial description of the data. As an example, frontal face detection is not capable of detecting all faces from a collection of holiday images in which human posture and distance are unconstrained by nature.

This thesis focuses on the development of ad-hoc retrieval methods for broad domain collections of digital video and images. Research experiments focus on testing the main hypothesis using both system-oriented automatic and task-oriented user tests. Experiments on video retrieval are realised in the context of annual international evaluation for content-based systems. The evaluation framework defines search tasks and provides the test and development corpus, relevance judgements and the evaluation of results in two distinct tasks: manual and interactive video retrieval.

Research objectives can be divided into the following subtasks: development and evaluation of low-level features; development and evaluation of semantic concept detectors; evaluation of combination techniques for text, low-level and semantic features; and development and evaluation of content-based interaction techniques between user and the search system. The overall improvement in semantic search effectiveness for different methods is achieved by the comparison against the conventional search paradigm of sequential queries and relevance feedback.

The comprehensive realisation of a video retrieval system calls for technology from the fields of video segmentation, automatic speech recognition and machine language

translation. However, they are not the focus of this thesis. Video segmentation and speech transcripts are received from external sources and employed in the prototype search system to construct the baseline video search functionality. The computation is generated off-line in this work and there is no specific focus on the computational cost of the developed methods, instead the focus is on retrieval methods that provide semantically the most relevant results.

## 1.4 Contributions and Summary of Original Papers

The following is a summary of the contributions of this thesis:

1. A novel method for extracting a low-level temporal colour description from video sequences.
2. A novel method for extracting a temporal structure description from video sequences.
3. A generic technique for training and detecting semantic concept confidences from video sequences.
4. A fusion technique for employing a content-based search based on visual, concept and text features in semantic retrieval from a large test video corpus.
5. A novel content-based browsing technique and application for interactive video database searching and navigation.

Paper I gives a comparison of low-level colour features in visual content-based image retrieval. Two colour feature paradigms were compared in HSV and RGB colour spaces: the autocorrelogram contained a spatial colour distribution and the colour histogram described the global colour structure. Selected features were experimented in semantic image category retrieval using data from Corel, QBIC and SUNET databases. The results showed that the spatial organization of colours in a perceptually oriented HSV colour space provide the best initial retrieval performance. The author was responsible for the development of colour descriptors and participated in the writing of the manuscript with Prof. Ojala whereas the actual experimentation was realised by Mr. Matinmikko and Mr. Aittola.

Paper II introduces the Temporal Color Correlogram (TCC) feature, which is an extension to the spatial HSV autocorrelogram as it operates in a temporal video domain. Semantic search experiments with 11 hours of video data from TREC 2001 video track showed that the TCC was more capable of finding relevant video sequences than the static key frame-based approaches, the HSV Color Correlogram and HSV Histogram. The author was responsible for the development of the descriptors and experiments whereas Dr. Doermann was the supervisor and participated in the finalisation of the manuscript.

Paper III introduces an adaptive colour change-based detector for detecting colour changes from image sequences. Two colour spaces, HSV and CIE Lab, were evaluated in sequences from easy, moderate and difficult environmental conditions. The adaptive detection of colour changes in the CIE Lab colour space lead to the most balanced performance in easy and moderate conditions, whereas HSV provided the largest proportion of true positives. This article offers directly a minor contribution to the main

scope of this work, which is the development of content-based retrieval methods. However, the experimental findings influenced the further development of colour features for visual retrieval. For example, the instabilities in colour changes discovered in low pixel intensities affected the design of colour space quantisation schemes. Nevertheless, the contribution of this paper to the main hypothesis is less than that of the other articles. The author was responsible for the development of the method and experiments. Dr. Ojala and Dr. Kauniskangas participated in the research discussion and finalisation of the manuscript.

Paper IV introduces Temporal Gradient Correlogram as a visual video feature extracting spatial edge information from luminance data in frame sequences. A simple training of semantic concepts using self-organizing maps and its efficiency in the detection of people, cityscape and landscape concepts was demonstrated in a five-hour test video collection from the TREC 2002 video track. Additionally, speech and instrumental sound detection was performed using an audio feature-based classification in the test video database. The author was responsible for developing the visual concept detectors. Mr. Penttilä and Mr. Peltola were responsible for the development of audio-based concept detectors. Prof. Seppänen was the supervisor and participated in the finalisation of the manuscript.

Paper V describes semantic concept detection experiments with Temporal Color Correlogram, Temporal Gradient Correlogram and Motion Activity features using several detector configurations. Experiments were operated on a 60-hour U.S. news video database from the TRECVID 2003 retrieval benchmark. Detection results for 12 concepts showed that the combination of low-level features based on ranked lists gave better detection performance than the combination of normalised feature point distances. The minimum rank fusion of temporal colour and structure features was found to provide the performance closest to manual validation with the evaluated concepts. The author was responsible for the development and testing of the concept detectors and Prof. Seppänen was the supervisor and participated in the finalisation of the manuscript.

Paper VI analysed the performance of visual, concept and text features in content-based video retrieval. Experiments with 24 semantic search topics in 60 hours of video data from TRECVID 2003 showed that content-based search with concept detectors and visual examples improve the overall search performance over conventional text search based on speech recognition and closed captions transcripts. The author designed and coordinated the video retrieval experiments that were carried out by the video retrieval research team. Prof. Seppänen and Prof. Ojala were the supervisors and participated in the finalisation of the manuscript.

Paper VII introduced cluster-temporal browsing as an interaction technique for video database navigation. The content-based search engine presented was based on self organising maps combined with semantic and lexical filters. Results obtained from the experiments in a 40-hour TREC 2002 video track test set indicated that the performance of the interactive search system was several times better than that of the non-interactive system. The author was responsible for the design of the cluster-temporal search paradigm and coordinated the video retrieval experiments that were carried out by the video retrieval research team. Mr. Kai Noponen implemented the lexical features. Mr. Jani Penttilä, Ms. Satu-Marja Mäkelä and Mr. Johannes Peltola from the VTT Technical Research Centre were responsible for the development of speech and instrumental music

detectors. Mr. Dmitri Vorobiev implemented the generic audio features. Prof. Seppänen and Prof. Ojala were the supervisors and participated in the finalisation of the manuscript.

Paper VIII extended the work from paper VII towards a more efficient interactive search system. Filtering was replaced with the combination of lexical and concept features. Interactive semantic search experiments with eight test users on a 60-hour test video database of TRECVID 2003 showed that the performance of cluster-temporal browsing based on visual similarity can be improved by combining it with text content similarity. The author designed and coordinated the video retrieval experiments with the cluster-temporal browser that was implemented by the video retrieval research team. Prof. Seppänen and Prof. Ojala were the supervisors and participated in the finalisation of the manuscript.

Paper IX described experiments with the cluster-temporal browsing of videos and content-based video queries with relevance feedback. The experiments were organised with novice test users on 24 semantic search tasks in a 70-hour TRECVID 2004 test video database. Results showed that the cluster-temporal browser gives good improvement in search performance over the traditional content-based query paradigm, which follows sequential queries with iterative relevance feedback between the user and the system. The author designed and coordinated the video retrieval experiments that were carried out by the video retrieval research team. Prof. Seppänen and Prof. Ojala were the supervisors and participated in the finalisation of the manuscript.

Paper X described the video retrieval experiments with the cluster-temporal browser in a large 80-hour multilingual video corpus from a TRECVID 2005 evaluation. Experiments with novice and expert system developers were conducted and cluster-temporal browser effectiveness was contrasted against the sequential queries and relevance feedback. Also, search effectiveness with content-based search features was evaluated against a baseline text search in non-interactive, manually created queries. Results showed that the novice users were able to improve their search effectiveness with the help of the cluster-temporal browser. Content-based search features were found to improve search effectiveness over the baseline speech transcript search. The author designed and coordinated the video retrieval experiments that were carried out by the video retrieval research team. Prof. Seppänen and Prof. Ojala were the supervisors and participated in the finalisation of the manuscript.

Paper XI described the statistical validation for the experimental results in the TRECVID 2004 and TRECVID 2005 evaluations. The Wilcoxon signed rank test was employed in the results of two system variants; a conventional search paradigm of sequential queries and relevance feedback was contrasted against the system augmented with a cluster-temporal browser. The cluster-temporal browser was found to bring statistically significant improvement over the conventional search paradigm with novice system users, whereas system developers did not have any statistically significant differences between the system configurations. The author carried out all statistical experiments described in the paper. Prof. Seppänen and Prof. Ojala were the supervisors.

The video retrieval experiments have been carried out using a large prototype search system that incorporates several search technologies and interface paradigms. The collaborative efforts of the system development team at the MediaTeam research group have been vital for the successful realisation of the system and the experiments. Several people have been contributing to the development of the retrieval system during the years

of development; Matti Hosio, Ilkka Hanski, Matti Varanka, Jialin Liu, Jukka Kortelainen, Anu Pramila, Kai Nojonen, Ilkka Juuso, Timo Koskela, Esa Matinmikko, Doctor Hannu Kauniskangas, Markus Aittola, Heikki Keränen and Kimmo Hagelberg are gratefully acknowledged.

## **2 Literature Review**

### **2.1 Information Retrieval**

When computerised document management became viable, system development focused on the retrieval and storage of data elements with structured databases. The database management systems were dependent on a numeric and constrained language structure for accessing the data. In order to provide more intuitive document access and management, the field of information retrieval (IR) began to emerge with a focus on the retrieval of documents that embody content relevant to a user's information need (van Rijsbergen 1979). IR research focused on developing text retrieval models and was initially used in bibliographic databases but afterwards retrieval methods were extended towards full text search. Concepts such as the relevance and probability ranking principle have made IR distinctive from exact data retrieval. (Sparck Jones & Willett 1997) The relevance of the information content to the search need can not be defined precisely. Instead, IR models provide a computational estimate, goodness of fit, of the document to meet the user's information need. This estimate is utilised to arrange the retrieved documents into a list of results with the most relevant documents on top.

IR system development focuses on maximising retrieval effectiveness by estimating the relevance. Computational matching of documents is hampered by both the vagueness of the representation of users' information need at the intermediary interface and the vagueness of information content representation as a provider of substance to the information need. This complexity of relevance leads to a semantic gap between the user and the system, thus affecting overall user satisfaction with the system. In this light, Sparck Jones & Willett (1997) describe the role of IR system concisely: A retrieval system seeks to capture the relevance relation by establishing a matching relation between the two expressions of information in the document and the request, respectively. In this definition the expressions of information may or may not define search task-related information with adequacy, thereby satisfying the information need. When a retrieval system is designed it is equally important to address it both from the user and document perspectives to make the different characterisations of information converge and produce better retrieval effectiveness.

Information retrieval research classifies two operational modes for an information retrieval system: ad hoc and filtering. Ad hoc retrieval consists of a scenario, in which new queries are directed to a rather static document collection thereby creating an information pull, i.e., the user takes the initiative to receive the required information. Filtering, conversely, is a push type activity, as the queries remain relatively static while new documents are being fed into the system. (Baeza-Yates & Ribeiro-Neto 1999) The work presented in this thesis focuses on the ad hoc retrieval scenario.

Several models in IR focus on improving the effectiveness of text document retrieval. Although the retrieval of text differs from the retrieval of other multimedia content, many of the IR models are generic and can be utilised in the retrieval of other types of media. The following is a brief description of the classical IR models. These models provide the theoretical framework for several contemporary IR techniques and they have also functioned as the source of inspiration for several CBR studies. Knowledge about these models will assist in gaining a better point of view on the latest work in CBR. A more thorough overview of these methods can be found from the literature (Baeza-Yates & Ribeiro-Neto 1999, Sparck Jones & Willett 1997).

Probably the most commonly used model for retrieval is a set-theoretic *Boolean retrieval model*. This model is mostly used in a data retrieval scenario in which the exact correspondence of the query parameters and database attributes is regarded as a match. There is no distinction between the degrees of relevance since all the matching documents are considered to be equally relevant to the search task. Since the retrieval model is set-theoretic, it is capable of constraining and expanding the result set with the use of the logical operators *or*, *and* and *not*. Therefore queries requesting documents with the terms ‘forest *and* pine’ will not retrieve documents in which one of the desired terms is missing. On the other hand, the query ‘forest *or* pine’ returns documents containing at least one of the search terms. With the Boolean model the user is left with an unordered set of results in which relevant documents have to be identified by hand. This will become laborious with queries that return large result sets.

As a solution to the limitations of the Boolean model, Salton *et al.* (1975) presented a *vector space retrieval model* for the retrieval of text documents. Instead of returning unordered sets, the vector space model assigns non-binary weights to the document index terms. These values are utilised in computing the degree of similarity between the query and index terms. The degree of similarity helps a retrieval system to return documents with a higher relevance at the top of the ordered list of results. In the vector space model the document  $d_j$  is associated with a vector  $\mathbf{d}_j$  of non-binary weights  $w_{i,j}$  for each index term  $k_i \in \mathbf{K}$  ( $\mathbf{K} = \{k_1, \dots, k_I\}$ ). Additionally, a query vector is constructed from the weights of the index terms in the query definition:  $\mathbf{q} = (w_{1,q}, \dots, w_{I,q})$ . In order to compute the degree of similarity between the query and document vectors, the correlation between vectors should be measured. In information retrieval, the most common measure is the cosine of the angle between the vectors (Baeza-Yates & Ribeiro-Neto 1999)

$$\text{sim}(d_j, q) = \frac{\mathbf{d}_j \bullet \mathbf{q}}{|\mathbf{d}_j| \times |\mathbf{q}|} = \frac{\sum_{i=1}^I w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^I w_{i,j}^2} \times \sqrt{\sum_{j=1}^I w_{i,q}^2}}, \quad (1)$$

here  $|\mathbf{d}_j|$  and  $|\mathbf{q}|$  are the vector norms. The former term gives the normalisation of the similarity measure whereas the latter remains constant throughout the query procedure. Baeza-Yates & Ribeiro-Neto (1999) write that the benefit of using vector models is the partial matching of documents to the query as it reduces the need for creating precise correspondences between the query definition and relevant documents. Ranking the results in a vector space may return documents with high relevance even when only parts of the query terms match the terms in the document index. There are many alternatives to the computation of weights for the document index, but the most fundamental in the retrieval of text documents is based on term frequency inverse document frequency (TF-IDF). Briefly, the weight of a single term  $w_{i,j}$  is computed using both the frequency of the term in the parent document  $j$  and the inverse of the frequency of documents containing the term

$$w_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^I n_{i,j}} \cdot \log\left(\frac{|DB|}{|(d_j \supset k_i)|}\right), \quad (2)$$

where  $n_{i,j}$  is the number of occurrences of  $k_i$  in document  $j$ ,  $|DB|$  is the number of the documents in the database, and  $|(d_j \supset k_i)|$  is the number of documents containing the term. More details about the vector model can be found from the work of Salton *et al.* (1975).

A generalisation to the vector space model in IR was introduced by Wong *et al.* (1985). Their model addresses the independence assumption of the index terms in a more theoretical setting. Instead of regarding index terms independent and orthogonal, as in the original vector space model, the principal idea of the generalised vector space retrieval is to use co-occurrence to map interdependent terms into independent minterm vectors. These vectors are pair-wise orthogonal, so they are theoretically more fitting for the similarity measurement. The ranking of the documents against the query is computed using minterm vectors and cosine similarity (see Eq. 1).

Another algebraic method maps the index term vector space to lower dimensional space. The *latent semantic indexing retrieval model* (Deerwester *et al.* 1990) reduces the dimensionality of the index term space using singular value decomposition. Firstly, the TF-IDF weights create a term correlation matrix. Second, singular value decomposition is employed to the index term matrix and the largest singular values outline the low-dimensional representation of the index terms. The reduced dimensionality can be considered as a conceptual view on the documents. Similarities between the query and the documents are measured by modelling the query as a pseudo-document and finding the most similar documents in the projected concept space.

Connectionist learning based on *neural networks* has also been used in IR. Belew (1989) introduced a neural network that used index terms, authors and documents to change its representation over time according to user feedback. The network adapted to the behaviour of a user group and created a representation of a consensual meaning of keywords and documents shared by a group of users. Wilkinson & Hingston (1991) introduced a three layer neural network consisting of queries, index terms and documents. As the query propagates through the network to the document nodes, another round of iteration is initiated through spreading activation of the term nodes using the document terms. Iteration continues as the term nodes may generate new activation signals to the document nodes. Through several iterations, activation signals attenuate and the network stabilises to a steady state in which relevant document nodes have been activated. As the documents may activate term nodes outside the query definition during the iterations, the network operates as a dynamic thesaurus and can find documents that are topically related but do not necessarily use terminology defined in the query. The authors also present a method to incorporate user feedback to the network.

Robertson & Sparck Jones (1976) presented the *binary independence retrieval model*. This naïve Bayes model estimates the probability that a user will find a document relevant to his search task. The model describes the existence of an index term as a binary value and the likelihood for each term appearance is assumed independent of the other index terms. The model is based on the concept of an ideal answer set that will have maximum relevance. Ranking of the relevant results is based on minimising the probability of a false judgement as it uses the ratio of the two probabilities: a document belonging to the relevant set and a document belonging to the non-relevant set. Initially, these probabilities have to be set arbitrarily and only after some user feedback on the document relevance can the probabilities be properly adjusted. It is worth noting that even though the independence assumption does not hold in reality for many situations, linguistic text as an example of this, naïve Bayes models have proven to be very competitive against more complex models (Domingos & Pazzani 1997).

Turtle & Croft (1991) have proposed the *Bayesian inference network model*, which assigns random variables with the index terms, texts, documents, query concepts and queries. According to this generative model, the document random variables represent the observation of the document during the search process. The network can be represented graphically as a directed acyclic graph that is formed from parent nodes, nodes and the edges that connect them. The nodes represent the various random variables and the parent nodes hold the prior probabilities. The graph is structured on conditional independence, i.e., the nodes are conditioned on their parents and their descendants are conditioned on them. In the graph, conditional relationships are shown by edges. The graph model helps in propagating the known probabilities through the network and enables the combination of multiple query configurations for improving retrieval performance. The model can adapt to various information retrieval ranking strategies, such as Boolean and TF-IDF.

Another group of IR models are based on fuzzy sets. The *Fuzzy information retrieval model* (Ogawa *et al.* 1991) is based on an assumption where the query is formed as a fuzzy set of terms and the relevance of a document is defined as a degree of membership for this set. The degree of membership is a value between 0 and 1, where 1 indicates full membership and 0 corresponds to non-existing membership. The real strength of the fuzzy sets is in the operators that combine separate sets using set-theoretical operations

similar to Boolean model. The most common operators are complement (*not*) of a single set, disjunction (*or*) and conjunction (*and*) of two separate sets. The mathematical descriptions of the fuzzy set-theoretic functions are

$$\mu_{\bar{A}}(u) = 1 - \mu_A(u), \quad (3)$$

$$\mu_{A \cup B}(u) = \text{MAX}(\mu_A(u), \mu_B(u)), \quad (4)$$

$$\mu_{A \cap B}(u) = \text{MIN}(\mu_A(u), \mu_B(u)). \quad (5)$$

The formulae above describe complement, union and intersection operators respectively. Symbols  $A$  and  $B$  describe two fuzzy sets in the universe of  $U$  of which  $u$  is an element. Ogawa *et al.* (1991) utilised fuzzy set theory in their fuzzy information retrieval model but instead of using the original definitions for fuzzy operators they introduced the use of an algebraic sum as a replacement for the *MAX* function in the union of sets as well as an algebraic product to replace the *MIN* function in the intersection of sets. In addition, they introduced a simple thesaurus for expanding the query. The correlation factors for index terms pairs were computed into a keyword connection matrix.

Salton *et al.* (1983) introduced the *extended Boolean retrieval model*, which, faithfully to its name, extends the traditional Boolean queries by altering the conjunctive (*and*) and disjunctive (*or*) operations towards a vector space model. The underlying principle is to relax the constraints of the original Boolean algebra by introducing partial matching and term weighting to the operations. Salton *et al.* proposed a general  $p$ -norm model that at the other extent operates as a fuzzy theoretic model using *MAX* and *MIN*, and at another extent as a vector space model using inner product. The formulae for the  $p$ -norm are

$$\text{sim}(\text{Q}_{\text{or}}, d_j) = \left( \frac{w_{1,d}^p + \dots + w_{m,d}^p}{m} \right)^{\frac{1}{p}}, \quad (6)$$

$$\text{sim}(\text{Q}_{\text{and}}, d_j) = 1 - \left( \frac{(1 - w_{1,d})^p + \dots + (1 - w_{m,d})^p}{m} \right)^{\frac{1}{p}}, \quad (7)$$

where  $\text{Q}_{\text{or}}$  and  $\text{Q}_{\text{and}}$  describe the disjunctive ( $k_1 \vee k_2 \dots \vee k_l$ ) and conjunctive ( $k_1 \wedge k_2 \dots \wedge k_l$ ) sets of query terms  $k_i$ .  $w_{i,d}$  is the associated term weight for the pair  $[k_i, d_j]$  that measures the degree where  $k_i$  is associated with the document  $d_j$ . The term weights may be computed using the product of term frequency and normalised inverse document frequency. The definitions of the  $p$ -norm formulae show that the parameter  $p$  can be used to adjust the operator characteristics. Setting  $p = 1$ , Eqs. 6 and 7 become

$$\text{sim}(\text{Q}_{\text{and}}, d_j) = \text{sim}(\text{Q}_{\text{or}}, d_j) = \frac{w_{1,d} + \dots + w_{m,d}}{m}, \quad (8)$$

when  $p = \infty$

$$\text{sim}(Q_{or}, d_j) = \text{MAX}(w_{i,d}), \quad (9)$$

$$\text{sim}(Q_{and}, d_j) = \text{MIN}(w_{i,d}), \quad (10)$$

when  $p = 2$ , the similarities are measured as the Euclidean distances from the point  $[0, 0]$  in the case of disjunctive query and from the point  $[1, 1]$  with conjunctive query. the extended Boolean retrieval model provides a flexible framework to adjust the level of strictness in the operators. It combines vector space and the fuzzy set-theoretic models orderly while maintaining the possibility of constructing complex queries using the Boolean notation. The flexibility of the extended Boolean retrieval model is utilised in this work in the fusion of content-based features for video retrieval.

The models presented here have become the basis for the advanced IR methods and can be considered as the classical set of IR techniques. Relevance feedback as an interaction method originates from IR but will be introduced later. This overview did not cover the recent advances in the field of IR, such as the popular language models for IR (Ponte & Croft 1998, Hiemstra 1998, Croft & Lafferty 2003), but the focus was instead on models that are related to the models used in machine learning and pattern classification (Duda *et al.* 2001). The models presented here have already been adapted to CBR. Some of them can be used directly for multimedia document retrieval by replacing or adding features from another modality, or modalities, whereas the rest require more thoughtful adaptation to the multimedia domain. Some recent examples of adapting language models to video retrieval can be found from the works of Westerveld & de Vries (2005) and Mc Donald & Smeaton (2005).

The vector space and extended Boolean models were described with closer scrutiny, as they are central to the experimental video retrieval system that is introduced in this work. Vector space based models have become popular in multimedia retrieval research. In these models the relevance can be translated as a measurement of geometric closeness between the points in multidimensional space. In contrast, probability-based retrieval models, for instance, are generative, i.e., they require *a priori* assumptions about the underlying distributions in the source data, which may be difficult to recognise from the complex structures in heterogeneous multimedia documents.

## 2.2 Content-based Retrieval of Video

Retrieval of multimedia data has surfaced in the past decade. As a definition, multimedia is an integration of at least two of the following: text, graphics, animation, full-motion images, still video images, or sound. Recorded video data is a synchronised presentation of audio and motion image tracks. This section introduces some background information for image, audio and multimedia retrieval at a level that is relevant for understanding the path that has led to the research of content-based video retrieval.

As was described earlier, research in CBR emerged from the need to address the problems with the early pictorial database management systems. Chang & Fu (1980) introduced the concept of query by pictorial example where satellite images were

assigned as the source of evidential relevance to the search task at hand. At that time, the query example was translated into a parametric query in a relational database. The idea of example-based queries has since become a common paradigm in the realisation of CBR queries (Faloutsos *et al.* 1994).

In the early 1990's researchers began to propose automatic indexing schemes (Kato 1992, Chang & Hsu 1992). In the same year, in a workshop on visual information management systems (Jain 1992), major research directions were identified in the fields of databases, computer vision, and knowledge representation and management tightly paired with concrete and practical applications. The workshop envisioned the need for highly parallel processing architectures, especially for the upcoming video database systems. Parallel processing is an important part of the retrieval interaction model presented in our work, but improvement of the parallel processing techniques as such will not be the emphasis of the scientific experimentation.

As the outcome of the workshop, computer vision research was advised to focus on interactive image understanding and interactive environments whereas the knowledge representation research focus was directed towards combining symbolic and non-symbolic representation at the same level. The guidelines from the workshop are largely valid for contemporary research on multimedia retrieval; although the CBR research field is still very much shaping the underlying theories, some of the most promising progress has been made in interaction techniques and methods that attempt to bridge the gap between the symbolic and non-symbolic representations of multimedia data.

Related work in CBR from the perspective of images can be found from the overview studies of Rui *et al.* (1999), Smeulders *et al.* (2000), Vasconcelos & Kunt (2001), Eakins (2002), Kherfi *et al.* (2004), Datta *et al.* (2005) and from the books of Chen *et al.* (2004), Dunckley (2003), Santini (2001), Lew (2001), and Del Bimbo (1999).

During the 1990's several prototype systems were introduced for image retrieval: Chabot (Ogle & Stonebraker 1995), Photobook (Pentland *et al.* 1994), QBIC (Flickner *et al.* 1995), VisualSEEk (Smith & Chang 1996), Virage (Bach *et al.* 1996). At that time, several novel techniques were introduced in an operational retrieval framework but they did not become popular in commercial products.

Images and the visual domain captured the majority of research interest in multimedia retrieval at the expense of auditory media. The ideas from pictorial databases and CBIR have typically been adaptable to video retrieval by simple modifications. Visual automation has inspired researchers in computer vision for decades. Although the theories for visual similarity and image understanding are still concealed in the complex entanglement of human visual system and cognition, the idea of sound computational models for visual understanding would enable a wider selection of application areas than the corresponding models for the auditory signals.

Nevertheless, audio plays a significant role in the understanding of video content. In the genres of movies, commercials and music videos it is used to transmit important information about the content. Audio analysis can help in video segmentation by detecting music, speech, silence and speaker segments as well as in the creation of video skims (Li *et al.* 2006). Sound effects assist in the recognition of the video setting or scene and musical styles have meaning in cinematic language. In contemporary movies, commercials and music videos, music and sound effects are used in the cinematic narrative together with the spoken dialogue to augment the main plotline via multimodal

storytelling. In these genres, audio analysis can help in recognising a cinematic structure. Overview studies on the methods that use audio content in multimedia indexing and retrieval can be found from Wang *et al.* (2000), Snoek & Worring (2005), Scaringella *et al.* (2006).

Traditionally in the genres of news and documentary videos most of the audio channel is used for speech whereas auditory sounds and music are used only to emphasise structural changes in the narrative. In order to attract more viewers, the genre of documentaries has shifted towards more a cinematic style of storytelling, in which music and sound effects are used to support the main storyline, similar to the genre of movies. News and documentaries are informative by their fundamental nature, even when they adopt contemporary techniques in their presentation. Therefore the most important content will typically be communicated linguistically either by the speech of a narrator, dialogue or textual transcripts overlaid on the visual surface. The transcription of the communicated language into text document indexes using automatic speech recognition and optical character recognition (Zhong *et al.* 2000, Li *et al.* 2000) has proven successful in realising more semantic retrieval of video. During the years of experiments with Informedia Digital Video Library, Hauptmann (2005) found that retrieval precision based on speech recognition transcripts is nearly as effective as with perfect transcripts when the word error rate is below 40 %. He also reports that the optical character recognition has been found useful in specific applications, especially in the association of faces with names in a video.

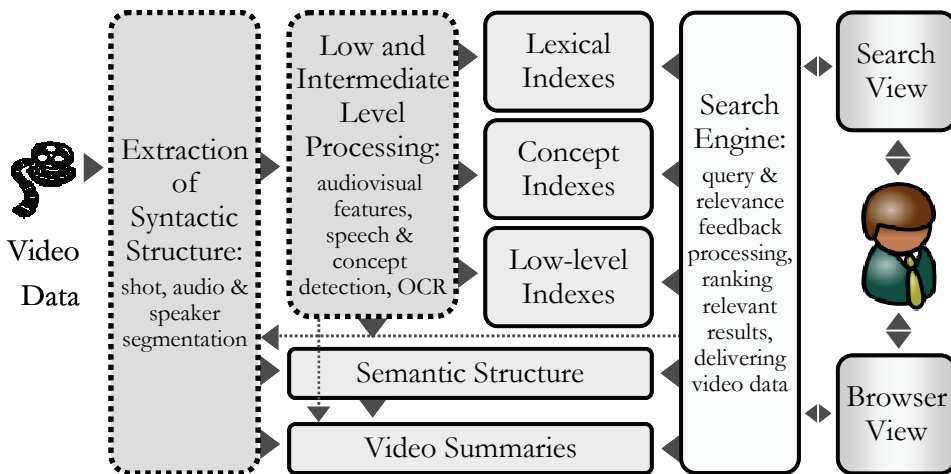
The early work on temporal logic by McDermott (1982) and Allen (1983) paved the way for the temporal event-based retrieval of video (Abe *et al.* 1989). Around that time an indexing method based on two-dimensional spatial relations, 2-D strings, of symbolic pictures was introduced by Chang *et al.* (1987). A few years later Day *et al.* (1995) combined spatial and temporal predicates into a directed graph model for spatio-temporal video retrieval. Due to the complexity involved in typical video objects and their spatio-temporal trajectories, these techniques have become popular in specialised domains, such as visual surveillance, but not in the generic CBVR.

Recent development in video retrieval has focused on models that combine several modalities for joint indexing and retrieval. Whereas Deerwester *et al.* (1990) proposed latent semantic indexing as a transformation of textual documents and queries to the conceptual space, Sclaroff *et al.* (1999) extended this approach to associate textual terms and visual indexes at the conceptual level. They reported improvement in effectiveness for the proposed combination of textual and visual features in the WWW image database retrieval using relevance feedback. Wang *et al.* (2000) describe several low-level features from audio and visual domains. First they show that the features for audio, colour and motion provide independent information about the video, and then review several audiovisual techniques for shot detection and classification, video scene content classification and film genre recognition. As their conclusive remark they note the difference between a text and an audiovisual document: “for the same semantic concept, there are relatively few text expressions, whereas there could be infinitely many AV renditions. This makes the latter problem more complex and dynamic.”

The above-mentioned problem has been widely acknowledged and semantic retrieval models for multimedia retrieval have become the subjects of increasing interest in recent years (Androustos *et al.* 2006). Generally retrieval based on speech and optical character

recognition transcripts offer a higher level of semantic video access than models based on the query-by-example with low-level feature similarities. But not all relevant information is entailed in the language transcripts. An overview study on TRECVID video retrieval evaluation (Hauptmann & Christel 2004) demonstrated that the semantic retrieval for visually-oriented search topics, such as ‘people walking in groups in an urban setting’, can not be effective by relying simply on a textual search with narrative transcripts.

Fig. 2 portrays a generic model for a content-based video retrieval system. As the source video is supplied to the system, processing commences with the extraction of the syntactic structure. The segmentation of the video divides it into shots from which the key frames are extracted. The audio track is segmented separately into speech and music, and speech is divided into speaker segments. Next, the analysis of the video continues with low-level feature extraction. A separate module shapes a summary or skim of the video. Automatic speech recognition and optical character recognition produce text transcripts. Semantic concept detection and semantic structure analysis process a higher level description from the audiovisual content. Low-level, lexical and concept indexes are created and associated with the syntactic and semantic structure. A retrieval module processes queries from the user, adapts to the relevance feedback from the user, accesses feature indexes, combines the multiple results into a final ranked list and returns the list to the user interface. The user interface offers access to the database through search and browsing views and delivers database videos using their logical and semantic structure.



**Fig. 2. The structure of a content-based video retrieval system.**

The syntactic structure of a video begins with static still images and short-time audio frames. The next syntactic unit is the sequence of images, a shot, which is comprised of a consistent visual flow created by a camera run or an animation sequence. The boundaries of shots contain gradual or abrupt visual transitions. A consistent audio sequence does not necessarily align with visual transitions and is usually carried over several visual shots, such as the narrative track in a documentary film. A group of shots creates another syntactic layer, a scene, where consistent visual sequences can be grouped with narrative

rules. An example of a scene is a collage of shots that depicts an overview of a parade. Higher level grouping of scenes into elements of a story is part of the video's semantic structure and may not necessarily follow any specific syntactic rules, albeit the reverse is also possible. An example of the latter is a news story which usually holds a clear structure: an introductory scene with news anchors is often followed by reportage outside the studio. By recognising the syntactic and semantic structure from the video, it can be summarised automatically. Zhang *et al.* (1995) presented a technique for shot boundary detection and key frame selection, which has become the first step in many automatic video indexing systems. A survey on video abstraction techniques that build upon a recognised video structure can be found from Li *et al.* (2006). Hoashi *et al.* (2005) introduced a technique for story segmentation using a support vector machine (SVM) classifier on automatically detected shots and extracted audiovisual features. The method was demonstrated in ad-hoc story segmentation as well as in a practical application using online training. The authors concluded that whereas good performance can be achieved with ad-hoc databases, segmentation of unknown television programs remains challenging for the current techniques. The technique of Hoashi *et al.* received top-level performance in a TRECVID 2005 story segmentation task.

Several knowledge-based techniques have been published for domain-specific indexing, typically for sports, news and surveillance videos. A more comprehensive overview of the field of CBVR can be found from the review studies of Yoshitaka & Ichikawa (1999), Brunelli *et al.* (1999), Al-Khatib *et al.* (1999), Hauptmann & Christel (2004), Snoek & Worring (2005), Lew *et al.* (2006) and Xiong *et al.* (2006). Several books have been recently published about multimedia retrieval (Shih 2003, Furht & Marques 2003, Feng *et al.* 2003). Several prototype video retrieval systems have been introduced for CBVR: Informedia (Wactlar *et al.* 1996), MARS (Huang *et al.* 1996), Virage (Hampapur *et al.* 1997), VideoQ (Chang *et al.* 1997), Fischlár (Smeaton *et al.* 2004).

### 2.3 Low-level Visual Features

Smeulders *et al.* (2000) covered a large collection of studies on low-level visual features. They outlined three groups of methods for processing the image data for the features: colour, local shape and texture. They categorised spatial image grouping techniques into global, image partitioning, sign location, strong and weak segmentation. The features for indexing were grouped into global and accumulative, of which the former is a holistic feature, such as the average colour, and the latter can be a (geometric) histogram, moments, or a compression transform. Within the scope of this thesis, this section introduces global techniques that are based on colour and texture. Shape features are not considered as they require grouping of data with strong segmentation or sign location techniques. These techniques are not used in our work. Reviews from Rui *et al.* (1999), Smeulders *et al.* (2000), Manjunath *et al.* (2001) and Datta *et al.* (2005) give a comprehensive overview of the studies on all low-level visual feature types.

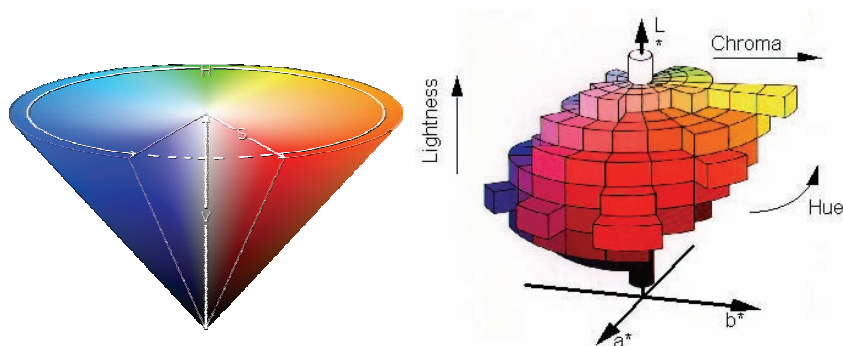
Mojsilovic *et al.* (2000) reported subjective experiments on psychophysical properties of the human visual system in the perception of colour patterns. The authors found that

the separate visual processing pathways based on luminance and chrominance as well as the processing of achromatic patterns in later stages affect the similarity perception between colour patterns. They showed that the processing of edge-based textures and colour information yields to retrieval performance and behaviour similar to human. Similarly in our work, visual correspondence between video shots is computed using colour features and local differential geometrical properties of luminance, i.e., local texture features.

Since the landmark work by Swain & Ballard (1991), colour has been identified as a powerful cue in the estimation of visual similarities. According to Smeulders *et al.* (2000) “image colour has superior discriminating potentiality of a three-dimensional domain compared to the single dimensional domain of grey-level images”. Possibly for that reason colour has been frequently utilised in visual image and video retrieval.

Colour space configuration plays a significant role in the use of colour. Some of the most common colour spaces are RGB, HSV and CIE Lab. The differences between the three are in the representation of chromaticity and luminance information. RGB colour space originates from the 1950’s colour television standards. In RGB space the colours are presented with the additive primaries of red, green and blue attempting to match the colour reception of the human eye. Luminance is incorporated into the encoding of the three colours. Therefore, colour constancy and perceived colour similarity do not correspond mathematically to the RGB colour representation and render RGB space less suitable for measuring visual correspondence. Smith (1978) introduced a non-linear transformation from RGB into HSV colour space. HSV is based on the Munsell colour system where colour is represented by hue, lightness and chroma. In the HSV model, H stands for colour hue, S for colour saturation and V for colour lightness. By separating the hue from the illumination-dependent saturation and luminosity a degree of illumination independence can be obtained. This has been found important for object recognition and colour indexing. In 1976, the International Commission on Illumination (CIE) introduced a reference colour model,  $L^*a^*b^*$  or Lab, as the derivation from the CIE 1931 XYZ colour space in order to create a linear presentation for the perceptible colour differences.

The CIE Lab colour space is closer to the human visual system as it is designed to be perceptually uniform. As with HSV, CIE Lab separates the luminosity and chromaticity:  $L^*$  signifies luminance,  $a^*$  shows the levels of magenta-green and  $b^*$  levels of blue-yellow. Although the colour space enables the measurement of similarity between two colours using Euclidean distance, it has a disadvantage. The CIE Lab colour space is irregularly shaped and its quantisation for indexing purposes is not trivial. Therefore it has not been extensively used in colour-based indexing unlike HSV, which retains regular geometric proportions (Mojsilovic *et al.* 2000). Several other colour spaces have been proposed especially for object recognition purposes where illumination invariance is a necessity. References to the related work can be found from Smeulders *et al.* (2000).



**Fig. 3. Left: HSV cone visualisation<sup>1</sup>. Right: visualisation of the CIE Lab 1976 colours<sup>2</sup>.**

In the 1990's colour-based features evolved from colour histogram statistics (Swain & Ballard 1991) to colour moments (Stricker & Orengo 1995), colour regions (Hsu *et al.* 1995), colour coherence vectors (Pass *et al.* 1996), HSV colour sets (Smith & Chang 1996) and colour correlograms (Huang *et al.* 1999). A colour histogram has limited discriminative power as it ignores the spatial organization of colours. Huang *et al.* demonstrated that the autocorrelogram, a subset of a colour correlogram, creates competitive performance to histograms and coherence vectors by utilising a local colour structure. Ma & Zhang (1998) benchmarked colour correlograms, colour histograms, colour moments and colour coherence vectors and came to similar conclusions.

Recently, new colour descriptors have been introduced and many incorporate spatial information as well. The MPEG-7 content description standard defines the following descriptors: dominant colour, scalable colour, colour structure and colour layout (Manjunath *et al.* 2001). Hadjidemetriou *et al.* (2004) showed good texture retrieval results with a multiresolution histogram that combines colour histogram characteristics and spatial colour structure.

Some of the well-known texture features are co-occurrence matrix (Haralick *et al.* 1973), wavelets (Daubechies 1990), MRSAR models (Mao & Jain 1992), Wold features (Liu & Picard 1996) and local binary patterns (Ojala *et al.* 2002). Good performance with local binary patterns in face localisation and recognition tasks has recently been reported in the works of Ahonen *et al.* (2006) and Hadid *et al.* (2004).

## 2.4 Semantic Concept Features

Studies about semantic modelling of video have focused on intermediary techniques for higher level conceptualisation of the information-rich audiovisual content. In Naphade *et al.* (1998) and Naphade & Huang (2000) a framework of probabilistic multimedia objects for modelling audiovisual semantics was proposed and demonstrated with the detection

<sup>1</sup> Source: [http://en.wikipedia.org/wiki/HSV\\_color\\_space](http://en.wikipedia.org/wiki/HSV_color_space)

<sup>2</sup> Source: <http://www.kodak.com/US/en/corp/researchDevelopment/technologyFeatures/gamut.shtml>

of concepts such as sky, water, forest, rocks and snow. The authors proposed Multinet, a network of probabilistic concept detectors, to enhance and infer the automatic description of semantic video content. Naphade & Smith (2003) extended Multinet into a hybrid framework for combining discriminative and generative models for concept detection. Also Chang *et al.* (1998) and Del Bimbo (2000) introduced automatic semantic concept detectors for describing video content with a higher level of semantics. Snoek & Worring (2005) have introduced the following semantic index hierarchy with varying descriptive granularity: named events, logical units, sub-genre and genre. In a survey of techniques for extracting audiovisual semantics, Naphade & Huang (2002) concluded that building methods for semantic detection can be divided into two parts: building models for the actual semantic concepts and modelling the knowledge base and contextual setting that govern the set of concepts being modelled.

Christel & Hauptmann (2005) evaluated the utility of various semantic concept types in video retrieval tasks and concluded that generic search topics are likely to benefit more from the concept queries than specific topics. Retrieval by specific concepts is prone to errors if the user has a different conception of the query concept, or if the concept does not overlap with any search need. However, in the few cases where specific concepts are relevant to a search topic, retrieval performance can be greatly improved. Two common semantic video features studied by the TRECVID participants are detection of person and monologue (Ikizler & Duygulu 2005, Snoek *et al.* 2004, Yang & Hauptmann 2004) as the person-related topics are common in queries on informational video genres, such as news.

A generic detector approach is proposed by Iyengar & Nock (2003), who describe a general information fusion algorithm for building arbitrary concept models from multimodal features. Super-kernel fusion of independent modalities (Wu *et al.* 2004) has been proposed as an optimal combination of multimodal information where statistically independent components of modalities are fused using two layers of classifiers. The authors report promising results for TRECVID 2003 semantic concept detection.

An interesting recent direction for modelling multimedia content is the establishment of the semantic description from the user perspective. Hanjalic & Xu (2005) propose a computational framework for affective video content representation and modelling based on two-dimensional emotion space using *arousal* and *valence* as the dimensional axis. Santini *et al.* (2001) propose that the image meaning can only be revealed by contrasting the image with other images in the feature space. The authors introduce emergent semantics as the result of interaction between the user, mediator interface and the database.

## 2.5 Combining Features and Modalities for Search

Section 2.1 described IR models for text document retrieval. In the area of multimedia retrieval, several models have been adopted and extended to work with multiple modalities. Combining the evidence from different modalities should result in better relevance by the search system. However, the most effective techniques for multimodal combination have not yet been identified.

Mc Donald & Smeaton (2005) adopted joint probability and evidence combination methods from text IR (Fox & Shaw 1994) and evaluated them with multiple example and feature combinations as well as multimodal retrieval by visual and lexical features. They concluded that the additive combination of separate feature and example scores or ranks offered better overall search performance than the joint probability of individual retrieval models. Snoek *et al.* (2005) evaluated early and late fusion techniques in semantic concept detection with text and visual features and concluded that late fusion results in better performance more often than early fusion but with increased training cost. Amir *et al.* (2004) described retrieval experiments with late fusion of speech transcripts and visual features. The best observed fusion configuration was a linear combination scheme with query-dependent weights.

De Vries *et al.* (2004) evaluated a generative probabilistic model for combining static and dynamic visual models with a language model in video retrieval experiments. The combination of dynamic visual and language models with joint probabilities was found robust and comparable with a language model alone and the combination of multiple examples was found to overcome single example queries. Several other combination techniques from IR have been proposed for video retrieval, such as latent semantic analysis (Souvannavong *et al.* 2004). Christel *et al.* (2004) evaluated content-based filters for image and video retrieval. They focused specifically on the effect of filter robustness in baseline text search performance. They concluded that the current quality of the semantic filters is not sufficient for making content-based retrieval and browsing more efficient.

The combination of multiple features typically requires adjustment of feature weights. Weighting places emphasis on the evidence that is most fitting for the document relevance assessment regarding the user's relevance expectations. The weights can be adjusted through relevance feedback from the user, but recent research has studied the possibility of automatic weight adjustment. Yan *et al.* (2004) studied the effect of query-independent and class-dependent weights in video retrieval. The authors proposed a technique for classifying user queries into four distinct categories: named person, named object, general object and scene query. Each category was trained to have different linear weights to improve the retrieval for each query class. The retrieval engine was based on a two-level hierarchical mixture-of-experts architecture with probabilistic outputs and linear expert combinations. The authors found significant performance improvement over configuration with query-independent weights.

By observing the related work, the successful techniques for multimodal combination in video retrieval have so far been late fusion, linear combinations, lexical and visual features, and query class-dependent weighting. The attractiveness of the linear methods is based on simplicity and robustness in the contemporary retrieval experiments. However, Yan & Hauptmann (2003) derived theoretical limits for linear combination methods and concluded that linear methods become insufficient when the number of retrieval components increases. As a solution, authors propose focusing on non-linear combination functions.

## 2.6 User Interaction and Relevance Feedback

Rodden *et al.* (2001) evaluated the effect of organisation by image similarity for information retrieval tasks. They created a simulated work task situation in which users were given alternative interfaces to accomplish given search tasks. The authors concluded that organisation by visual similarity seemed to be useful for graphic designers as it helped to divide images into simple genres, albeit causing some items to merge and disappear if the image group was considered too homogeneous. Overall, the results of the study encourage constructing alternative search views based on content-based techniques. Hollink *et al.* (2005) assess user behaviour in news video retrieval. They recorded user actions during a search and found that the number of selected items is the most significant contributing factor in the retrieval outcome. In the development of future retrieval systems, the authors recommended focusing on effective visualisation of the dataset and improved facilities for browsing.

Rocchio (1971) introduced relevance feedback, an IR technique for improving retrieval performance by optimising queries automatically through user interaction. 19 years later, Salton & Buckley (1990) evaluated different variations of relevance feedback techniques in text document retrieval and reported a 50% - 150% improvement over initial performance after first iteration. Relevance feedback has also been successfully employed in CBR. In brief, the interactive relevance feedback technique adjusts subsequent queries based on user relevance judgements in order to direct retrieval towards more relevant results. Yan & Hauptmann (2004) propose co-retrieval as an automatic weighting and re-ranking scheme for video retrieval. As with relevance feedback, co-retrieval adjusts query weights according to the relevance information. However, the user does not provide the relevance information interactively. Instead it is automatically estimated from the noisy relevance labels in the initial text retrieval results. The authors show good retrieval performance on TRECVID 2003 video retrieval data.

Adcock *et al.* (2005) described the design and implementation of an interactive video search system that focuses on assisting humans in locating relevant video results. Their system uses story segmentation, key frame montages, tooltips and colour coded indicators of relevance on a video timeline. The authors report good results in a TRECVID 2004 interactive task. Christel & Conescu (2005) presented an interactive Informedia interface, which consists of a storyboard showing chronologically ordered key frames, a playback window, a panel for collecting answers and a panel for defining query parameters. The authors report that novice users preferred text queries and a view configuration that displayed both lexical and visual information from the videos. Heesch & Ruger (2004b) introduced three different techniques for content-based access to image collections. They proposed a two-step  $NN^k$  search technique that was found to improve quantitatively on traditional relevance feedback techniques and on the traditional query-by-example paradigm. Sav *et al.* (2005) introduce an interactive object-based video retrieval technique that follows ostensive video retrieval. The proposed technique allows for the collection of sets of query objects, which are grouped in clusters based on content-based similarity to represent the user's object-based information need. The user can further continue searching with any of the created groups and return to previous groups at any time, as the ostensive model for information retrieval suggests.

Interactive systems provide alternative routes for searching and browsing as well as informative interface elements to facilitate user-based navigation towards the relevant results. Many of the techniques introduced here have been placed well in interactive video retrieval in TRECVID evaluations (TREC Video Retrieval Evaluation Home Page n.d.).

## 2.7 Similarity Measures

Vector space models in content-based retrieval are based on computational similarity between points in a feature space. Similarity measurement is effectively a computation of inverted dissimilarity by measuring the distances. In this work, similarity refers to the inverse of normalised computational dissimilarity  $D$ , which can be computed as  $1 - D$ . For the statistical features, an alternative is to use non-parametric statistical tests as the measure of similarity between two distributions. Selecting the best similarity metric for complex multi-dimensional data is not a straightforward choice. Eidenberger (2003) evaluated 38 similarity measures to find the best for eight MPEG-7 visual descriptors in clustering annotated image categories. The report showed that among the best general purpose metrics, albeit not the best, are the Minkowski norms. Therefore selecting them as the similarity metric for the retrieval of unknown multimedia data permits an assumption of reasonable performance. The computational simplicity of Minkowski norms and their resulting popularity in content-based retrieval is also reflected in this work as the metric of choice for most of the experiments. The mathematical definition of the Minkowski norm is

$$D_p(Q, R) = \left[ \sum_{j=1}^n |\alpha(Q, j) - \alpha(R, j)|^p \right]^{\frac{1}{p}}, \quad (11)$$

where  $\alpha(I, j)$  denotes  $j^{\text{th}}$  feature vector value of image  $I$ ;  $Q$  is the query image and  $R$  the reference image; and  $p$  is the Minkowski norm. By setting  $p = 1$ , the norm becomes the city block ( $L_1$ ) distance whereas  $p = 2$  is equal to the Euclidean distance ( $L_2$ ). A recent study by Howarth & Rüger (2005) suggests that the retrieval performance of  $L_1$  and  $L_2$  could be improved by fractional distance metrics where  $p < 1$ . This work adopts Huang *et al.* (1999) by using city block ( $L_1$ ) as the similarity measure for spatial correlation statistics.

A work by Swain & Ballard (1991) with colour histograms proposed a histogram intersection as the similarity measure. A few years later, Hafner *et al.* (1995) proposed a weighted Euclidean distance for histogram similarity as follows:

$$D_H(Q, R) = [h(Q) - h(R)]^t A[h(Q) - h(R)], \quad (12)$$

where  $A$  is a weight matrix for the colour pairs and  $[h(Q), h(R)]$  are the histograms. A weighted Euclidean has been utilised in our work to compute dissimilarities between colour histograms.

## 2.8 Video Retrieval Evaluation

The National Institute of Standards and Technology (NIST) has organised Text Retrieval Conferences since 1992. The focus has been on evaluating information retrieval systems in a common framework on large datasets. In 2001, TREC started a track for video retrieval evaluation that expanded to a separate event in 2003 (TRECVID).

Principally, the TRECVID evaluation constitutes of a video collection divided into development and test databases; a common segmentation and speech recognition transcripts on the data; a set of 24 or 25 semantic search topics with lexical definitions and examples designed to imitate real users' requests, such as finding specific people, objects or instances of object types, activities or locations or instances of activity or location types; search experiments at the participating research groups; submission, pooling and manual evaluation of the participants' result lists at NIST; and finally evaluation of the results using standard retrieval measures such as average precision, r-precision and precision at varying recall levels. Each year the search topics change and the size of the video collection increases. Several video genres have been included in the collections with the main emphasis being on broadcast news. Lately, the collection has been expanding towards multi-lingual content (TREC Video Retrieval Evaluation Home Page n.d.).

In the TREC 2002 video track, the second running of the video retrieval evaluation, the video database contained 68 hours of advertising, educational, industrial, and amateur films produced between the 1930's and the 1970's by corporations, non-profit organizations, trade associations, community and interest groups, educational institutions, and individuals from the Open Video (Open Video Project n.d.) collection and Internet Archive's Prelinger Archives (Internet Archive: Moving Image Archive n.d.). 17 research groups participated in the experiments worldwide and a total of 27 runs were submitted for the manual retrieval task. 13 runs were submitted for the interactive task. (Smeaton & Over 2002)

The TRECVID 2003 video retrieval evaluation used 133 hours of MPEG-1 video of U.S. broadcast news from ABC and CNN, commercials and C-SPAN programming captured in 1998. The test material was divided in half for the development and test collections. 24 groups participated in the 2003 experiments and 38 runs were submitted for the manual retrieval task. The interactive runs had 36 submissions. For the TRECVID 2004, test and development databases of 2003 were swapped and a total of 52 manual runs and 61 interactive runs were submitted. (Smeaton *et al.* 2003) (Kraaij *et al.* 2004)

The TRECVID 2005 evaluation had a collection of U.S. news, Arabic and Chinese broadcast video totalling about 169 hours. 74 hours was English, 43 hours was Arabic and 52 hours was Chinese content. A total of 43 participants attended 2005 experiments and 26 runs were submitted for the manual retrieval task. The interactive task received 49 submissions. (Over *et al.* 2005)

### 2.8.1 Evaluation Measures

TRECVID uses a set of established measures to evaluate the retrieval effectiveness from the retrieval results. In order to enable numerical evaluation, truth information about the relevant and non-relevant items is needed. The most common measures are precision and recall. Precision describes the proportion of items that are relevant within a result set. Recall indicates the proportion of all relevant items in the collection that has been successfully retrieved.

$$\text{precision} = \frac{|\{\text{relevant items}\} \cap \{\text{retrieved items}\}|}{|\{\text{retrieved items}\}|}, \quad (13)$$

$$\text{recall} = \frac{|\{\text{relevant items}\} \cap \{\text{retrieved items}\}|}{|\{\text{relevant items}\}|}. \quad (14)$$

In order to evaluate the ordered result sets at length, precision can be measured as a function of fixed rank points or as a function of recall. The fixed rank precision can be measured at specific rank values, such as precision at rank 5, rank 10, rank 100, and rank 1000. These values portray the development of precision as the size of the result set grows. As an example, if three of the top five results are considered relevant, the precision at rank 5 is 3/5. If recall is used, it is typical to measure an 11-point recall using recall values from 0 to 1 with 0.1 intervals. Commonly, precision decreases when recall grows and it is not simple to grasp the performance behind the series of precision values. Therefore, recall-precision curves are used to plot the performance visually. (Harman 1995) The F-measure is another formula that summarises precision and recall using the weighted harmonic mean. Mean average precision (MAP) is commonly used in this work as an overall performance measure. It is computed as the mean of the average precisions over a set of queries. Average precision (AP) is the average of the precision values measured after each relevant document is retrieved. The relevant documents that have not been retrieved are counted with zero precision. Relevant items located at the top of the list have typically more weight in average precision.

Statistical tests are important in validating the significance of absolute performance differences between two system configurations for the same topics. Zobel (1998) recommends using a non-parametric Wilcoxon signed rank test for information retrieval experiments in order to achieve better robustness and reliability of results. The Wilcoxon test is computed using result pairs. First, the absolute differences between pairs are computed and an ordered list is created. The differences are replaced with rank values but the sign is retained. The Wilcoxon value  $W$  is computed by aggregating the ranks. The boundary for the statistical significant value of  $W$ , rejection of null hypothesis set at the level of 5%, depends on the number of available topic pairs  $n$ . Zobel (1998) recommends the normalisation of average precisions using the highest obtained performances for the queries in order to eliminate varying difficulty between topics.

## 3 Manual Video Retrieval

Manual retrieval is a process in which a user manually creates a query based on a specific information need and receives a direct system response in return. By definition, it does not take into account in any further interaction between the user and the search system. More importantly, manual retrieval supplies an initial ranked list for the interaction procedure in an actual search scenario. The measurement of the effectiveness of the manual retrieval gives an indication of search engine performance before applying any task related learning to the system. From the interactive system perspective, it is essential to enhance performance in manual retrieval to provide more relevant seeds for the interaction phase and allow the search process to converge towards the relevant results faster. Since the initial query is defined by the user, search effectiveness is dependent on the capability to express the task-related search need with the syntax specified by the system.

This chapter introduces the fundamental components of a search engine and a model for combining several features. Section 3.1 introduces low-level visual features for the video search engine and experiments in content-based visual retrieval. Section 3.2 describes a technique for propagating concept labels using small example sets and provides results for the concept detection experiments. Section 3.3 introduces a model for the prototype search engine utilising a combination of visual, concept and semantic features and reports experimental results of the proposed search engine in manual search tasks of the TRECVID video retrieval evaluation.

### 3.1 Low-level Visual Feature Descriptors

Visual features were introduced in several papers. Papers I, II and III described colour-based descriptors for images and video. Paper I compared HSV and RGB colour spaces with histograms and correlograms and Paper III evaluated CIE Lab and HSV colour spaces in colour change detection. Papers II and IV introduced novel video features based on local colours and edges in the temporal video data.

In addition to features based on chrominance and luminance, Papers V and VII described content-based indexing experiments with motion activity (Manjunath *et al.*

2002), which is a composition of intensity, direction, spatial distribution and temporal distribution statistics extracted from the estimated motion data. (1) Intensity of motion is a discrete value where high intensity indicates high activity and vice versa. Intensity is defined as the variance of motion vector magnitudes normalised by the frame resolution and quantised in the range between 1 and 5. (2) Average intensity measures the average length of all macro block motion vectors. (3) Spatial distribution of activity indicates whether the activity is scattered across many regions or in one large region. This is measured using the short, medium and long runs of zero motion blocks and it provides information about the size and number of moving objects in the scene. Each attribute is computed using the thresholded flow vectors from the video data.

The contribution of this section is primarily on the experiments of low-level colour features, secondarily on the experiments of low-level structural features and thirdly on experiments with the motion features. The selected emphasis ignores low-level feature types that typically are extractable from a video source, such as features from audio and object shapes. The texture properties of a video image are represented to an extent by the introduced features as they measure the correlations of spatial edges and colours.

### 3.1.1 Detecting Temporal Colour Changes

Low-level colour features capture the characteristics of the chromatic information channel in an image. Humans experience colours as a result of the complex processing of data from the retinal colour receptors and antagonistic receptive fields. The higher-level colour correspondences are experienced through the arrangement of colours and illumination in the visual plane. The development of perceptually oriented colour spaces has allowed better mathematical approximations for colour similarity measurement. Paper III studied two perceptually oriented colour spaces in the detection of visual changes from sequential images. The objectives of this study were primarily to develop a method for autonomous colour change detection from image sequences and secondarily to discover the characteristics of pixel colours as the first order descriptors of visual changes.

Paper III studied the applicability of two colour spaces, HSV and CIE Lab, for identifying colour changes in sequential images by automatic thresholding. The presented approach comprises the following steps: reference image acquisition, sample image acquisition, calculation of the difference image, spatial thresholding of the difference image, morphological post-processing of the resulting binary image, and finally colour change detection from the post-processed binary image.

Given a reference image  $R$  and a sample image  $I$ , the difference image  $Diff$  is computed in the CIE Lab colour space as

$$Diff_{Lab} = \sqrt{(\Delta a^*)^2 + (\Delta b^*)^2 + (\Delta L^*)^2}, \quad (15)$$

where  $\Delta L^* = L_R^* - L_I^*$ ;  $\Delta a^* = a_R^* - a_I^*$  and  $\Delta b^* = b_R^* - b_I^*$ .  $L^*$ ,  $a^*$  and  $b^*$  are the three colour channels of CIE Lab space.

The pixel-wise difference in HSV colour space is computed with the following formula

$$Diff_{HS} = \sqrt{(\Delta S_X)^2 + (\Delta S_Y)^2 + (\Delta V)^2}, \quad (16)$$

where  $\Delta S_X = S_R \cos(H_R) - S_I \cos(H_I)$ ;  $\Delta S_Y = S_R \sin(H_R) - S_T \sin(H_T)$ ;  $\Delta V = V_R - V_T$ ;  $H, S$  and  $V$  are the hue, saturation and brightness channels.

The automatic spatial thresholding of the colour differences is obtained using stochastic Poisson modelling for the colour noise (Paper III). The underlying assumption is that in a difference image, the noise and detected colour changes follow the Poisson distribution. The assumption holds when the noise is uniformly distributed and real colour changes from objects clump together into local clusters. The automatic algorithm selects the threshold by testing different levels using relative variance

$$var_R = \frac{s^2}{\bar{x}}, \quad (17)$$

where  $s^2$  and  $\bar{x}$  correspond respectively to the variance and mean of the thresholded pixel distribution in local  $32 \times 32$  windows. When  $var_R \rightarrow 1$ , the sample distribution approaches the Poisson distribution. The objective is to find the optimal threshold value that would minimise the appearance of global noise in the sample windows while maximising the regions with real colour changes. Therefore a threshold value maximising the relative variance is selected as the global threshold for the difference image.

Paper III compares the performance of the proposed colour change detection in two colour spaces using test images from shipping docks. The test sets contain reference images of the empty docks followed by sequential images containing objects typical to the context, such as empty cardboard boxes and packing material. In order to make chromaticity relevant for detecting changes, many of the objects are matte surfaced and have luminosity levels close to the original scene. The test image sets contain three different types of scenes categorized as ‘easy’, ‘moderate’, and ‘difficult’, based on properties such as heterogeneity of the background, existence of shadows, illumination changes, and the reflectivity and chroma properties of the objects. 0 illustrates the results with HSV and CIE Lab colour spaces using the proportion of detected object pixels of all true object pixels ( $T_P$ ) and the proportion of erroneously detected background pixels of all detected pixels ( $F_P$ ). If  $Det$  denotes the set of detected pixels,  $GT$  the set of ground truth pixels with real colour changes caused by objects,  $\overline{GT}$  the background pixels, and  $|\cdot|$  the amount of elements in a set then performance measures  $T_P$  and  $F_P$  can be defined as

$$T_P = \frac{|Det \cap GT|}{|GT|}, \quad (18)$$

$$F_P = \frac{|Det \cap \overline{GT}|}{|Det|}. \quad (19)$$

Table 1 indicates that the CIE Lab colour space produces 18 percentage units less true positives than the HSV over all scene types. Overall false positive rates, in contrast, favour the CIE Lab colour space. These values reveal that the change detection using relative variance in CIE Lab space has a higher tendency to over-threshold an image, which makes it more conservative in complex scenes. This can be seen by examining the false positive rates of the ‘moderate’ scenes; HSV shows a considerable increase in false positives whereas CIE Lab does not. ‘Difficult’ scenes, however, are too complex for both configurations. Although the overall results support HSV, Paper III describes CIE Lab as more robust in change detection at the lower luminance levels, against reflections and in the detection of uniform colour surfaces.

*Table 1. Averages and standard deviations for the true and false positive rates in HSV and CIE Lab colour spaces. The two columns on the far right display the average over the test sets.*

Colour Space	Scene						Overall	
	Easy		Moderate		Difficult		TP	FP
	TP	FP	TP	FP	TP	FP		
CIE Lab								
AVG	0.68	0.10	0.62	0.17	0.37	0.35	0.56	0.21
STD	0.15	0.10	0.28	0.33	0.28	0.38	0.24	0.27
HSV								
AVG	0.82	0.12	0.77	0.41	0.62	0.36	0.74	0.30
STD	0.12	0.06	0.31	0.33	0.36	0.37	0.26	0.25

The system described in Paper III shows the low-level feature extraction phase in the application of environmental scene event detection. The detected regions at the shipping docks separate relevant foreground objects from the background. Further processing of the regions makes the extracted information useful for surveillance and retrieval applications. For example, when detected object boundaries exceed a limit of tolerance, a surveillance system creates an alert for the waste management to clean up the area. In a retrieval scenario, centralised database systems collect encumbrance data from the regional loading areas and create driving routes for the waste transport by retrieving and organising the areas based on the measured encumbrance.

### ***3.1.2 Colour Histograms and Colour Correlograms***

The previous section introduced first order characteristics of pixel colours in the detection of visual changes in sequential images. For the use of content-based indexing, every pixel in an image can not be indexed. Therefore more compact descriptors are needed to represent the chromatic information of images. The objective of Paper I was to evaluate compact colour feature configurations in two colour spaces, HSV and RGB, for the application of content-based indexing of images. The most common feature for describing image colour is a histogram that portrays a quantised colour distribution. The histogram of image  $I$  for colour  $c_i$  is defined as

$$\text{hist}(c_i, I) \triangleq \Pr_{p \in I} [p \in I_{c_i}], \quad (20)$$

where  $I$  is an  $X \times Y$  image that is comprised of pixels  $p(x, y)$ .  $C = \{c_1, \dots, c_N\}$  denotes the set of quantified colours that can occur in the image. For a pixel  $p$ ,  $I(p)$  denotes its colour  $c$ , and  $I_c$  corresponds to a set of pixels  $P$ , for which  $I(P) = c$ . By definition, a histogram corresponds to the probability of any pixel in  $I$  being of colour  $c_i$ . This means that the spatial relationship of colours is ignored. Consequently, this increases ambiguity in the interpretation feature similarities. The colour correlogram (Huang *et al.* 1999) describes the spatial correlation of colours as a function of spatial distance. A correlogram is described as a three-dimensional histogram indexed by triplets  $(c_i, c_j, d)$ . The mathematical definition of a correlogram is

$$\gamma_{c_i, c_j}^{(d)}(I) \triangleq \Pr_{p_1 \in I_{c_i}, p_2 \in I} [p_2 \in I_{c_j} \mid |p_1 - p_2| = d], \quad (21)$$

which describes the probability of finding any pixel  $p_1$  of colour  $c_i$  at a distance  $d$  from a pixel  $p_2$  of colour  $c_j$  in the image. Spatial constraints are defined by a set of fixed distances ( $D = \{d_1, \dots, d_D\}$ ) that are measured using the  $L_\infty$  norm, which is also known as the chessboard or Chebyshev distance. The size of the correlogram is  $O(C^2D)$ .

Huang *et al.* (1999) concluded that a subset of the correlogram capturing the spatial correlation between identical colours, the autocorrelogram, is sufficient for the purpose of image retrieval. The autocorrelogram gives the probability of finding identical colours at distance  $d$

$$\alpha_c^{(d)}(I) \triangleq \gamma_{c,c}^{(d)}(I). \quad (22)$$

The size of the autocorrelogram is  $O(CD)$ . Being smaller than the correlogram, the autocorrelogram is both storage-wise and computationally a more efficient representation for multimedia indexing. In essence, the colour autocorrelogram is a measure of colour compactness in an image. Therefore uniform colour regions in an image create peaks in the autocorrelogram. The colour correlogram experiments in this work are conducted using an autocorrelogram instead of full correlograms. In order to simplify the naming convention they are still denoted as correlograms in the text.



**Fig. 4. Uniform colour regions in an image create peaks in a colour autocorrelogram. Left: Colours of Burano by Temka<sup>3</sup>. Right: Apple Stealer by OliYoung<sup>4</sup>.**

In order to compute a colour histogram or autocorrelogram, the designated colour space needs to be quantised into distinct containers. The granularity of the discretised space determines the compactness of the feature descriptors. Overly dense quantisation may lead to insufficient sampling distribution and the measurements between distributions become statistically unreliable. Varying the granularities of the hue, saturation and brightness in the HSV colour space makes it possible to emphasise the chromatic characteristics over luminosity. This increases illumination invariability and robustness against variations in scene characteristics.

### ***3.1.3 Temporal Colour and Gradient Correlograms***

The previous section introduced colour features with first and second order statistical characteristics for indexing still images. For the purpose of video retrieval, features from the static content neglect the dynamic changes occurring in a video sequence. The objective of Paper II was to develop a feature for short video sequences by extending the colour correlogram feature into a temporal domain. The temporal colour correlogram (TCC) was introduced for indexing video shots by colour, and experimental results in video retrieval tasks were reported. However, colour features are useless in indexing when a video does not contain chromatic data. In order to index monochromatic video sequences, Paper IV introduced the temporal gradient correlogram (TGC) to describe edge information from the achromatic visual channel. Similar to the TCC, the TGC is dynamically computed from the video sequence and can be used as the complementary visual descriptor in content-based video indexing. A short description of the extensions to the temporal domain follows.

---

<sup>3</sup> <http://www.flickr.com/photos/temka/16619242/>

<sup>4</sup> <http://www.flickr.com/photos/oliyoung/98203904/>

Digital video is constructed of image frames  $I$ . The longest sequence of frames that creates continuous, consistent movement within the mechanical limitations of a recording device constitutes a shot  $S$ . The characteristics, such as the chromatic and textural transitions over time as well as consistent motion flow describe useful structural information about the shot. Temporal colour correlograms were developed to describe colour-spatial transitions in a shot. As a feature that describes shot properties, the benefits over traditional approaches, such as static colour histograms or correlograms, are its ability to encapsulate the temporal changes of small spatial colour relations. Fig. 5 illustrates the temporal colour transition. While the temporally computed colour histogram would only portray the proportional amount of colour in the frames, a temporal colour correlogram will capture information about the spatial changes in all pixels with the same colour throughout the entire frame sequence.



**Fig. 5. Temporal colour change illustrated by a frame sequence. A temporal correlogram captures the dispersion of the colour element whereas a histogram does not.**

Let  $N$  be the number of frames  $I$  sampled from a shot  $S$ . The TCC formula is

$$\bar{\gamma}_{c_i, c_j}^{(d)}(S) \triangleq \Pr_{p_1 \in I_{c_j}^S, p_2 \in I^S} \left[ p_2 \in I_{c_j}^S \mid |p_1 - p_2| = d \right], \quad (23)$$

which is a derivation from the Eq. 21.  $I^S$  denotes the set of  $I$  frames taken from the shot  $S$ . As with the static colour correlograms, the computation of the TCC is condensed to an autocorrelogram by setting  $c_i = c_j$  and making it effectively the same as the Eq. 22.

Colour is an important cue in determining visual similarity, but it is not inclusive. As pointed out by Mojsilovic *et al.* (2000) the human visual system uses also achromatic patterns to evaluate similarities between figures. Paper IV introduced the temporal gradient correlogram feature to describe achromatic texture patterns in a video. The TGC is based on image gradients to adopt functionality of the human visual system, in which antagonist retinal receptive fields excite impulses to achromatic neural channels from stimuli caused by spatial lightness contrasts, such as object edges or texture patterns. The TGC is computed from local correlations of specific edge orientations as an autocorrelogram whose elements correspond to the probabilities of fixed edge directions occurring at discrete spatial distances during a video frame sequence. Similarly to the TCC, the TGC is also able to capture temporal changes in spatial edge orientations. From each sample frame, the edge orientations are discretised into four segments depending on whether their orientation is horizontal, vertical or either of the diagonal directions. The TGC feature is thus computing higher-order texture statistics similar to a well-known grey level co-occurrence matrix (Haralick *et al.* 1973). However, the TGC is less dependent on overall luminance levels since the correlation is computed from gradient

edge directions instead of the quantified luminance values as Haralick *et al.* have proposed. First, the gradient image frames are extracted from the sampled video shot using Prewitt edge detection kernels (Prewitt 1970). Prior to the computation of an autocorrelogram the gradients of each pixel with an average magnitude exceeding a pre-defined threshold  $t$  are used to compute a discrete direction image  $Dir$ . Empirically,  $t$  is set to 12 in our experiments. Each  $Dir(x, y)$  contains integer values from 0 to 4. Value 0 indicates that the gradients do not exceed the threshold  $t$ . Values 1 to 4 represent orientations of horizontal, vertical and two diagonal lines with steps of  $\pi/4$ . Finally, the TGC for  $S$  is computed from a set of  $N$  direction images ( $Dir^S$ ). The temporal gradient correlogram formula is

$$\bar{\gamma}_{c_i, c_j}^{(d)}(S) \triangleq \Pr_{p_1 \in Dir_{c_j}^S, p_2 \in Dir^S} \left[ p_2 \in Dir_{c_j}^S \mid |p_1 - p_2| = d \right], \quad (24)$$

which follows the operations described in Eq. 23 with the difference that the correlation is computed from the edges instead of colours.

Two important parameters are required for the computation of the TCC and TGC features.  $N$  defines the number of frame samples from the shot  $S$ . A larger value has a negative effect on computational cost but increases statistical robustness. The theoretical maximum value for  $N$  is the number of frames in a shot whereas the minimum of 1 equals the static computation of a correlogram from a single frame. The proper values for  $N$  can be approximated from a typical encoded video structure, such as the MPEG-1 VCD format, which limits frames in a group of pictures to 15 (PAL) or 18 (NTSC). A typical shot length for a produced video is approximately five seconds. In order to take a sample from every group of pictures in a shot,  $N$  should be above 10. The frames are sampled with a fixed sampling delay to obtain  $N$  images. Paper I experiments with static correlograms ( $N = 1$ ) whereas Paper II uses dynamic sampling with  $N$  less than or equal to 40.  $N$  is set to less than or equal to 20 in the rest of the experiments with temporal correlograms.

Another parametric requirement is to define the set of spatial distances  $D = \{d_1, \dots, d_D\}$ . Specific rules for defining the set of distances do not exist; Huang *et al.* (1999) used  $D = \{1, 3, 5, 7\}$  distances in their experiments with good results. In this work, the same distances are adopted for all correlogram features. Varying the correlation distances affects the scale of the operator. This has to be considered when the source data size varies. The experiments presented in this work normalise the size of the source material prior to computing the feature values. The distances proposed by Huang *et al.* (1999) were originally experimented with images that were  $232 \times 168$  in size. For this work, the same distances have been empirically validated with the frame size of Common Intermediate Format (CIF), in which frame width is fixed at 352 pixels. In the retrieval experiments of this work, every source image is scaled proportionally to the CIF width before visual features are computed.

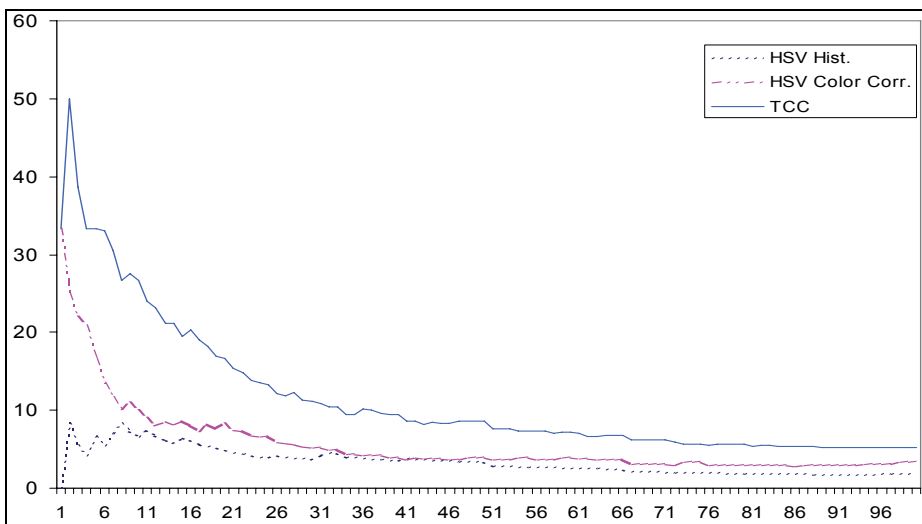
### 3.1.4 Experiments in Visual Similarity Retrieval

Papers I and II describe retrieval experiments with the visual colour features that have been introduced: the HSV and RGB colour correlograms and the TCC. Paper I evaluates the effect of different colour spaces and their quantisation in the semantic category retrieval of images. Eight semantic image categories and their images were used in the category retrieval experiments on the database of 2445 images. 822 images belonged to the semantic categories of ‘sunset’, ‘underwater’, ‘building’, ‘portrait’, ‘item on black background’, ‘item on white background’, ‘urban view’ and ‘beach’ whereas the rest of the data was non relevant to the retrieval task. Each of the category samples was in turn used as a query example to retrieve relevant images from the same category. Both histograms and colour correlograms were computed as described in Eqs. 20 and 22. The experiments evaluated varying quantisations for the HSV colour space using correlograms. Additionally, both RGB- and HSV-based correlograms were contrasted against a HSV colour histogram. The quantisations of the hue, saturation and brightness value (H, S, V respectively) were (7,2,2), (9,3,3), (9,4,4), (9,5,5) and (12,3,3). The colour hue was always more densely quantised than the saturation and brightness value. The computation of similarity to the example image was based on inverted metric dissimilarity in the feature space. For the HSV colour histogram, a weighted Euclidean distance was employed as in Eq. 12.

Paper I reports experimental results using precision-recall curves, in which precision is plotted as the function of recall. From the HSV quantisation experiments a configuration in which hue was divided into 12 bins and saturation and value into 3 bins created the best precision-recall values. For the quantisation of hue, the centre of the first bin was set to  $H=0$  due to the position of red colours in the HSV hue plane. The comparison of the HSV colour histogram, RGB correlogram and HSV correlogram showed that the correlogram is more effective than the histogram in semantic category retrieval. The precision-recall of the correlogram in HSV colour space (12,3,3) was slightly better than the best RGB colour correlogram configuration (4,4,4) up to the 20% recall. Although the difference in precision is less than 5 percentage units, the quality of results was better with HSV correlograms according to visual observation. The intra-category variance of the image visual content was large; therefore, the quality of result sets with different features could vary a lot even if the retrieval precision would have been the same. The intra-category colour variance gave large performance differences between semantic categories. Paper I reported that the ‘underwater’ and ‘sunsets’ categories provided the best retrieval results, whereas ‘beach’ and ‘portrait’ were the worst. The latter categories had too large a variance in structural colour whereas the former enclosed a distinctive colour structure within the category images, a dominant aquamarine blue background texture in the ‘underwater’ images as an example.

The experiments in Paper II were conducted in the video retrieval track of the 2001 TREC Text Retrieval Conference (TREC Video Retrieval Evaluation Home Page n.d.). 11 hours of MPEG-1 video material from NIST, Open Video Project and BBC stock shot archives created the test collection, which was segmented into 7,375 shot segments approximately 5 seconds in length. Three features were computed from the shots, the HSV colour correlogram (CC), HSV colour histogram and TCC. The first two features

were computed using Eqs. 22 and 20 from the representative key frame that was selected as the first frame of a shot. The TCC was computed using Eq. 23 from the dynamically sampled shot frames. The actual test queries were executed using either example videos or an example image from the respective TREC 2001 video track topic specifications (TREC Video Retrieval Evaluation Home Page n.d.). The search tasks in the experiment contained 74 topics, such as finding footage of a lunar vehicle travelling on the moon, snow-capped mountains, water-skiers, or speakers in front of the U.S. flag. More abstract retrieval topics were included as well, such as finding scenes about environmental degradation, which was challenging for colour-based retrieval methods. In the evaluation, a match was declared if a video shot contained a part of a video segment that was annotated as relevant for that topic in the ground truth. Fig. 6 depicts the averaged search precision percentages for the computed features as a function of the result rank.

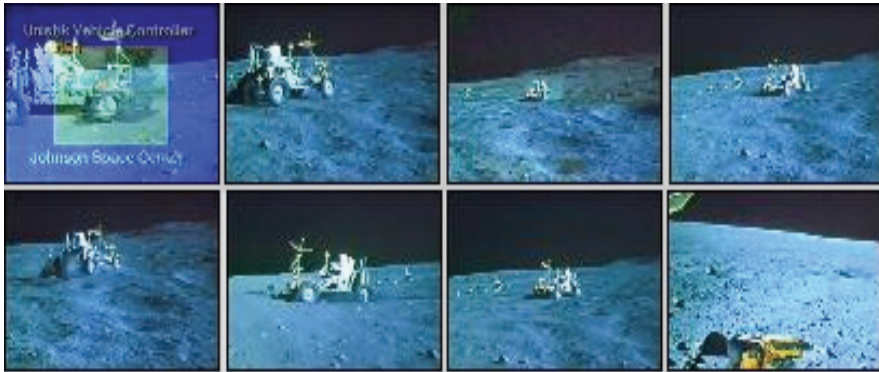


**Fig. 6. Averaged precisions of the three colour-based features in TREC 2001 video track search experiments. The curve depicts the averages for the six best performing topics. The vertical axis indicates the precision percentage as a function of result rank (Paper II).**

The performance of the TCC surpasses the other features. The HSV correlogram gives a much higher initial precision but levels with the precision of the HSV histogram at the end. Fig. 6 demonstrates that the best semantic search performance is obtained when temporal characteristics are utilised in the colour feature. The TCC also beats the static CC feature in recall. Paper II reports 0.181 (TCC) and 0.014 (CC) as the recall at 100 using soft overlapping criterion.

The effectiveness of the TCC indicates that the temporal sampling of shot frames gives robustness over single key frame selection techniques. In the experiments, the key frame was selected from the beginning of the shot close to the shot boundary, which may expose the key frame to transitional errors. More sophisticated key frame detection methods exist in the literature, such as the adaptive technique by Zhuang *et al.* (1998).

Fig. 7 depicts some exemplary results for a TCC-based query on the topic of finding lunar vehicles.



**Fig. 7. The highest ranked retrieval results for the query ‘Find lunar vehicle travelling on the moon’ using the TCC feature. Most similar shot on top left, similarity decreases row-wise from left to right. Key frames from the TREC 2001 video track test collection (Paper II).**

### 3.1.5 Summary

Paper III compared the HSV and CIE Lab colour spaces in colour change detection. The HSV colour space was found to give the highest amount of true positives at the expense of increased false positives. The CIE Lab was found to be more robust in change detection at the lower luminance levels, against reflections and in the detection of uniform colour surfaces. Additionally, unlike HSV, the CIE Lab colour space is not symmetric, which makes it difficult to quantise uniformly for the purpose of colour indexing.

The image category retrieval experiments in Paper I showed that the colour correlogram performs well compared to conventional histograms. The comparison of the HSV and RGB colour spaces with colour correlograms was favourable for the HSV, which has the advantage of separating the chromaticity and lightness information in the colour description, allowing the preference of chromaticity over luminance in the computation of colour similarities.

The temporal colour correlogram was found to improve similarity-based video retrieval performance over the key frame-based colour correlogram feature in Paper II. The TCC feature does not require an additional key frame selection technique and therefore is a more robust visual feature, albeit at an increased computational cost. However, adaptive key frame selection could be employed in the sampling of shots to reduce the computational requirements for the TCC and TGC if the selection process is computationally less expensive than the processing of 20 sampled frames.

### 3.2 Semantic Concept Descriptors

Video retrieval based on global visual similarity to a given example is effective in search topics for finding a specific setting with distinctive visual properties. However, there are more challenging search topics that treat video content in a higher abstraction layer. It is difficult to find an all-embracing visual example for topics such as finding shots depicting environmental degradation. Therefore video content analysis should result in a more semantically meaningful, symbolic presentation that will reduce the semantic gap between the raw content data and its linguistic interpretations. This section introduces semantic concept features that can be trained from the lower level video characteristics and used as a part of query definition. Papers IV and V introduce a set of techniques for detecting semantic concepts from videos and describe experiments in the framework of the TRECVID semantic feature detection task (TREC Video Retrieval Evaluation Home Page n.d.).

#### 3.2.1 Visual Concept Detectors

This section introduces a computationally lightweight method for detecting semantic concept detectors using content-based similarities. Papers IV and V describe the detection experiments for sets of visual concepts and detector configurations. An overview of the method introduced in Paper V is shown in Fig. 8. The method utilises user-selected positive examples ( $k$ ) embodying a concept  $f$ . Label confidences for database items ( $n$ ) are created from low-level dissimilarities and nearest neighbour rank lists.

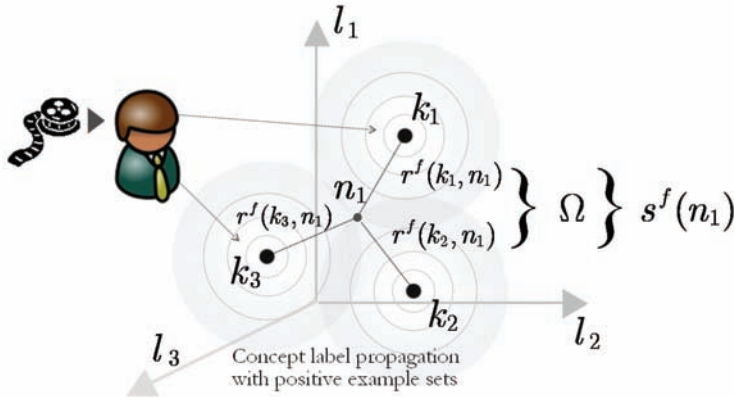


Fig. 8. An overview of the concept propagation method using positive example sets.

Paper IV introduced several techniques for detecting semantic concepts from audiovisual content. This dissertation addresses exclusively the contributions on the development of visual techniques for ‘people’, ‘cityscape’ and ‘landscape’ detection. The concept training technique introduced in Paper IV was based on self-organising maps (SOM) of low-level

features. A simple procedure was introduced that utilised SOMs in the detection of concepts from video shots. To ‘train’ the map with a certain semantic concept, sets of characteristic example shots were selected from the training video set. A total of 13 example shots were selected for ‘people’ ( $K = 13$ ), and 10 for both ‘cityscape’ and ‘landscape’ ( $K = 10$ ). TGC feature (Eq. 24) was used in the SOMs.

To create confidences for the shots in the collection for a specific concept, we used every selected example shot to locate sets of best matching units from the SOM. Therefore the procedure for detecting concepts was practically a propagation of concept label confidences to the closest neighbour nodes and samples within. From these nodes, the closest shots for each example were selected using the shortest  $L_1$  distances. The term propagation refers to the process of filling in concept confidences from the ordered list of nearest database items and their similarities to the positive examples. An increase in overall similarity results in greater concept confidence. Every distance value in a result set  $R_k$  for each example  $k$  was normalised with the largest distance in the set. Finally, the sets of results  $\{R_1, \dots, R_k\}$  were combined using a minimum operator for the final ranked list of concept confidences. In practise, the minimum operator sorts the final result set using the closest normalised distance values in the result lists  $\{R_1, \dots, R_k\}$  for each shot  $S$ .

Paper IV described the experiments using only a single TGC feature, whereas Paper V expanded to a larger set of visual features. The TGC, TCC and motion activity were used in the concept detection. Moreover, Paper V omitted SOMs from the propagation of label confidences and instead computed the nearest neighbours directly from the low-level feature spaces. A description of the technique follows. See Fig. 8 for an overview.

Initially, each detector has  $K$  positive examples for the propagation procedure. The propagation of labels and confidences is as follows: First,  $L_1$  dissimilarities to the example  $k$  in low-level feature space  $l$  results in rank ordered list  $D_l^f(k)$  where the nearest neighbour of  $k$  has the highest rank and confidence. Subsequently  $D_l^f(k)$  lists for every  $l \in 1 \dots L$  are combined using either a combination of ranks based on Borda voting (Ho *et al.* 1994) or an extended Boolean combination of dissimilarity values

$$r^f(k, n) = \Theta\left(\frac{d_1^f(k, n)}{D_{1\max}^f(k)}, \dots, \frac{d_L^f(k, n)}{D_{L\max}^f(k)}\right), \quad (25)$$

where  $r^f(k, n)$  is the overall rank or dissimilarity of a result shot  $n$  to the example  $k$  using  $L$  features ;  $d_l^f(k, n)$  is the rank or dissimilarity to the example  $k$  of concept  $f$  in feature space  $l$  ;  $d_{l\max}^f(k)$  is the largest rank or dissimilarity value to the query example  $k$  in its result set ;  $\Theta$  is the selected fusion operator: minimum rank ( $MIN$ ), aggregation of ranks ( $SUM$ ) or minimum dissimilarity ( $MINDIST$ ).

In the next phase, the result sets obtained from the independent modalities are combined into the final confidence list  $R^f(k)$  which contains an ordered list of database shots.  $R^f(k)$  is a manifestation of confidence votes for a concept  $f$  based on overall similarity to the example  $k$ . Next, the ordered lists  $\{R^f(1), \dots, R^f(K)\}$  are combined with a fusion operator  $\Omega$  to form the final confidence  $s^f(n)$  for each database item  $n$ . Finally, the top  $X$  results are submitted to the evaluation procedure. The following formulae describe the combination of confidences from the example queries

$$s^f(n) = \Omega\left(\frac{r^f(1, n)}{R_{\max}^f(1)}, \dots, \frac{r^f(K, n)}{R_{\max}^f(K)}\right), \quad (26)$$

$$S^f = \left[ \text{sort} \{s^f(1), \dots, s^f(N)\} \right]_X, \quad (27)$$

where  $s^f(n)$  is the confidence of a shot  $n$  to have concept  $f$ ;  $r^f(k, n)$  is the rank or dissimilarity of shot  $n$  to the query example  $k$  of concept  $f$ ;  $R_{\max}^f(k)$  is the maximum rank or dissimilarity to the query example  $k$  in its result set;  $S^f$  is the final ranked set of results for the concept  $f$ ; the *sort* operator creates an ordered list by the similarity value;  $X$  defines the cut-off for the top ranked relevant items in a list;  $N$  is the amount of items in the feature index;  $\Omega$  is a combination operator: minimum rank (*MIN*), aggregation of ranks (*SUM*) or minimum of distances (*MINDIST*).

The propagation of concept labels and confidences using small example sets is an efficient way to create trained confidences for the concept detection. The rationale behind the selected approach is to have several simplified concept detectors that are trained using small sets of positive example shots, each propagating labels to their nearest neighbours in selected feature spaces. The concept confidence is proportional to the measurable low-level feature dissimilarity between the example and the target.

The two significant differences between the combination techniques in Paper IV and Paper V were substituting the self-organizing maps with nearest neighbour dissimilarities and the introduction of multi-feature detection. Another difference is that the latter technique computes concept confidences to every item in the database, whereas the former operates with a subset of the database that is collected from the closest matching units in the SOM.

### 3.2.2 Experiments in Semantic Concept Detection

Paper IV describes the experiments with the semantic concept (feature) detection task in TREC 2002 video track. The video database contained 68.5 hours of videos from the 1930's to the 1960's and contained large variations in audiovisual quality, some of the material being from the narrow-film source. Two separate collections were available for training and testing, sized as 23 and 5 hours respectively. The test videos for the semantic feature detection were segmented into 1,800 shots and training database into 7,800 shots. At the evaluation, the task of the semantic feature detection was to create a list of at most 1,000 video shots ranked by the highest likelihood of finding the semantic concept present in the shot. In the evaluation, each concept was assumed binary, i.e., it is either present or absent in the shot. The visual concept categories that were experimentally evaluated in Paper IV were 'people', 'cityscape' and 'landscape'.

The effectiveness of the concept detectors was measured using the average precision (AP) measure. Table 2 shows the AP detection results for the SOM-based propagated training algorithm using a single TGC feature. The results are contrasted with median, maximum and minimum performances of the TREC 2002 semantic feature experiments.

Table 2. Concept detection AP results for the SOM-based propagated training algorithm using a TGC feature. The maximum, median and minimum performances for the TREC 2002 video track are reported as well.

Concept	Max	Median	Min	TGC
People	0.274	0.071	0.008	0.248
Cityscape	0.374	0.299	0.036	0.299
Landscape	0.198	0.190	0.128	0.193

Table 2 demonstrates that the proposed concept detection technique was suitable for the detection of ‘people’. This was due to the successful detection of material containing crowds, marches and assemblies. The characteristic properties of these types of shots had survived the passage of time better than other types that seemed to have suffered more from eroded chromaticity and sharpness in the test videos.

Paper V describes semantic concept detection experiments with a larger set of low-level features: the TCC, TGC and motion activity. The self-organising map in the propagated labelling algorithm of the Paper IV is replaced with Eqs. 25, 26 and 27. The experiments were conducted in the framework of the TRECVID 2003 semantic feature extraction task (TREC Video Retrieval Evaluation Home Page n.d.). The task consisted of returning 2,000 top ranked video clips for preset semantic features from a database of approximately 32,000 shots containing news video from ABC, CNN and C-SPAN. The following 12 semantic concept detectors were evaluated: ‘outdoors’, ‘people’, ‘building’, ‘road’, ‘vegetation’, ‘animal’, ‘car/truck/bus’, ‘aircraft’, ‘non studio setting’, ‘sporting event’, ‘weather news’ and ‘physical violence’. The training sets were kept small and consisted of a total of 217 positive examples, with the sizes of independent concept sets ranging from 7 to 26. As was done in the experiments in Paper IV, the average precisions were computed from the ranked concept detector output list.

The evaluation consisted of several detector configurations and the objective was to understand the significance of independent parameters in the detector effectiveness. The experiments evaluated different configurations of low-level features, fusion operators and the effect of feature validation on detector performance. Additionally, the effect of the reduced training set was tested with 106 examples. A single detection run consisted of detector outputs for all 12 features. Table 3 depicts the overall run performance as the run-wise mean and median of the average precisions for every feature and fusion configurations. The first row shows the performance of the validated feature configuration. Validation stands for a procedure where during the training phase different feature and operator configurations are applied and the best configurations are selected for each concept. The second row contains a fixed configuration for all features using the TCC and TGC with  $\Theta$  (feature combination) and  $\Omega$  (result set combination) configured to the *MIN* operator. The mean, median and maximum of the average precisions of 32 runs that reported results for the selected 12 semantic features in TRECVID 2003 test set are 0.132, 0.076 and 0.314, respectively.

Table 3. Detector configurations and their run-wise mean and median average precisions using 12 semantic concepts.

RunID	Used Features	$\Theta$	$\Omega$	Mean	Med
MT1	VALIDATED	VALIDATED	VALIDATED	0.090	0.047
MT2	TCC/TGC	MIN	MIN	0.075	0.056
MT3	MA/TGC/TCC	SUM	MIN	0.037	0.025
MT4	MA/TGC/TCC	MIN	MIN	0.063	0.053
MT5	TGC	-	MIN	0.043	0.043
MT_extra1	TCC	-	MIN	0.078	0.033
MT_extra2	TCC/TGC	SUM	MIN	0.066	0.030
MT_extra3	TCC/TGC	MIN	MIN (106 ex.)	0.057	0.045
MT_extra4	TCC/TGC	MIN	SUM	0.039	0.018
MT_extra5	TCC/TGC	MINDIST	MINDIST	0.036	0.032

Detector configuration of the TCC and TGC features using the *MIN* operator resulted in performance closest to the validated configuration and performed better than the *SUM* operator. Adding the motion activity feature decreased overall detection performance. The *MINDIST* operator, which uses the minimum of normalized  $L_1$  distances to combine low-level features and result lists, is performing worse than the rank-based operators, although the combination by distance preserves the dynamics of the feature space unlike the combination by ranked lists. Paper V demonstrated that although the overall average precisions were low, the precision-recall curves show much higher precision values within the first 200 shots.

### 3.2.3 Summary

Paper IV described the initial performance of the semantic concept detector technique based on concept propagation with small example sets. Above median performances were reported in the three concept types: ‘people’, ‘cityscape’, and ‘landscape’. The evaluated concepts contained emblematic textural characteristics, which contributed to the overall performance. Also, the degraded chromatic quality of the test video collections did not support concept detection with colour features.

Paper V developed the proposed concept detector technique further and conducted an evaluation using a larger set of concepts and low-level feature combinations. Overall, the rank-based combination of the TCC and TGC features using *MIN* operators gave the best non-validated performance. The combination using normalised distances did not result in good performance, indicating that the different dynamics of the independent feature spaces are not well suited to the extended Boolean combination. The number of training examples did have an effect on overall performance. When the example set was reduced to half, there was a reduction in overall performance as well. This is characteristic of the detection by propagated concept labels, the more archetypal examples selected, the better concept labels propagate to the potentially relevant areas in the feature space.

### 3.3 Experiments in Manual Video Retrieval

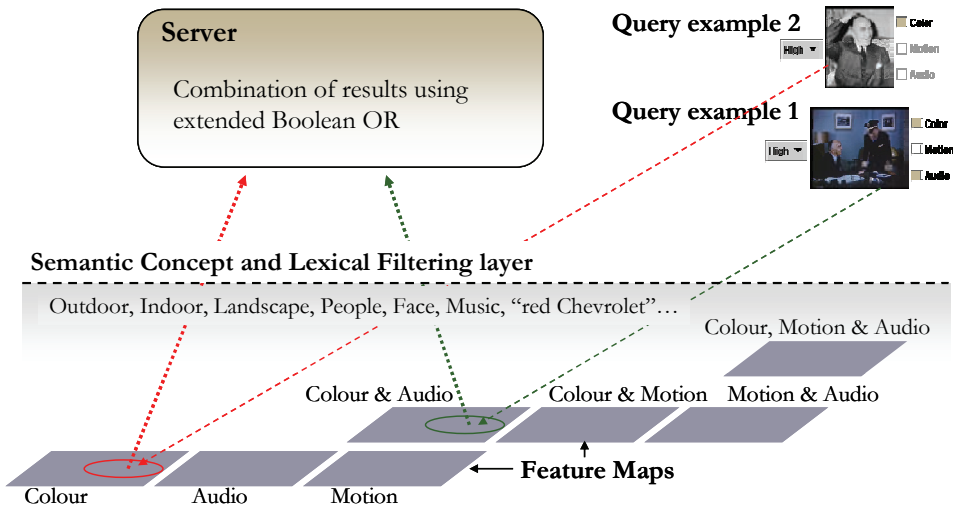
Previous sections introduced content-based features that have been developed to support the semantic retrieval of video. Papers VI, VII and X describe experiments with these techniques in the task of manual video retrieval. Manual retrieval experiments are based on the definition of the annual TRECVID retrieval evaluation. The retrieval experiments presented in this section are based on the TRECVID evaluations for the years 2002, 2003 and 2005.

A manual retrieval experiment is based on a pre-defined search topic definition, which consists of the lexical definition of the search problem and a number of image and video examples visualising the media of interest  $k \in \{1, \dots, K\}$ . The topics arrive from NIST and become the source for a test user to create a query definition  $t$ . The test user can create a lexical query  $t_{\text{txt}}$ , define a set of searchable concepts ( $c \in \{1, \dots, C\}$ ) that complement the lexical description and select any subset or the complete set of example media from the original definition ( $k \in \{1, \dots, K\}$ ) for a content-based similarity search. The completed query becomes the definition of search topic for the search system that computes similarities from its index structure and ultimately returns an ordered list of relevant video shots as the final result for the query.

The manual experiments described in Papers VI, VII and X have followed a grand theme in which the focus is on improving the overall semantic retrieval effectiveness with the combination of multiple feature types: visual, concept and lexical features.

#### 3.3.1 Experiments with Semantic Filtering

Paper VII describes the experiments with the retrieval search engine where self-organizing maps are utilised for structuring the video database with varying low-level feature configurations. The first phase in the retrieval with semantic filtering uses  $K$  image or video examples from the query definition and, based on the configuration of enabled search features, directs each to the respective SOM for retrieval based on content-based similarity. The best matching units are found from the SOMs for the feature configurations and similar shots are identified as the shots that were revealed within the best matching units. The shots retrieved from SOMs are subjected to optional semantic filtering, which determines the presence of pre-defined semantic concepts  $f$  and lexical key words ( $t_{\text{txt}}$ ) in the retrieved shots. Concept presence is expressed as a real confidence value, where 0 equals not present and 1 equals certainly present. For the purpose of binary filtering, the confidence values are thresholded to obtain a decision rule about the presence or absence of concepts. Lexical filtering is based on automatic speech recognition transcripts. The  $t_{\text{txt}}$  definition of the query contains lexical search terms that are required to exist in the speech transcripts of a result shot for it to be included in the final result set. The final result set contains only the shots that have passed the semantic filtering. Fig. 9 illustrates the semantic and lexical filtering with self-organising feature maps.



**Fig. 9. Semantic filtering with self-organising maps.**

In the experiments, the low-level index structure consisted of seven indices using self-organising maps with the TCC colour feature (see Eq. 23), motion activity (Manjunath *et al.* 2002), audio features and their combinations. An additional layer of semantic feature filters was employed to reduce semantically irrelevant matches from self-organised index maps. The semantic filters were established on lexical and concept definitions of the video content. The existence of a particular concept was expressed as a confidence value between not present (0) and certainly present (1). The confidence values were further thresholded to the binary value of presence or absence using a simple rule where one third of the shots with the highest confidences were defined as containing the concept.

Lexical filtering was based on automatic speech recognition indexes where the terms were pruned using common pre-processing operations for lexical data. Several operations were employed to relax the lexical filtering. First, query words were pre-processed and disambiguated using WordNet. Second, the filtering threshold was determined from the term frequency inverse document frequency that was temporally expanded using a heuristic four-second time frame. Shots that did not receive a score above zero were filtered out.

In the manual search experiments of the TREC 2002 retrieval evaluation a user initiated a query by indicating relevant visual examples for the topic and selected a set of fitting semantic filters. The system created a single pass search returning a list of the 100 most relevant shots from the test collection. The reported manual retrieval results for different search techniques were: content-based retrieval without semantic filtering 38/0.03, content-based retrieval with semantic concept filters 22/0.02 and content-based retrieval with lexical filters 12/0.01. The values describe the number of true hits at depth 10/mean average precision. The number of hits at depth 10 equals the number of true matches within the top 10 retrieved results. The mean average precision denotes the mean of the average precisions in 25 search topics.

The results obtained with the self-organising maps and semantic filtering demonstrate that the proposed filtering structure reduces retrieval effectiveness. This result agrees with the work of Christel *et al.* (2006), in which the overall effect of semantic filtering for image and video retrieval was measured using state-of-the-art concept detectors from the IBM T.J. Watson Research Center. Semantic filtering reduced recall at 200 with the current levels of concept detector performance. Recall improves over the non-filtering system only with extreme accuracy levels of MAP 0.7 and higher. The high accuracy requirements for the semantic filters have led to the development of a more flexible fusion model for semantic features, which is introduced in the following section.

### 3.3.2 Experiments with Visual, Semantic and Lexical Feature Combinations

Paper VI introduced an experimental retrieval system based on combinations of multiple independent search engines. Fig. 10 illustrates the operational model of the system. Principally, the engine serves three modalities: text, semantic concept and visual search. In order to process a query, a query definition  $t$  is dissected into separate feature modalities and processed by the independent search engines whose results are combined in a later stage to create the final result list.

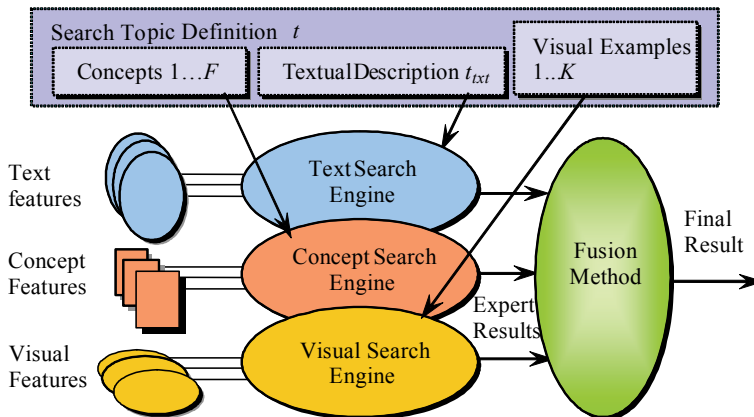


Fig. 10. Search modalities and their combination in the retrieval engine.

Item lists from the independent search modalities are combined using feature fusion at the end of the retrieval process. The model for the search engine illustrated in Fig. 10 is a design evolution from the experimental findings with a search model introduced in Section 3.3.1. The previous model consisted of self-organising multi-modal indexes and semantic filtering for the application of cluster-temporal video browsing.

The visual search module considers each query example  $k$  independently by creating queries in low-level feature spaces, namely the TCC and TGC. The ranked result list

$D_i^t(k)$  is computed using  $L_1$  distance between the vectors in respective feature space. To combine feature spaces, the ranks from the TCC and TGC search results are aggregated using Eq. 25, where  $\Theta$  is set to  $SUM$ . Finally the result lists for each visual example are combined using Eq. 26, where extended Boolean *or* operator is used by setting  $\Omega$  to  $MIN$ , and Eq. 27.

The concept search module uses 15 detected concepts and their confidence values: 'outdoors', 'news subject face', 'people', 'building', 'road', 'vegetation', 'animal', 'female speech', 'car/truck/bus', 'aircraft', 'news subject monologue', 'non studio setting', 'sporting event', 'weather news' and 'physical violence'. Query definition  $t$  contains a user-defined list of concepts with  $F$  elements assigned with Boolean values depending on whether a concept is required (1) or not required (0) in the result shots. Each concept query item  $f$  that is set to 1 retrieves a list of database shots ordered by the largest confidences for having the concept in the actual content. Finally, the overall dissimilarity between the concept query definition and a shot is computed as the aggregation of individual ranks, as described in Eq. 25 with  $\Theta$  set to  $SUM$ .

The text search module uses the automatic speech recognition (ASR) data. The words of ASR transcripts are first pre-processed with stop word removal and stemming and then indexed into the database. Words are grouped into speaker segments, which expand the contextual scope for the text query. The similarity between lexical query description  $t_{\text{txt}}$  and the ASR transcript of a shot is computed using term frequency inverse document frequency (Salton & Yang 1973). Stemming (Porter 1980) is used to remove any suffixes from the query words. The result of a text query is a rank ordered list of database items  $L^t$ . Initially,  $L^t$  contains only the results that have matched the search terms or are in the same speaker segment with the matching shot. However, in the later stage of combining results, processing the rest of the items is required. To prepare for the combination, the original text search results are normalised and database items that are not in  $L^t$  are appended to the list with ranks that do not affect the combination. Below is the algebraic description of the text search normalization procedure

$$l^t(n) = \begin{cases} \left( \left[ l^t(n) - 1 \right] \cdot \frac{N}{L_{\max}^t} \right) + 1, & \text{if } n \in L^t \\ N, & \text{if } n \in \bar{L}^t \end{cases}, \quad (28)$$

$$L^t = L^t \cup \bar{L}^t, \quad (29)$$

where  $l^t(n)$  is the rank of a database item  $n$ ;  $\bar{L}^t$  equals non-retrieved database items.

After the independent feature modules have produced their ranked shot list, separate lists are combined to create the final result list for the query definition  $t$

$$\psi^t(n) = \text{sum} \left( \frac{w_v \cdot v^t(n)}{V_{\max}^t}, \frac{w_s \cdot s^t(n)}{S_{\max}^t}, \frac{w_l \cdot l^t(n)}{L_{\max}^t} \right), \quad (30)$$

$$\Psi^t = \left[ \text{sort} \{ f^t(1), \dots, f^t(N) \} \right]_X \quad (31)$$

where  $\psi^t(n)$  is the overall rank of a result shot  $n$  to the search topic definition;  $v^t(n), s^t(n), l^t(n)$  contain the ranks of a result  $n$  in visual, concept and lexical result lists respectively;  $w_v, w_s, w_l$  are the weights of the search engines;  $V_{\max}^t, S_{\max}^t, L_{\max}^t$  are the sizes of the independent result lists;  $\Psi^t$  is the final ranked set of results for the query  $t$  and  $[\text{sort}()]_X$  denotes the selection of  $X$  top-ranked items from the sorted list.

Paper VI describes extensive search experiments with several search engine configurations for the TRECVID 2003 search data. The objectives were to measure the effect of different combination weights on semantic search effectiveness, find out the significance of query term selection in a lexical search and measure if the combined speech recognition and closed caption transcripts improve the overall search performance. Original text queries were constrained only to the lexical terms that were given in the TRECVID 2003 search topic definitions. Expansions to the text queries were created manually by a new user, who did not formulate the original manual queries. He doubled the original number of keywords using terms that were outside of the original topic description, but still relevant to the context of the topic. The 1,000 best results were evaluated for each topic.

Table 4 shows the results of the following combined search runs: (A) baseline search using only a text query; (B) visual search using a single search example from the search topic definition; (C) visual search using all search examples from the search topic definition; (G) combination of concept and text search; (H-J) combination of visual with all examples and a text search using varying weights; (K-O) combination of visual search with all examples, concept search and text using varying weights; (Q) text search using a manually expanded word list; (P) text search after speech recognition transcripts were patched with closed captions text; (R) a full run that uses all search features with combined closed captions and speech recognition transcripts and expanded lexical query.

*Table 4. Search engine configurations, combination weights for the independent feature modules and their run-wise median average precisions (MAP) with TRECVID 2003 data.*

Run Configuration	$w_l$	$w_s$	$w_v$	MAP
A. Text Baseline	1	0	0	0.098
B. Visual (single example)	0	0	1	0.005
C. Visual (all examples)	0	0	1	0.023
G. Concept + Text	1	1	0	0.079
H. Visual (all examples) + Text	1	0	1	0.114
I. Visual (all examples) + Text	2	0	1	0.116
J. Visual (all examples) + Text	1	0	2	0.103
K. Visual (all examples) + Concept + Text	1	1	1	0.101
L. Visual (all examples) + Concept + Text	6	1	3	0.123
M. Visual (all examples) + Concept + Text	6	2	3	0.119
N. Visual (all examples) + Concept + Text	8	1	4	0.122
O. Visual (all examples) + Concept + Text	6	1	1	0.109
P. Combined Text Transcripts	1	0	0	0.100
Q. Expanded Text Query	1	0	0	0.146
R. Combined Transcripts + Expanded Text Query + Concept + Visual (all ex.)	6	1	3	0.169

Paper VI reports the effect of weights on different topic groups. Common noun topics search for shots with a more generalised theme, whereas known location and person topics are focused on finding specific targets. Table 5 shows the average precisions for groups with different feature weights (A, H, J, K, L) and with combined transcripts and an expanded text query (R).

*Table 5. The effect of weighting the different search features in topic groups. Mean average precisions are reported.*

Run + Weights ( $w_l, w_s, w_v$ )	Common Noun Topics, 15 Topics	Known Location Topics, 3 Topics	Known Person Topics, 5 Topics
A. (1,0,0)	0.038	0.248	0.170
H. (1,0,1)	0.063	0.329	0.127
J. (1,0,2)	0.060	0.308	0.085
K. (1,1,1)	0.057	0.301	0.098
L. (6,1,3)	0.061	0.375	0.150
R. (6,1,3)	0.122	0.434	0.173

With properly weighted multi-modal queries, it is possible to improve the performances in common noun and known location topics whereas known person topics benefit the least from the combination of features.

Paper X describes experiments with the multi-lingual video database in the TRECVID 2005 data. The visual modules use a configuration similar to the previous description. The concept search module is extended with a more comprehensive set of concept features designed to better suit a larger topic variety than the concept sets from the previous experiments. The following concept detectors were employed: entertainment, faces, newsroom, outdoor, desert, natural-disaster, snow, fire-explosion-smoke, maps-charts, meeting-footage, nature-footage, sports, water, and weather. The lexical search module is based on automatic speech recognition and machine translation transcripts for the non-English data, both provided by NIST. Feature results were combined using Eqs. 30 and 31 and the weights  $w_v, w_s, w_l$  were set to 1, 1 and 2, respectively. The 1,000 best results were evaluated for each search topic. Table 6 shows that both visual and concept searches are able to give advanced performance over the baseline lexical search.

*Table 6. The search engine configurations and respective mean average precisions in manual retrieval experiments with multilingual TRECVID 2005 data.*

Run Configuration (Ratio of Feature Weights)	MAP
Text search baseline performance	0.081
Combination of text and concept search (2:1)	0.097
Combination of text and visual example search (2:1)	0.102

### **3.3.3 Summary**

The goal in the manual retrieval experiments was to understand how multi-modal search features should be used in order to improve semantic retrieval effectiveness. The experiments in Section 3.3.1 with self-organising maps and semantic filtering showed a reduction in search effectiveness when a content-based low-level feature search was combined with semantic filters. Section 3.3.2 described experiments with a retrieval model in which linear weights can be applied to emphasise different features in the search. Several findings can be derived from the results. First, good baseline performance can be achieved using text queries against automatic speech recognition transcripts. Second, the significance of global visual similarity search is not essential, but can be improved by defining several examples for the search. Third, by adjusting the linear weights, both concept and visual searches can improve the search effectiveness over the baseline text. By creating better concept detectors, the significance of semantic features can be further increased. Using weighted multiple feature configurations, the mean average precision in common noun and known location topics is multiplied approximately by 1.5 whereas the known item topic performance does not improve. Overall, a 2:1 weight ratio between lexical and content features respectively appears to be a good estimate for the search engine. When content-based features become more descriptive and concept vocabulary increases, this ratio approaches 1:1. Finally, enriching the definition of the text query has a greater impact on search performance than reducing errors in the speech transcripts using external information, such as closed captions text.

The experiments proved that the content-based features can improve semantic retrieval performance over conventional techniques, namely text search from speech transcripts. The results revealed design guidelines for the search engine, which lays down the foundation for effective interactive search.

## **4 Interactive Video Retrieval**

The process in which a user with an information need interacts with the search system by creating requests, receiving responses as the relevant results and using navigational tools to approach data from various angles, is identified as interactive retrieval. In contrast to manual retrieval, interactive retrieval enables multifaceted interaction between the user and the search system. In its simplest form, interactive retrieval is sequences of queries in which the user redefines query parameters based on the previously retrieved results. Since the search process is a complex chain of communication between the user and the system, search effectiveness is highly dependent on the user's capability to utilise system characteristics in creating task-related interaction steps and to make the system converge towards relevant results.

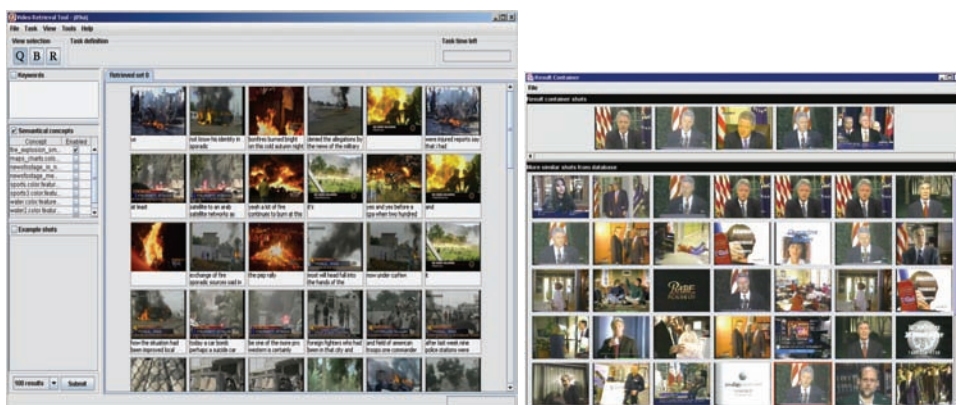
This chapter introduces content-based interaction techniques for the prototype video retrieval system and reports the results of the experimental evaluation in the interactive search task of the TRECVID video retrieval evaluations. Section 4.1 presents an interface for creating video queries manually and an interface for communicating relevance information between the user and the system. Section 4.2 introduces cluster-temporal browsing, a novel technique for content-based video database access. Section 4.3 reports experimental results with the different search interfaces. Statistical validation for the main results is also provided. Finally, Section 4.4 presents the obtained search results in the light of annual TRECVID performance scores.

### **4.1 Sequential Queries and Relevance Feedback**

The most straightforward way to add interactivity to a search is with sequential queries. The results retrieved by a search system give a user a better insight on how to set the query parameters to better suit her search task. This iterative revision of queries based on system response will eventually make the search converge towards a more relevant set of results. The manual refinement of query parameters during the iteration can be circumvented using automated relevance feedback methods (Rocchio 1971, Salton & Buckley 1990). Papers IX and X introduced interfaces for creating video queries and providing relevance feedback. The search interface enables the manual definition of a

video query. The user can select examples for a visual search, configure a semantic query from the list of given semantic concepts and create a text query by typing words to a text box. The search interface delivers the query definition to the search engine, which returns a ranked list of relevant shots back to the interface. At any time, the user can select relevant shots from the list and modify and submit new queries to the system.

The selected relevant shots are transferred into the result container interface. The relevant items are regarded as positive feedback from the user and transformed into a new example query, which is directed to visual and text search engines. The relevance feedback query follows the description of Section 3.3.2 where the creation of a query with  $K$  positive examples in visual and lexical search modalities is explained. Each time the set of positive examples is modified, the feedback query is regenerated and updated results are displayed under the selected shots. The retrieved results present an additional resource to assist in locating more relevant shots from the database. Fig. 11 depicts the search and relevance feedback interfaces.



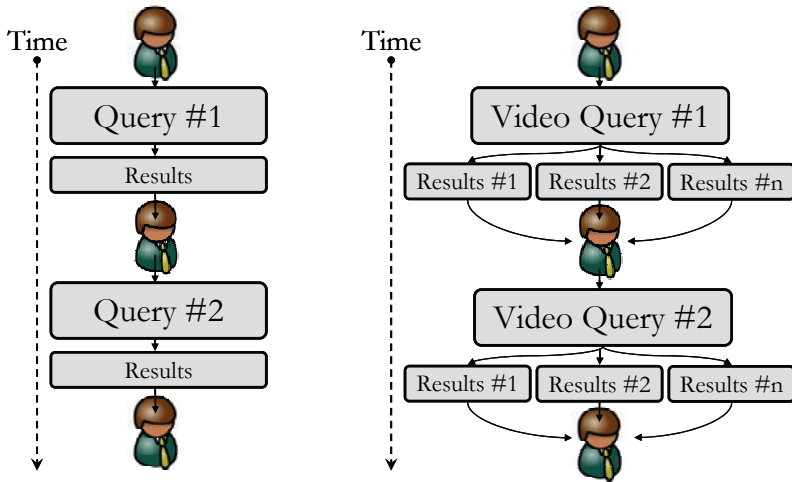
**Fig. 11. Search interface (left) enables the creation of manual queries. The result container interface with positive relevance feedback (right).**

## 4.2 Cluster-temporal Browsing

Content-based search is believed to retrieve relevant content from videos. However it can not guarantee complete relevance due to the semantic gap between the actual search need and the representation of data in the system. In order to allow the user the possibility to revise the query, interactive techniques are needed. The process of iterative queries is essentially a sequential process in which two delays have significance on the overall search efficiency and effectiveness. The first delay is the time expended on creating and refining the query parameters. The second delay is the time used in waiting and inspecting the ambiguities in the response from the system. The former delay can be reduced with relevance feedback techniques, where query refinement is adjusted according to user relevance selections. The latter delay can not be much shortened

computationally, but the time consumed can be made more useful by providing efficient navigational techniques, where the user is given competent tools to acquire different views on data (Heesch & R ger 2004a, Yeung *et al.* 1995, Zhang *et al.* 1997, Cox 1992, Campbell & van Rijsbergen 1996).

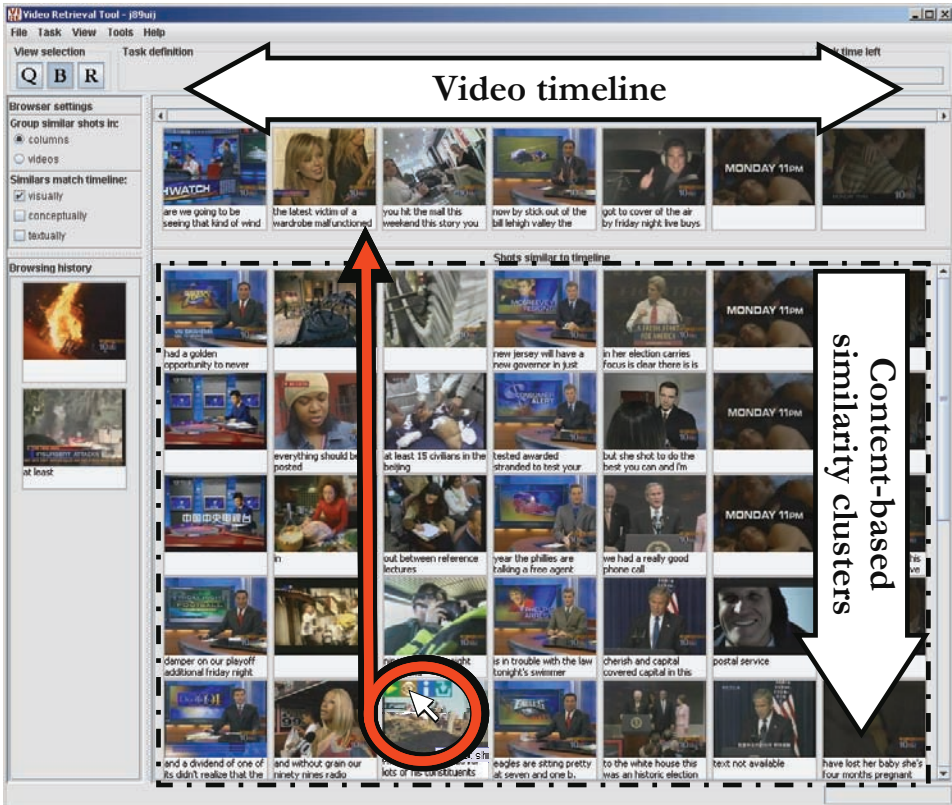
Papers VII, VIII, IX, X and XI describe a novel browsing technique for video data in which temporally adjacent shots of a video are used to concurrently retrieve and display a number of nearest neighbours from the content-based feature clusters. The objective of simultaneous retrieval of multiple feature clusters is to overcome the uncertainty of content-based retrieval by offering multiple views on the feature clusters. Temporally adjacent shots in produced video content often have the same contextual setting. Therefore, sequential shots work as an ample source of example content for a common semantic theme. Cluster-temporal browsing helps the user to navigate through the vast search space towards relevant objects. Cluster-temporal browser operates unobtrusively as it does not enforce users to learn system-specific query syntax, whereas typical content-based search systems make users dependent on the successful query parameterisation. The name cluster-temporal browsing implies that the content-based feature clusters and video timelines are integrated into an application view and present two degrees of freedom for navigating in the database.



**Fig. 12. Interaction process for sequential queries (left) and cluster-temporal browsing (right). Cluster-temporal browsing gives more results to the user during each iteration.**

Fig. 13 illustrates the organisation of cluster-temporal browsing. The browser view presents two alignments for the video shots. Horizontally, a portion of the video timeline is shown at the top row. Each shot on the timeline is a source for the shots in vertical columns. The contents of the vertical columns are acquired from the results of content-based queries in which the timeline shots act as a query example. The dotted rectangle shows the similarity view, where multiple concurrent queries have contributed to the contents within. The user is given the possibility to select simple parameters by which the similarity view is generated. The configuration can be selected from any or a combination

of visual, conceptual and lexical similarity. A parent video of any visible shot can be expanded to the timeline row, which is illustrated with the red circle and arrow.



**Fig. 13. The structure of cluster-temporal browsing.** The top row displays the shots from the selected video timeline. The dotted rectangle portrays the area with the nearest neighbour results from parallel similarity queries. Any video shot can be expanded to the video timeline using the superimposed buttons.

By overlaying the mouse cursor on a key frame, a set of quick buttons are superimposed on the shot key frames. The buttons launch the following actions: initiate shot playback, open parent video in the browser, display additional information and append the shot to the result container's list of relevant items. The text below the key frames shows automatically recognised spoken content in the shot. The browser application keeps track of the browsed shot history to enable reversing the browsing process. Any change in the top row's configuration of shots causes the similarity view to update itself using the new query examples.

### 4.3 Interactive Retrieval Experiments

The interactive retrieval task in the TRECVID retrieval benchmark follows the same specification as the manual retrieval. The search topic definitions as well as the test and development collections were equivalent to those in the manual retrieval; result pools were also unified. The details of the 2002, 2003, 2004 and 2005 experimental settings are described in Section 2.8 . Starting from TRECVID 2003, the time for interactive experiments was limited to 15 minutes per topic.

The procedure for the experiments consisted of test users searching for video shots that were perceived as relevant to a given topic definition. A typical interactive search configuration was a pair-wise test between two system configurations with four simultaneous test users. The test users were non-English speaking undergraduate or graduate students, having experience with the proposed search system ranging from novice to system designer. At the beginning of the experiments, each test participant received 30 minutes of training with the search system. A total of 12 topics were assigned for each participant and search time was limited to 10-12 minutes for searching each topic. The first six of the topics were carried out with one system configuration, which was followed by a break with refreshments to reduce the effects of fatigue. The final half of the topics was completed with another system configuration after the break. In order to eliminate random proficiency for any topic types, each user/set of topics was organised into a Latin Square configuration as shown in Table 7.

*Table 7. A Latin Square configuration for a group of four test users in interactive experiments.*

Run ID	Searcher ID [Topic Set IDs]			
System A	S1[TG1]	S3[TG2]	S2[TG3]	S4[TG4]
System B	S2[TG1]	S4[TG2]	S1[TG3]	S3[TG4]

At the end of the experiments, the selected relevant results for each topic were selected as the source for a content-based query, which expanded the final result list to the maximum allowed by the TRECVID evaluation rules.

Paper VII described the first interactive experiments with the cluster-temporal browser using self-organising maps and semantic filtering techniques and contrasted the interactive search with manual retrieval results. Paper VIII gave semantic search experiments with two cluster-temporal browser configurations: browsing with visual features only and with combined visual and lexical features. The search engine was changed from that of Paper VII to the independent feature combination with rank-based fusion (see Section 3.3 for details). Paper IX described more experiments with visual vs. visual and lexical browsing though with two different user groups: expert and novice users. Experiments with principal search strategies were also reported: A conventional content-based search with relevance feedback was evaluated against a search system augmented with cluster-temporal browsing. Paper X reported evaluations with two user groups: novice users and system designers as users. The user groups evaluated the system with and without cluster-temporal browsing. The objective of the experiments in Paper IX and X was to find out if the conventional search model benefits from content-based

browsing techniques in semantic video retrieval tasks. Paper XI reported statistical significance tests on the browser experiments.

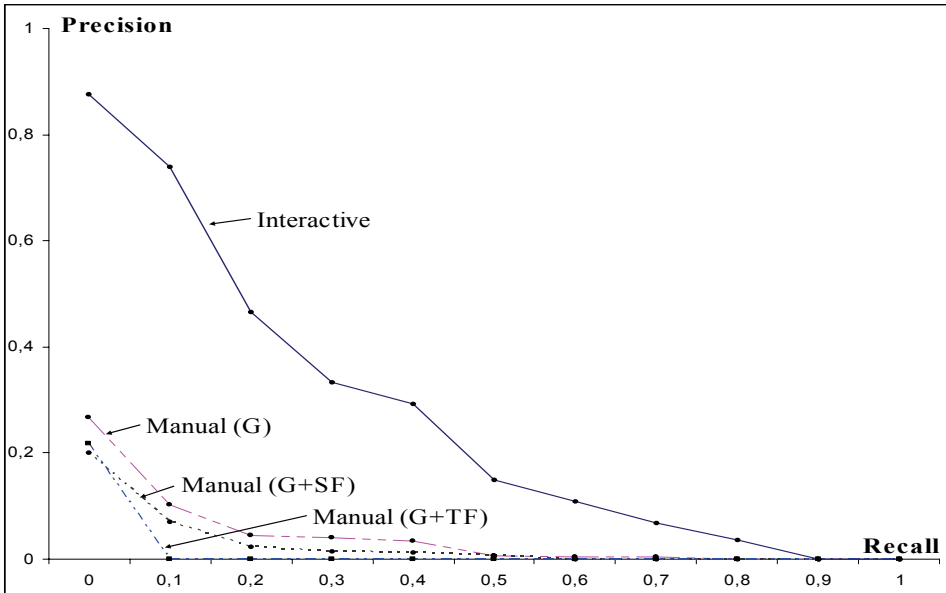
### ***4.3.1 Experiments with Cluster-temporal Browsing***

Paper VII describes the first semantic retrieval experiments with cluster-temporal browsing on TREC 2002 video track data. A group of eight users carried out the interactive search. The test users, most of them males, were novice information engineering students with good skills in using computers but less experience in searching video databases. Every user reported being somewhat familiar with the 25 search topics. The topics were divided into four sets that were randomly given to the test users so that two users carried out the same set of topics. Fig. 14 depicts the precision-recall curves for the manual and interactive search experiments. The manual system configurations were based on (G) low-level features colour, motion and audio, (G+SF) low-level features with semantic filters and (G+TF) low-level features with lexical filtering. See Section 3.3.1 for the description on semantic filtering with self-organising maps. The interactive system provided searching and cluster-temporal browsing and a ten-minute time frame was given per search topic. Manual retrieval experiments were performed by a different test user.

The precision-recall curve in Fig. 14 illustrates that the difference between the interactive and manual search paradigm is major significant. This indicates that stressing the interactive techniques is important for boosting overall semantic search effectiveness.

Papers IX and X report evaluations on the cluster-temporal browsing with the traditional content-based search paradigm: sequential queries with relevance feedback. In Paper IX, the experiments were carried out by a group of five test users in a 70-hour database of U.S. news and commercials from TRECVID 2004. The test users were novices with the search system. They had good skills in using computers yet little experience in searching video databases. None of the test participants was a native English speaker. All of the participants were used to searching the web.

Paper X evaluated the cluster-temporal browser on a database of 80 hours of English, Arabic and Chinese video data from TRECVID 2005. The interactive experiments were carried out with a group of eight test users: four were novices with the system and the other four were involved in the retrieval system development but had not seen the test search topics or any content from the test database. The novice users were mainly information engineering undergraduate students with good skills in using computers and searching the web but with little experience in video database searching.



**Fig. 14. Precision-recall curves for manual video retrieval performance vs. interactive retrieval using cluster-temporal browsing and sequential queries.**

Two system variants were tested. With the variant IxQ users were only allowed to use a search application for sequential queries, where they had to manually define search parameters to generate results for a query. In addition to the search application, users were able to use the result container's relevance feedback view. The cluster-temporal browser was not available. The second variant, IxB, allowed users to utilise an additional cluster-temporal browser during the search. The search time was limited to 12 minutes. During that time users were supposed to do their best to use the available system configuration to locate results for the given search task. Table 8 shows the search results.

*Table 8. Mean average precisions and relevant shots returned for sequential searching with relevance feedback (IxQ) and the same configuration augmented with cluster-temporal browsing (IxB). Experiments are from the years 2004 and 2005.*

Search Run ID	MAP	# Relevant Returned
2004 (novice users): 11Q/12B	0.165/0.202	650/681
2005 (novice users): 13Q/14B	0.202/0.226	1907/1998
2005 (expert users): 11Q/12B	0.264/0.242	2284/1916

The results in Table 8 show that the cluster-temporal browsing improves search performance for novice users over conventional search paradigm. Expert users did not benefit from the browser. Section 4.3.3 validates the obtained results with statistical tests.

### 4.3.2 Experiments with Browser Configurations

Papers VIII and IX evaluate different browser configurations for cluster-temporal browsing. Paper VIII tests two different system variants in order to find out whether the combination of lexical and visual browsing cues improves the search efficiency over browsing by visual features only. System variants IxV excluded the lexical browsing feature and the visualisation of speech recognition text so that searching was based entirely on visual cues. System variants IxVT included lexical searching and speech transcript visualisation of shots (see Fig. 13). The interactive experiment was carried out by a group of eight novice test users in the TRECVID 2003 evaluation. The test users, two of them females, were mainly information engineering undergraduate students with good computers skills but little experience in searching video databases.

Paper IX describes interactive experiments in the TRECVID 2004 setting with two cluster-temporal browser configurations: browsing using only lexical similarity (configurations IxT) vs. browsing with combined visual and lexical features (configurations IxVT). The experiment was carried out by a group of 12 test users, from which four users had prior experience in searching with the proposed system and the rest being novice level users.

Table 9 shows the results for different configuration pairs in the years 2003 and 2004. Mean average precisions and the number of retrieved relevant shots are reported. It demonstrates that visual features were defeated by combined visual and lexical configurations, but the best performance was obtained using a single lexical feature.

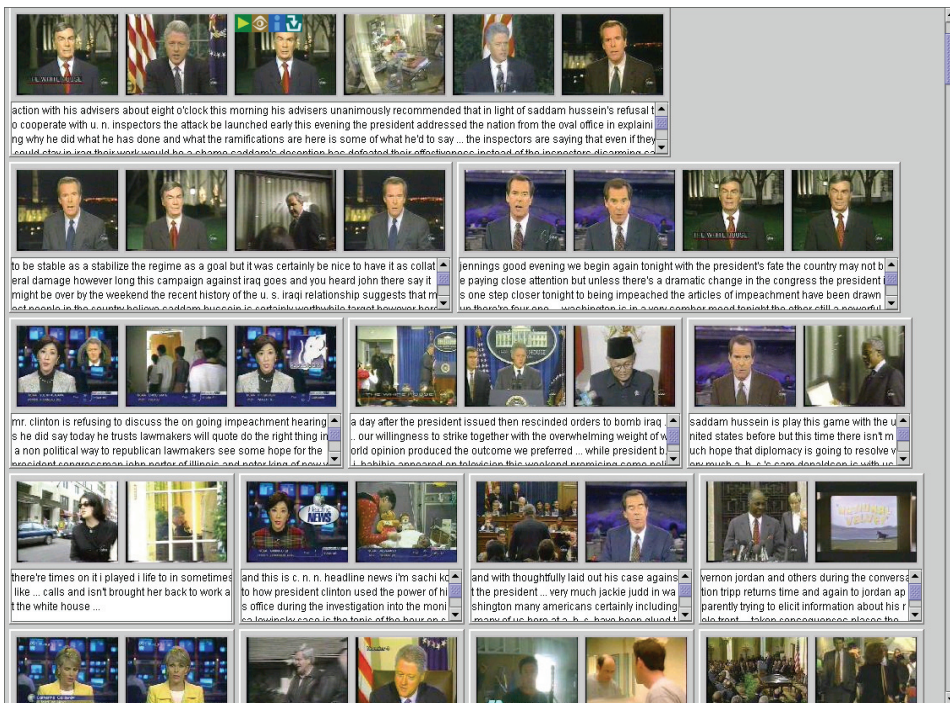
Table 9. TRECVID 2003 and 2004 experiments with different cluster-temporal browser configurations: visual vs. combined visual and lexical features.

Search Run ID	MAP	Number of Relevant Returned
2003 (novice users): I1V/I2VT	0.172/0.241	623/633
2003 (novice users): I3V/I4VT	0.156/0.207	565/679
2004 (novice users): I1T/I2VT	0.210/0.179	726/678
2004 (novice users): I5T/I6VT	0.212/0.201	723/721
2004 (experienced users): I3T/I4VT	0.212/0.212	767/776

The results of Table 9 show that low-level visual features do not contribute to overall browser performance. However, a closer examination of topic performances by their sub-category shows that two categories have better performance with visual features enabled: finding generic scenes and actions. For example, topics of finding ‘one or more buildings with flood waters around it/them’, ‘a person hitting a golf ball that then goes into the hole’ and ‘one or more buildings on fire, with flames and smoke visible’ have performed better with the browser configuration in which both text and visual features are used. Also, Christel *et al.* (2006) concluded in their study that the current state-of-the-art concept detectors show immediate utility in video browsing by delivering double the performance of not using any concept detectors for browsing. Therefore, future work for cluster-temporal browser entails the integration of semantic concepts with lexical and visual browser features.

Paper IX also presents an alternative order for the similarity view (See Fig. 15). The view groups result shots according to their originating video file. The largest group of results from the same video file receives the highest rank. The ranked shot groups are displayed row-wise starting from left to right and from top to down. Each shot group shows a selection of key frames and speech transcripts. The grouping of results is assumed to aid in identifying the contextual setting of the results and to return structured information to the users of the browser.

During the experiment, users had the possibility to switch between the original (Fig. 13) and the proposed video-based grouped view, and the times of use for each were recorded. The default view was set to the grouped view. In total, the grouped video view (Fig. 15) was used 436 minutes whereas the original view was used 1,069 minutes during the interactive experiments. The elapsed times show that the user preference was towards ungrouped similarity view.



**Fig. 15. Alternative similarity view for the cluster-temporal browser. Shots are grouped as videos to create skimmed presentations of matching video clips.**

### 4.3.3 Statistical Validation of Cluster-temporal Browsing Experiments

The TRECVID retrieval evaluations used 24 or 25 semantic search topics on large collections of videos. The tables in previous sections show the mean of the average precisions obtained in the semantic retrieval experiments. An increase in mean average precision indicates that the precision and recall for the test system configuration are higher on average, which may be construed as improvement in semantic retrieval effectiveness. The remaining question is how large the improvement in mean average precision has to become to be statistically significant. In his work, Zobel (1998) studied the behaviour of large-scale information retrieval experiments and recommended the Wilcoxon signed rank test as the preferred statistical test to evaluate performance differences between two search system configurations.

Paper XI studied the significance of the reported results from the 2004 and 2005 experiments using statistical validation tests. In order to evaluate the standing of the cluster-temporal browsing against conventional search and relevance feedback tools, Table 10 describes the overall contributions of different search interfaces for the user-collected relevant shots during the search experiments, both from novice and expert users. When the browser is not available, most results are selected from the search interface but as many as three shots out of ten are collected from the relevance feedback tool. When cluster-temporal browser is available, half of the shots are being selected from the browser view.

*Table 10. Distribution of the collected shots by the types of search interfaces in TRECVID 2005 experiments.*

Experimental Setup	Search Interface	Browser	Relevance Feedback
With Browser	29%	52%	19%
Without Browser	68%	-	32%

Paper XI describes pair-wise statistical tests to learn if the improvement in novice user search performance with the cluster-temporal browser is significantly different. The non-parametric Wilcoxon signed-rank test is used to validate the observation. An alternate hypothesis is that the average precision differences with the cluster-temporal browser are statistically significant from the baseline configuration that consists of search interface with relevance feedback. Table 11 shows the results of the Wilcoxon tests with normalised and non-normalised average precisions. The first test uses non-normalised average precisions directly from the original NIST results (values  $W, z, p$ ) whereas in the second test topic-wise average precisions were normalised using the maximum average precisions across all the results evaluated by NIST for that query (values  $W_n, z_n, p_n$ ). The normalisation of average precisions removes the variance in difficulty between different topics and results in better statistical robustness.

The first column in Table 11 reports the results of statistical tests with 23 search topics from the 2004 set-up, where one topic was removed from the official evaluation. The second column is computed from the results of 24 topics in the 2005 experiments. The third column is a combination of novice user tests from the 2004 and 2005 experiments totalling in 47 search topics. The fourth column portrays the experiment with system

developers. The overall test was computed as the combination of the 2004 and 2005 results. Normalisation was based on the maximum performance of the respective year's experiments. The first line of Table 11 shows the Wilcoxon signed rank differences for the configurations without and with the browser. Negative values indicate that the pairs are in favour of the cluster-temporal browser whereas the positive values favour the conventional search paradigm. The last two rows display the most interesting information; the probability that the null hypothesis will not be rejected for the pairs in that set.

*Table 11. Wilcoxon tests for TRECVID experiments with novice and expert users.*

Wilcoxon Test	2004 (Novices)	2005 (Novices)	2004/2005 (Novices)	2005 (Experts)
$W$	-134	-46	-358	106
$W_n$	-144	-46	-360	104
$n$	23	24	47	24
$z$	-2.03	-0.65	-1.89	1.51
$z_n$	-2.18	-0.65	-1.9	1.48
$p$	0.0212	0.2578	0.0294	0.0655
$p_n$	0.0146	0.2578	0.0287	0.0694

The following conclusions can be drawn from the reported statistical tests:

1. The 2004 experiments with novice users demonstrate improvement in search efficiency using the cluster-temporal browser. The results are statistically significant within a margin of error of 2.5%;
2. For the 2005 experiments, the null hypothesis cannot be rejected within a margin of error of 5%. The browser configuration does not provide statistically significant benefits;
3. The overall combined experiments of 2004 and 2005 with 47 semantic query topics show that the cluster-temporal browser improves the semantic search for novice users. This result is statistically significant within a margin of error of 5%;
4. The 2005 experiments with expert users favour a non-browser configuration. Statistical tests show that the 5% margin of error is not reached therefore results are not statistically significant.

The most important result is the third observation. It concludes that the cluster-temporal browser improves novice users' overall search efficiency by a statistically significant proportion. The results from the system designer user group are the opposite, but statistically not significant. Also, Paper XI prompts evidence that experimenting with system designers has lower internal validity than experimenting with novice users.

#### **4.3.4 Summary**

Interactive techniques yield significantly better search effectiveness than a manual search. The browser experiments show that the best overall performance is obtained using lexical features. However, some topics involving finding generic scenes and actions benefit from

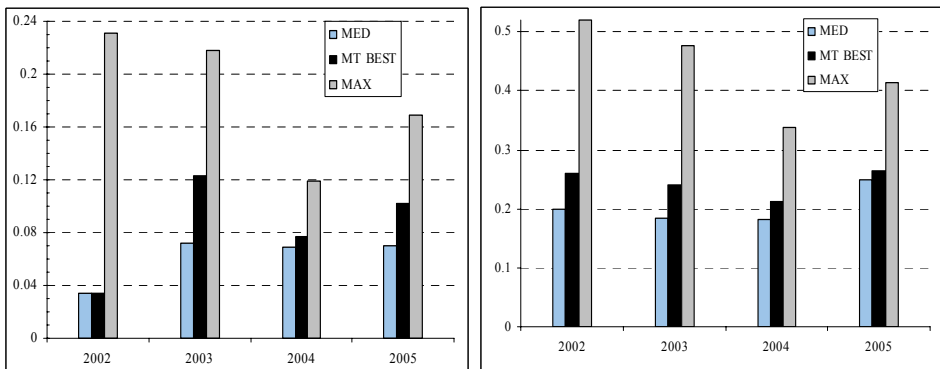
the combined visual and lexical browsing over the bare lexical configuration. It is also possible that a video does not contain text transcripts at all, in which case the significance of visual and/or auditory features increases. The combination of visual and lexical search generally performs better than using visual similarity only.

The grouping of results by video clips was not favoured by the test users. The results from the manual retrieval indicate that it is possible to improve performance by combining visual, concept and lexical features. Future work involves integrating semantic concept features with the lexical and visual features in the cluster-temporal browser.

The distribution of the relevant shots between search interfaces (Table 10) indicates that the cluster-temporal browser has been equally favoured with the conventional search paradigm and relevance feedback. Statistical tests with the results show that novice users obtain better search effectiveness with the browser than without it.

#### 4.4 Overview of the Results in TRECVID Video Retrieval Benchmarks

An international video retrieval benchmark provides unified data and search experiments to the participating teams. This makes it possible to examine the effectiveness of the proposed techniques on a global scale. Fig. 16 summarises the average precision results from participating groups for each evaluation year.



**Fig. 16.** Average precisions of the best runs from the proposed search system (MT BEST) in each annual TRECVID evaluation are displayed with the maximum and median scores. On the left are manual retrieval experiments. Interactive retrieval experiments are on the right.

The figure shows the yearly maximum and median performances in the TRECVID experiments and contrasts them with the techniques proposed in this thesis. The diagrams show that the proposed techniques stay at or above the median in every evaluation.

The diagrams demonstrate that the interactive search achieves an average precision that is more than two times higher than that of the manual search. It is worth noting that differences between participating user groups around the world, for example in their

English language skills or specific skills with a search system, affect retrieval performance and make it harder to contrast different interactive systems. The yearly fluctuations in the maximum value are caused by varying test collections, search topic sets and participant group lists. Consequently the advancement of video search systems is not obvious in the diagrams, albeit the results of 2004 and 2005 indicate a small progressive trend.

## 5 Summary

With the surge of digital video comes the challenge of data management. While manual annotation can not keep up with the accumulation of video data, automatic content-based techniques concentrate on alleviating the problem of semantic access. This thesis addressed the problem by focusing on the following sub-problems.

What is the significance of low-level visual descriptors in semantic retrieval effectiveness? The colour correlogram, temporal colour correlogram and temporal gradient correlogram use perceptually oriented colours and local contrasts in image brightness information independently, which is similar to the behaviour of the human visual system. The features introduced here incorporated temporal changes in colour and brightness for improved retrieval effectiveness. The results showed that low-level visual features improved semantic search effectiveness when combined with concept and lexical features using an appropriate weighting scheme. Low-level visual features were effective in queries where a salient visual structure is predominant.

How should the computational low-level content descriptors be connected with higher-level concepts to create more semantic indexes for retrieval? A fast training method using label propagation with small example sets and visual features was introduced as a concept detection technique. The rank-based combination of temporal colour and gradient correlograms resulted in performance that is competitive with manual feature validation.

How to combine search modalities in a search system? The linear aggregation of ranks was found to be effective in combining lexical, concept and visual features for multi-modal video retrieval. By giving the greatest weight to lexical features, the combination of features from multiple modalities surpasses the performance of the conventional video retrieval technique: text search on speech transcripts.

What types of user interface techniques can aid the user towards semantically meaningful search results efficiently? A novel content-based technique for interactive browsing of video shots was introduced in this thesis. Cluster-temporal browsing was found to improve the semantic search performance over sequential queries and relevance feedback by a statistically significant proportion. Future work involves evaluating weighting techniques to improve cluster-temporal browsing with multi-modal features.

## References

- Abe S, Tonomura Y & Kasahara H (1989) Scene retrieval method for video database applications using temporal condition changes. Proc. International Workshop on Industrial Applications of Machine Intelligence and Vision, Tokyo, Japan, 355-359.
- Adcock J, Cooper M, Girgensohn A & Wilcox L (2005) Interactive Video Search Using Multilevel Indexing. Proc. International Conf. on Image and Video Retrieval (CIVR 2005), LNCS 3568, 205-214.
- Ahonen T, Hadid A & Pietikäinen M (2006) Face description with local binary patterns: application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12): 2037-2041.
- Al-Khatib W, Day YF, Ghafoor A & Berra PB (1999) Semantic modeling and knowledge representation in multimedia databases. IEEE Transactions on Knowledge and Data Engineering 11(1): 64-80.
- Allen JF (1983) Maintaining knowledge about temporal intervals. Communications of the ACM 26(11): 832-843.
- Amir A, Iyengar G., Lin C-Y, Naphade MR, Natsev A, Neti C, Nock HJ, Smith JR & Tseng B (2004) Multimodal video search techniques: late fusion of speech-based retrieval and visual content-based retrieval. Proc. of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 3: 1048-1051.
- Androutsos D, Ling Guan & Venetsanopoulos AN (eds) (2006) Semantic retrieval of multimedia [from the Guest Editors]. IEEE Signal Processing Magazine 23(2): 14-16.
- Bach JR, Fuller C, Gupta A, Hampapur A, Horowitz B, Humphrey R, Jain R & Shu CF (1996). The Virage image search engine: An open framework for image management. In: Sethi IK & Jain RJ (eds) Proc. of SPIE Storage and Retrieval for Image and Video Databases IV, 2670: 76-87.
- Baeza-Yates RA & Ribeiro-Neto B (eds) (1999) Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, United States.
- Belew RK (1989) Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents. Proc. of the 12<sup>th</sup> ACM SIGIR International Conference on Research and Development in Information Retrieval, Cambridge, MA, June 25-28, 11-20.
- Blaser A (ed) (1979) Data base techniques for pictorial applications. Lecture Notes in Computer Science, vol. 81, Springer Verlag GmbH.
- Brunelli R, Mich O & Modena CM (1999) A survey on the automatic indexing of video data. Journal of Visual Communication and Image Representation 10(2): 78-112.
- Bush V (1945) As we may think. The Atlantic Monthly 176(1): 101-108.

- Campbell I & van Rijsbergen K (1996) The ostensive model of developing information needs. Proc. of CoLIS 2, Danmarks Biblioteksskole, Copenhagen, 251-268.
- Chang NS & Fu KS (1980) Query-by-Pictorial-Example. IEEE Transactions on Software Engineering 6(6): 519-524.
- Chang SF, Chen W, Meng HJ, Sundaram H & Zhong D (1997) VideoQ: An automated content based video search system using visual cues. Proc. of the fifth ACM International Conference on Multimedia, ACM Press, Seattle, Washington, United States, 313-324.
- Chang SF, Chen W & Sundaram H (1998) Semantic visual templates – linking features to semantics. Proc. of IEEE International Conference on Image Processing, 3: 531-535.
- Chang SK, Shi QY & Yan CW (1987) Iconic indexing by 2-D strings. IEEE Transactions on Pattern Analysis and Machine Intelligence, 9(3): 413-428.
- Chang SK & Hsu A (1992) Image information systems: Where do we go from here? IEEE Transactions on Knowledge and Data Engineering 4(5): 431-442.
- Chen Y, Jia Li & Wang JZ (2004) Machine learning and statistical modeling approaches to image retrieval. Springer-Kluwer, 1st edition.
- Christel M & Conescu R (2005) Addressing the challenge of visual information access from digital image and video libraries. Proc. ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, CO, United States, 69-78.
- Christel M & Hauptmann A (2005) The use and utility of high-level semantic features. Proc. International Conference on Image and Video Retrieval (CIVR), in Lecture Notes in Computer Science 3568, Singapore, 134-144.
- Christel M, Moraveji N & Huang C (2004). Evaluating content-based filters for image and video retrieval. Proc. of ACM SIGIR '04, Sheffield, South Yorkshire, UK, 590-591.
- Christel M, Tesic J, Natsev A & Naphade M (2006) Assessing the Filtering and Browsing Utility of Automatic Semantic Concepts for Multimedia Retrieval. Proc. of International Workshop on Semantic Learning applications in Multimedia (SLAM), in association with IEEE Computer Society Conference on Computer Vision (CVPR'06), New York, NY, United States.
- Cox K (1992) Information retrieval by browsing. Proc. of the 5th International Conference on New Information Technology, Hong Kong.
- Croft WB (1993) Knowledge-based and statistical approaches to text retrieval. IEEE Expert, 8(2): 8-12.
- Croft W & Lafferty J (eds.) (2003) Language Modeling for Information Retrieval. Kluwer.
- Datta R, Jia Li & Wang JZ (2005) Content-based image retrieval - approaches and trends of the new age. Proc. of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval MIR, Hilton, Singapore, 253-262.
- Daubechies I (1990) The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory 36(9): 961-1005.
- Day YF, Dagtas S, Iino M, Khokhar A & Ghafoor A (1995) Object-oriented conceptual modeling of video data. Proc. of the 11th International Conference on Data Engineering, Taipei, Taiwan, 401-408.
- Deerwester S, Dumais ST, Furnas GW, Landauer TK & Harshman RA (1990) Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6): 391-407.
- Del Bimbo A (1999) Visual information retrieval, Morgan Kaufmann Publishers, Inc.
- Del Bimbo A (2000) Expressive semantics for automatic annotation and retrieval of video streams. Proc. IEEE International Conference on Multimedia and Expo, 2: 671-674.
- de Vries AP, Westerveld T & Ianeva TI (2004) Combining multiple representations on the TRECVID search task [video retrieval system]. Proc. of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing 3: 1052-1055.
- Dimitrova N., Zhang HJ, Shahraray B, Sezan I, Huang T & Zakhora A (2002) Applications of video content analysis and retrieval. IEEE Multimedia 9(3): 42-55.

- Domingos P & Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29: 103-130.
- Duda RO, Hart PE & Stork DG (2001) *Pattern Classification*. Wiley, New York, United States.
- Dunckley L (2003) *Multimedia Databases: An Object-Relational Approach*. Addison-Wesley.
- Eakins JP (2002). Towards intelligent image retrieval. *Pattern Recognition* 35(1): 3-14.
- Eidenberger H (2003) Distance Measures for MPEG-7-based Retrieval. *Proc. of ACM Multimedia Information Retrieval Workshop, ACM Multimedia Conference Proceedings, Berkeley, United States*, 130-137.
- Faloutsos C, Barber R, Flickner M, Hafner J, Niblack W, Petkovic D & Equitz W (1994) Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3(3/4): 231-262.
- Feng D, Siu WC & Zhang HJ (eds) (2003) *Multimedia information retrieval and management: Technological fundamentals and applications (Signals and Communication Technology)*. Springer, 1st edition.
- Flickner M, Sawhney H, Niblack W, Ashley J, Huang Q, Dom B, Gorkhani M, Hafner J, Lee D, Petkovic D, Steele D & Yanker P (1995) Query by image and video content: The QBIC system. *IEEE Computer* 28(9): 23-32.
- Fox E & Shaw J (1994) Combination of multiple searches. *Proc. of the 2nd Text REtrieval Conference TREC-2, NIST Special Publications 500-215*, 243-252.
- Furht B & Marques O (eds) (2003) *Handbook of video databases: design and applications*. CRC Press, 1st edition.
- Hadid A, Pietikäinen M & Ahonen T (2004) A discriminative feature space for detecting and recognizing faces. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington, D.C., United States*, 2: 797-804.
- Hadjidemetriou E, Grossberg MD & Nayar SK (2004) Multiresolution histograms and their use for recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence* 26(7): 831-847.
- Hafner J, Sawhney HS, Equitz W, Flickner M & Niblack W (1995) Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. Pattern Analysis and Machine Intelligence* 17(7): 729-736.
- Hampapur A, Gupta A, Horowitz B, Shu CF, Fuller C, Bach J, Gorkani M & Jain R (1997) Virage video engine. *Proc. SPIE Storage Retrieval Image Video Databases V, Sethi IK, Jain RC (eds), San Jose, CA, United States*, 3022: 188-198.
- Hanjalic A & Xu LQ (2005) Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1): 143-154.
- Haralick R, Shanmugam K & Dinstein I (1973) Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3: 610-621.
- Harman DK (1995) Evaluation techniques and measures. *Proc. of the third Text Retrieval Conference TREC-3, A5-A13*.
- Hauptmann AG (2005) Lessons for the Future from a Decade of Informedia Video Analysis Research. *International Conference on Image and Video Retrieval (CIVR'05), National University of Singapore, Singapore*, In: Springer LNCS 3568: 1-10.
- Hauptmann AG & Christel MG (2004) Successful approaches in the TREC video retrieval evaluations. *Proc. of the 12th Annual ACM International Conference on Multimedia, New York, NY, United States*, 668-675.
- Heesch D & Rüger SM (2004a)  $NN_k$  Networks for content-based image retrieval. *Proc. of 26th European Conference on IR Research, Sunderland, UK*, 253-266.
- Heesch D & Rüger SM (2004b) Three interfaces for content-based access to image collections. in Enser P *et al.* (eds) *Proc. of International Conf. on Image and Video Retrieval (CIVR 2004)*, Springer LNCS 3115, Dublin, Ireland, 491-499.

- Hiemstra D (1998) A linguistically motivated probabilistic model of information retrieval. Proc. European Conference on Digital Libraries, 569-584.
- Ho T, Hull J & Srihari S (1994) Decision combination in multiple classifier systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 16(1): 66-75.
- Hoashi K, Sugano M, Naito M, Matsumoto K & Sugaya F (2005) Video story segmentation and its application to personal video recorders. Proc. of International Conference on Image and Video Retrieval, LNCS 3568, 39-48.
- Hollink L, Nguyen GP, Koelma D, Schreiber AT & Worring M (2005) Assessing user behaviour in news video retrieval. IEE proceedings on Vision, Image and Signal Processing, 152(6): 911-918.
- Howarth P & R ger S (2005) Fractional distance measures for content-based image retrieval. 27th European Conference on Information Retrieval (ECIR), Springer LNCS 3408, Santiago de Compostela, Spain, 447-456.
- Hsu W, Chua TS & Pung HK (1995) An integrated color-spatial approach to content-based image retrieval. Proc. of 3rd International ACM Multimedia Conference (ACM MM '95), San Francisco, CA, United States, 305-313.
- Huang J, Kumar SR, Mitra M, Zhu WJ & Zabih R (1999) Spatial color indexing and applications. International Journal of Computer Vision, 35(3): 245-268.
- Huang TS, Mehrotra S & Ramchandran K (1996) Multimedia analysis and retrieval system (MARS) project. Proc. of 33rd Annual Clinic on Library Application on Data Processing - Digital Image Access and Retrieval, Urbana-Champaign, IL, USA.
- Ikizler N & Duygulu P (2005) Person Search Made Easy. Proc. of the 4th International Conference on Image and Video Retrieval (CIVR 2005), Singapore, 578-588.
- Internet Archive: Moving Image Archive (n.d.) Referenced November 5<sup>th</sup> 2006 from: <http://www.archive.org/movies/>.
- Iyengar G & Nock HJ (2003) Discriminative model fusion for semantic concept detection and annotation in video. Proc. of the 11th ACM International Conference on Multimedia, Berkeley, CA, United States, 255-258.
- Jain RC (1992) Proceedings of US NSF workshop on visual information management systems, ed., ACM SIGMOD Record, 22(3): 57-75.
- Kato T (1992) Database architecture for content-based image retrieval. Proc. of SPIE Image Storage and Retrieval Systems, San Jose, CA, USA, In: Jamberdino AA & Niblack CW (eds), 1662: 112-123.
- Kherfi ML, Ziou D & Bernardi A (2004) Image retrieval from the world wide web: issues, techniques and systems. ACM Computing Surveys 36: 35-67.
- Kraaij W, Smeaton AF, Over P & Arlandis J (2004) TRECVID 2004 - an overview. Online Proceedings of the TRECVID 2004, Gaithersburg, MD, United States., <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>
- Lew MS (ed) (2001) Principles of visual information retrieval, Springer-Verlag.
- Lew MS, Sebe N, Djeraba C & Jain R (2006) Content-based multimedia information retrieval: state of the art and challenges. ACM Transactions on Multimedia Computing, Communications and Applications 2(1): 1-19.
- Li H, Doermann D & Kia O (2000) Automatic text detection and tracking in digital video. IEEE Transactions on Image Processing, 9(1): 147-156.
- Li Y, Lee SH, Yeh CH & Kuo CCJ (2006) Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques. IEEE Signal Processing Magazine 23(2): 79-89.
- Liu F & Picard RW (1996) Periodicity, directionality and randomness: Wold features for image modeling and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(7): 722-733.

- Ma W & Zhang H (1998) Benchmarking of image features for content-based retrieval. Proc. of 32<sup>nd</sup> Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, United States, 1: 253-257.
- Manjunath BS, Ohm JR, Vasudevan VV & Yamada A (2001) Color and texture descriptors. IEEE Transactions on Circuits and Systems for Video Technology 11(6): 703-715.
- Manjunath BS, Salembier P & Sikora T (eds) (2002) Introduction to MPEG-7: multimedia content description language, John Wiley & Sons, New York.
- Mao J & Jain AK (1992) Texture classification and segmentation using multiresolution simultaneous autoregressive models. Pattern Recognition 25(2): 173188.
- McDermott D (1982) A temporal logic for reasoning about processes and plans. Cognitive Science 6(2): 101-155.
- Mc Donald K & Smeaton AF (2005) A comparison of score, rank and probability-based fusion methods for video shot retrieval. (CIVR'05) - International Conference on Image and Video Retrieval, Leow WK *et al.* (eds), LNCS 3569, Singapore, 61-70.
- Mojsilovic A, Kovacevic J, Hu J, Safranek RJ & Ganapathy SK (2000) Matching and retrieval based on the vocabulary and grammar of color patterns. IEEE Trans. Image Processing, 9(1): 38-54.
- Naphade MR & Huang TS (2000) Semantic video indexing using a probabilistic framework. Proc. of 15th International Conference on Pattern Recognition, 3: 79-84.
- Naphade MR & Huang TS (2002) Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. IEEE Transactions on Neural Networks, 13(4): 793-810.
- Naphade MR, Kristjansson T, Frey B & Huang TS (1998) Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval in multimedia systems. Proc. of International Conference on Image Processing, 3: 536-540.
- Naphade MR & Smith JR (2003) A hybrid framework for detecting the semantics of concepts and context. Bakker EM *et al.* (eds) CIVR 2003 - International Conference on Image and Video Retrieval, LNCS 2728, 196-205.
- Oard DW & Dorr BJ (1996) A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, Univ. of Maryland, Inst. for Advanced Computer Studies.
- Ogawa Y, Morita T & Kobayashi K (1991) A fuzzy document retrieval system using the keyword connection matrix and a learning method. Fuzzy sets and systems, 39: 163-179.
- Ogle VE & Stonebraker M (1995) Chabot: retrieval from a relational database of images, IEEE Computer 28: 40-48.
- Ojala T, Pietikäinen M & Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7): 971-987.
- Open Video Project (n.d.) Referenced November 5<sup>th</sup> from: <http://www.open-video.org/>.
- Over P, Ianeva T, Kraaij W & Smeaton AF (2005) TRECVID 2005 an overview. Online Proceedings of the TRECVID 2005, Gaithersburg, MD, U.S., <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>
- Pass G, Zabih R & Miller J (1996) Comparing images using color coherence vectors. Proc. of the 4th International ACM Multimedia Conference (ACM MM '96), Boston, MA, United States, 65-73.
- Pentland A, Picard RW & Sclaroff S (1994) Photobook: Tools for content-based manipulation of image databases. Storage and Retrieval for Image and Video Databases II, Proc. of SPIE, San Jose, CA, USA, 2185: 34-47.
- Ponte JM & Croft WB (1998) A language modeling approach to information retrieval system. Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, New York, United States, 275-281.
- Porter M (1980) An algorithm for suffix stripping program. Program, 14(3): 130-137.

- Prewitt JMS (1970) Object enhancement and extraction. In Lipkin BS & Rosenfeld A (eds) *Picture Processing and Psychopictorics*, Academic Press, New York.
- Robertson S & Sparck Jones K (1976) Relevance weighting of search terms. *Journal of the American Society for Information Science* 27: 129-146.
- Rocchio JJ (1971) Relevance feedback in information retrieval. In Salton G (ed) *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice Hall, 313-323.
- Rodden K, Basalaj W, Sinclair D & Wood K (2001) Does organisation by similarity assist image browsing? *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, Seattle, WA, United States, 190-197.
- Rui Y, Huang TS & Chang SF (1999). Image retrieval: current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation* 10(1): 39-62.
- Salton G & Buckley C (1990) Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4): 288-297.
- Salton G, Fox E & Wu H (1983) Extended Boolean information retrieval. *Communications of the ACM* 26(11): 1022-1036.
- Salton G, Wong A & Yang CS (1975) A vector space model for automatic indexing, *Communications of the ACM* 18(11): 613-620.
- Salton G & Yang C (1973) On the specification of term values in automatic indexing. *Journal of Documentation*, 29: 351-372.
- Santini S (2001) *Exploratory Image Databases*. Academic Press.
- Santini S, Gupta A & Jain R (2001) Emergent semantics through interaction in image databases. *IEEE Transactions on Knowledge and Data Engineering* 13(3): 337-351.
- Sav S, Lee H, Smeaton AF & O'Connor NE (2005) Using segmented objects in ostensive video shot retrieval. In: Detyniecki M *et al.* (eds) *Proc. of 3rd International Workshop on Adaptive Multimedia Retrieval*, 155 - 167.
- Scaringella N, Zoia G, Mlynek D (2006) Automatic genre classification of music content: a survey *IEEE Signal Processing Magazine* 23(2): 133-141.
- Sclaroff S, Cascia M & Sethi S (1999) Unifying textual and visual cues for content-based image retrieval on the world wide web. *Computer Vision and Image Understanding* 75(1-2): 86-98.
- Sebe N, Lew MS & Huijsmans DP (2000) Toward improved ranking metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(10): 1132-1143.
- Shih TK (ed) (2003) *Distributed Multimedia Databases: Techniques and Applications*, Idea Group Publishing.
- Smeaton AF, Gurrin C, Lee H, Mc Donald K, Murphy N, O'Connor N, O'Sullivan D, Smyth B & Wilson D (2004) The fischlár-news-stories system: personalised access to an archive of TV news. *RIA0 2004 - Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, Avignon, France.
- Smeaton AF, Kraaij W & Over P (2003) TRECVID 2003 – an overview. *Online Proceedings of the TRECVID 2003*, <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>, Gaithersburg, MD, United States.
- Smeaton AF & Over P (2002) The TREC-2002 video track report. *NIST Special Publication, SP 500-251: Proc. of the Eleventh Text REtrieval Conference TREC 2002*, Gaithersburg, MD, United States, 69-85.
- Smeulders A, Worring M, Santini S, Gupta A & Jain R (2000) Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12): 1349-1380.
- Smith AR (1978) Color gamut transform pairs. *Proc. of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, 12-19.

- Smith JR & Chang SF (1996) VisualSEEK: a fully automated content-based image query system. Proc. of the 4th International ACM Multimedia Conference ACM MM '96, Boston, MA, United States, 87-98.
- Snoek C & Worring M (2005) Multimodal video indexing: a review of the state-of-the-art. *Multimedia Tools Appl.* 25(1): 5-35.
- Snoek C, Worring M & Hauptmann AG (2004) Detection of TV news monologues by style analysis. Proc. of International Conference on Multimedia and Expo, Taipei, Taiwan, 2: 1103-1106.
- Snoek C, Worring M & Smeulders AWM (2005) Early versus late fusion in semantic video analysis. Proc. of ACM Multimedia Conference, Singapore, 399-402.
- Souvannavong F, Merialdo B & Huet B (2004) Latent semantic analysis for an effective region-based video shot retrieval system. Proc. of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval MIR, New York, NY, United States, 243-250.
- Sparck Jones K & Willett P (eds) (1997) *Readings in Information Retrieval*. Morgan Kaufmann Publishers Inc.
- Stricker M & Orengo M (1995) Similarity of color images. Storage and Retrieval for Image and Video Databases III, Proc. of SPIE, San Jose, CA, USA, 2420: 381-392.
- Swain MJ & Ballard DH (1991) Color indexing. *Int. Journal of Computer Vision* 7: 11-32.
- Tamura H & Yokoya N (1984) Image database systems: a survey. *Pattern Recognition* 17(1): 29-43.
- Tonomura Y, Akutsu A, Tangiguchi Y & Suzuki G (1994) Structured video computing. *IEEE MultiMedia* 1(3): 34-43.
- TREC Video Retrieval Evaluation Home Page n.d. Cited November 5<sup>th</sup> 2006 from: <http://www-nlpir.nist.gov/projects/trecvid/>
- Turtle H & Croft WB (1991) Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems* 9(3): 187-222.
- van Rijsbergen CJ (1979) *Information retrieval*. London: Butterworths (2nd ed)
- Vasconcelos N & Kunt M (2001) Content-based retrieval from image databases: current solutions and future directions. Proc. of IEEE International Conference on Image Processing ICIP, Thessaloniki, Greece, 3: 6-9.
- Wactlar HD, Kanade T, Smith MA & Stevens SM (1996) Intelligent access to digital video: informedia project. *Computer* 29(5): 46-52.
- Wang Y, Liu Z & Huang JC (2000) Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6): 12-36.
- Westerveld T & de Vries AP (2005) Generative probabilistic models for multimedia retrieval: query generation against document generation. *IEE Proceedings - Vision, Image, and Signal Processing*, 152(6): 852-858.
- Wilkinson R & Hingston P (1991) Using the cosine measure in a neural network for document retrieval. Proc. of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago, Illinois, United States, 202-210.
- Wold E, Blum T, Keislar D & Wheaton J (1996) Content-based classification, search, and retrieval of audio. *IEEE Multimedia* 3(3): 27-36.
- Wong SKM, Ziarko W & Wong PCN (1985) Generalized vector space model in information retrieval. Proc. 8<sup>th</sup> ACM SIGIR Conference on Research and Development in Information Retrieval, New York, United States, 18-25.
- Wu Y, Chang EY, Chang KCC & Smith JR (2004) Optimal multimodal fusion for multimedia data analysis. Proc. of ACM Multimedia (MM'04), New York, NY, United States, 572-579.
- Xiong Z, Zhou XS, Tian Q, Rui Y & Huang TS (2006) Semantic retrieval of video - review of research on video retrieval in meetings, movies and broadcast news, and sports. *IEEE Signal Processing Magazine*, 23(2): 18-27.

- Yan R & Hauptmann AG (2003) The combination limit in multimedia retrieval. Proc. of the 11th ACM International Conference on Multimedia, Berkeley, CA, USA, 339-342.
- Yan R & Hauptmann AG (2004) Co-retrieval: a boosted reranking approach for video retrieval. In Enser P *et al.* (eds) Proc. International Conf. on Image and Video Retrieval (CIVR'04), LNCS 3115, 60-69.
- Yan R, Yang J & Hauptmann A (2004) Learning query-class dependent weights in automatic video retrieval. Proc. of ACM Multimedia Conference (MM'04), New York, NY, United States, 548-555.
- Yang J & Hauptmann A. (2004) Naming every individual in news video monologues. Proc. of ACM Multimedia Conference MM'04, New York, NY, United States, 580-587.
- Yeung M, Yeo BL, Wolf W & Liu B (1995) Video browsing using clustering and scene transitions on compressed sequences. Proc. of Multimedia Computing and Networking, 399-413.
- Yoshitaka A & Ichikawa T (1999) A survey of content-based retrieval for multimedia databases. IEEE Transactions on Knowledge and Data Engineering 11(1): 81-93.
- Zobel J (1998) How reliable are the results of large-scale information retrieval experiments? Proc. of the 21st annual international ACM SIGIR conference on research and development in information retrieval, Melbourne, Australia, 307-314.
- Zhang HJ, Smoliar SW & Wu JH (1995) Content-based video browsing tools. Proc. SPIE Storage and Retrieval for Image and Video Databases, 389-398.
- Zhang HJ, Wu J, Zhong D & Smoliar SW (1997) An integrated system for content-based video retrieval and browsing. Pattern Recognition 30(4): 643-658.
- Zhuang Y, Rui Y, Huang T & Mehrotra S (1998) Adaptive key frame extraction using unsupervised clustering. Proc. of IEEE International Conference on Image Processing, Chicago, IL, United States, 866-870.
- Zhong Y, Zhang HJ & Jain AK (2000) Automatic caption localization in compressed video. IEEE Transactions on Pattern Recognition and Machine Intelligence 22(4): 385-392.



## Original Publications

- I Ojala T, Rautiainen M, Matinmikko E & Aittola M (2001) Semantic image retrieval with HSV correlograms. Proc. 12th Scandinavian Conference on Image Analysis, Bergen, Norway, 621-627.  
Reprinted, with permission, from the proceedings of 12th Scandinavian Conference on Image Analysis.
- II Rautiainen M & Doermann D (2002) Temporal color correlograms for video retrieval. Proc. 16th International Conference on Pattern Recognition, Quebec, Canada, 1: 267-270.  
© 2002 IEEE. Reprinted, with permission, from the proceedings of 16th International Conference on Pattern Recognition.
- III Rautiainen M, Ojala T & Kauniskangas H (2001) Detecting perceptual color changes from sequential images for scene surveillance. IEICE Transactions on Information and Systems, Special Issue on Machine Vision Applications, E84-D: 1676-1683.  
Copyright 2001 IEICE. Reprinted, with permission, from IEICE Transactions on Information and Systems.
- IV Rautiainen M, Seppänen T, Penttilä J & Peltola J (2003) Detecting semantic concepts from video using temporal gradients and audio classification. International Conference on Image and Video Retrieval, Urbana, IL, 260-270.  
Copyright 2003. Reprinted with kind permission of Springer Science and Business Media.
- V Rautiainen M & Seppänen T (2005) Comparison of visual features and fusion techniques in automatic detection of concepts from news video. Proc. IEEE International Conference on Multimedia & Expo, Amsterdam, Netherlands, 932-935.  
© 2005 IEEE. Reprinted, with permission, from the proceedings of 2005 IEEE International Conference on Multimedia & Expo.

- VI Rautiainen M, Ojala T & Seppänen T (2004) Analysing the performance of visual, concept and text features in content-based video retrieval. Proc. 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York, NY, 197-205.  
© 2004 ACM, Inc. Included here by permission.
- VII Rautiainen M, Ojala T & Seppänen T (2003) Cluster-temporal video browsing with semantic filtering. Proc. 5th International Conference on Advanced Concepts for Intelligent Vision Systems CD-ROM, Ghent, Belgium, 116-123.  
Copyright 2003. Reprinted, with permission, from the proceedings of 2003 Advanced Concepts for Intelligent Vision Systems.
- VIII Rautiainen M, Ojala T & Seppänen T (2004) Cluster-temporal browsing of large news video databases. Proc. 2004 IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, 2: 751-754.  
© 2004 IEEE. Reprinted, with permission, from the proceedings of 2004 IEEE International Conference on Multimedia & Expo.
- IX Rautiainen M, Ojala T & Seppänen T (2005) Content-based browsing in large news video databases. 5th IASTED International Conference on Visualization, Imaging and Image Processing, Benidorm, Spain, 731-736.  
Copyright 2005 ACTA Press. Reprinted, with permission, from the proceedings of 2005 IASTED International Conference on Visualization, Imaging and Image Processing.
- X Rautiainen M, Seppänen T & Ojala T (2006) Advancing content-based retrieval effectiveness with cluster-temporal browsing in multilingual video databases, Proc. 2006 IEEE International Conference on Multimedia and Expo, Toronto, Canada, 377-380.  
© 2006 IEEE. Reprinted, with permission, from the proceedings of 2006 IEEE International Conference on Multimedia & Expo.
- XI Rautiainen M, Seppänen T & Ojala T (2006) On the significance of cluster-temporal browsing for generic video retrieval – a statistical analysis. Proc. ACM Multimedia Conference, Santa Barbara, CA, United States, 125-128.  
Copyright 2006 held by authors. Reprinted with permission.

Original publications are not included in the electronic version of the dissertation.

247. Jortama, Timo (2006) A self-assessment based method for post-completion audits in paper production line investment projects
248. Remes, Janne (2006) The development of laser chemical vapor deposition and focused ion beam methods for prototype integrated circuit modification
249. Kinnunen, Matti (2006) Comparison of optical coherence tomography, the pulsed photoacoustic technique, and the time-of-flight technique in glucose measurements *in vitro*
250. Iskanius, Päivi (2006) An agile supply chain for a project-oriented steel product network
251. Rantanen, Rami (2006) Modelling and control of cooking degree in conventional and modified continuous pulping processes
252. Koskiahho, Jari (2006) Retention performance and hydraulic design of constructed wetlands treating runoff waters from arable land
253. Koskinen, Miika (2006) Automatic assessment of functional suppression of the central nervous system due to propofol anesthetic infusion. From EEG phenomena to a quantitative index
254. Heino, Jyrki (2006) Harjavallan Suurteollisuuspuisto teollisen ekosysteemin esimerkkinä kehitettäessä hiiliteräksen ympäristömyönteisyyttä
255. Gebus, Sébastien (2006) Knowledge-based decision support systems for production optimization and quality improvement in the electronics industry
256. Alarousu, Erkki (2006) Low coherence interferometry and optical coherence tomography in paper measurements
257. Leppäkoski, Kimmo (2006) Utilisation of non-linear modelling methods in flue-gas oxygen-content control
258. Juutilainen, Ilmari (2006) Modelling of conditional variance and uncertainty using industrial process data
259. Sorvoja, Hannu (2006) Noninvasive blood pressure pulse detection and blood pressure determination
260. Pirinen, Pekka (2006) Effective capacity evaluation of advanced wideband CDMA and UWB radio networks
261. Huuhtanen, Mika (2006) Zeolite catalysts in the reduction of NO<sub>x</sub> in lean automotive exhaust gas conditions. Behaviour of catalysts in activity, DRIFT and TPD studies

Book orders:  
OULU UNIVERSITY PRESS  
P.O. Box 8200, FI-90014  
University of Oulu, Finland

Distributed by  
OULU UNIVERSITY LIBRARY  
P.O. Box 7500, FI-90014  
University of Oulu, Finland

S E R I E S E D I T O R S

**A**  
**SCIENTIAE RERUM NATURALIUM**  
*Professor Mikko Siponen*

**B**  
**HUMANIORA**  
*Professor Harri Mantila*

**C**  
**TECHNICA**  
*Professor Juha Kostamovaara*

**D**  
**MEDICA**  
*Professor Olli Vuolteenaho*

**E**  
**SCIENTIAE RERUM SOCIALIUM**  
*Senior Assistant Timo Latomaa*

**E**  
**SCRIPTA ACADEMICA**  
*Communications Officer Elna Stjerna*

**G**  
**OECONOMICA**  
*Senior Lecturer Seppo Eriksson*

**EDITOR IN CHIEF**  
*Professor Olli Vuolteenaho*

**EDITORIAL SECRETARY**  
*Publications Editor Kirsti Nurkkala*

ISBN 951-42-8299-X (Paperback)

ISBN 951-42-8300-7 (PDF)

ISSN 0355-3213 (Print)

ISSN 1796-2226 (Online)

