VITALAS

State of the Art in Cross-Media Indexing
D 3.1.1

Project Number:	FP6 - 045389
Deliverable id:	D 3.1.1
Deliverable name:	Literature Review Looking at the State-of-the Art in Cross-Media Indexing
Date:	10 Sept 2007

Information Society Technologies

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	VITALAS
Project Full Name:	Video & image Indexing and Retrieval in the Large Scale
Document id:	D 3.1.1
Document name:	Literature Review Looking at the State-of-the Art in Cross-Media Indexing
Document type (PU, INT, RE)	RE
Version:	V 3.0
Date:	10 Sept 2007
Authors: Organisation: Email Address:	J. Tait & M. Abusalah University of Sunderland {John.Tait, mustafa.abusalah}@sunderland.ac.uk

Document type PU = public, INT = internal, RE = restricted

ABSTRACT:

This deliverable overviews the state of the art in Cross-Media Indexing. A wide range of literature has been reviewed in the course of the associated workpackage. Media for which previous work is analysed in depth are still-image, video, audio, and text. The deliverable concludes that the most promising techniques all rest on supervised machine learning and this should be the focus of work in VITALAS.

KEYWORD LIST: Cross-media indexing, Video indexing, Audio indexing, Image Annotation, Supervised Machine Learning, Unsupervised Machine Learning, Dimensionality reduction.

MODIFICATION CONTROL			
Version	Date	Status	Author
1.01	27 February 2007	Draft	J. Tait
1.02	11 March 2007	Draft	J. Tait
1.03	16 March 2007	Draft	M. Abusalah
1.04	30 March 2007	Draft	J. Tait
1.05	4 April 2007	Draft	M. Abusalah
1.06	18 April 2007	Draft	J. Tait
1.07	18 April 2007	Draft	M. Abusalah
1.08	24 April 2007	Draft	M. Abusalah
1.9	5 May 2007	Draft	J. Tait
2.0	7 May 2007	Draft	M. Abusalah
3.0	10 Sept 2007	Final	M. Oakes

List of Contributors

- Mustafa Abusalah, University of Sunderland
- Anastasios Delopoulos, CERTH-ITI
- Arjen De Vries, CWI
- Christos Diou, CERTH-ITI
- Alexis Joly, INRIA
- Panagiotis Panagiotopoulos, CERTH-ITI
- Christos Papachristou, CERTH-ITI
- Michael Oakes, University of Sunderland
- Daniel Schneider, Fraunhofer IAIS
- John Tait, University of Sunderland
- Theodora Tsikrika, CWI
- Thijs Westerveld, CWI

Table of Contents

1 Introduction

Information Retrieval systems are conventionally divided into two main components: an ***indexing engine*** which works continuously in the background to extract what we will call here document *signatures* and stores them in an index database; and an interactive ***query*** or ***retrieval engine*** which allows searchers to type in queries, browse search results and select queries, present example "relevant" documents and so on.

The offline indexing processes allows the retrieval engine to operate at interactive speeds. Without it would not be possible to provide an acceptably rapid interactive experience for the searcher with the scale of document numbers needed by systems like VITALAS. It relies on the signatures being good representations of the documents for the purpose of retrieval: in other words they should (for most queries and most searchers, most of the time) allow relevant and irrelevant documents to be distinguished easily and with a minimal amount of stored data.

For modern text search engines the almost universal signature is a feature vector of weighted terms plus ancillary information like position of document in the hypertext topology [Brin 1998]. Weighting schemes include TF-IDF and BM25 [Robertson 1995].

For VITALAS many documents are composites of several different media (audio, transcription and video). A challenge for the project is the bringing together of these into single integrated signatures.

In text retrieval both the signature and the query can be simply translated into terms derived from words. Non-text media require an additional step of assigning terms or concepts to sections of related media, a process which is called here annotation.

This literature review therefore focuses on scientific and technological challenges associated with the first major objective of the VITALAS project, "*Cross-media indexing and retrieval*", as outlined in Annex I - "Description of Work". The major targets of the deliverable are a) to identify which are the major methodologies used in cross-media indexing, in particular supervised learning, b) what are the strengths and weaknesses of these methodologies, and how these weaknesses can be overcome, and c) in which subtasks of the project we shall decide whether particular state-of-the-art tools should be adopted.

Cross-media indexing refers to the process of automatic or semi-automatic annotation of data objects composed of several media inputs. For example, a multimedia object consisting of audio, video and text channels should be annotated using a single unified representation for indexing purposes. The goal of the VITALAS project is to produce such annotations at the semantic level i.e., multimedia objects should be described in terms of meaningful concepts, as these are understood by humans.

Hybrid visual and content representation is a well studied area in cross media indexing today. We therefore focus only on higher level issues in this section. However, audio data is less well developed – therefore we cover that area in more depth in this document.

This process is essentially different from indexing using a single medium (mono-media indexing), since (i) One must unify the different representations used for each medium into a single cross-media representation and (ii) the semantic interrelations between the various media can be exploited.

The general problem of cross-media indexing can be broken down into the following sub-problems:

1. *Automatic extraction of low-level features describing each medium.*

   This is performed in each medium separately and is one of the major tasks of WP2. Section 2 provides a short summary of such features for reference purposes.

2. *Multi-cue, mono-media document representation.*

   Given a single-medium data object, a set of low-level descriptions can be extracted. "Multi-cue, mono-media document representation" refers to the problem of providing a single representation of the object using all of its low level descriptions. As an example, given an image, the MPEG-7 dominant colour and edge-histogram descriptors can be extracted, but for indexing purposes one must unify those two low-level representations of the image (either into a semantic conceptual representation or low-level description of its content). Section 3 presents the existing state of the art techniques for multi-cue, mono-media document representation.

3. *Cross-media document representation.*

   Most mono-media data objects appear in the same context with other data objects (e.g., text with images in a webpage, audio with video in the same multimedia stream). Exploiting the semantic correspondence between the various media composing a single multimedia object can increase the amount of information available, leading to more accurate semantic-level descriptions. Cross-media document representation refers to the problem of deriving multimedia document representations using all available descriptions of each single medium (low-level features, multi-cue and concept-level representations). Related state of the art techniques are presented in Section 4.

Acceptable solutions to problems 2 and 3 require the existence of a mapping process that associates low-level features and descriptions with high-level concepts. Within VITALAS, supervised machine learning techniques will be considered to this end. Section 5 provides a review of related previous work.

2 Mono-Media Features and Annotation

The first stage in producing a cross-media index is the process of extracting features from the raw incoming media and possibly the generation of annotation data for each medium separately. In VITALAS this is the focus of WP2, and a report on the state of the art in the production of mono-media annotation appears in D2.0.

In audio media indexing, which is part of WP2 it is necessary to investigate the work done in [Ohtsuki 2006] [Adams 2003] [Neti 2000] [Albiol 2002] [Wang 2000] [Chang 2005] [Gagnon 2005].

For VITALAS a particular problem is that much of the true multimedia is likely to be Video data in which important cues for the generation of signatures are contained in the audio track. Audio features state of the art concepts for extracting rich transcriptions from audio media such as broadcast news or TV programmes are discussed in detail in deliverable D2.0 in WP2.

3 Multi-Cue Document Representation

3.1 Single-Medium Multi-Cue Document Representation

Each single feature extracted from a multimedia document is a description that provides visual, auditory or textual cues regarding the document's content. As pointed in [Nilsback 2004], however, in most scenarios a single feature does not provide a satisfactory document representation. For instance, global features like colour or texture histograms tend to suffer from clutter and light changes. Local features are sensitive to view changes. Shape descriptors do not handle occlusions very well. Even humans perform poorly in recognizing objects when forced to use a single visual cue, as several experiments have shown (see [Yuille 1994] and references therein).

The purpose of single-medium multi-cue document representation is to combine various cues or features (including the original data itself) into a single document representation. Within VITALAS, the modalities considered are text, audio (including speech), images and video. The targeted representations should be useful for multimedia indexing purposes and can be broadly classified into low and high level representations.

Low level representations include:

Integration of the original data (e.g., image pixels) into low-dimensional embeddings that can be used for indexing.

Unification of the extracted feature sets into a single feature vector or low-dimensional embeddings.

Aggregation of multiple cues to characterize a document on the basis of low-level entities (e.g., fast or slow motion, video shot detection).

High level representations, on the other hand, aim at using concepts to characterize the document's content and are thus better suited to content based retrieval systems since: (i) The representation is understood by humans, (ii) is more informative in terms of the document's content and (iii) indexing concepts overcomes the problem of the features' high dimensionality. The downside, of course, is that concept detection is a hard problem currently receiving intense research attention, but to which no established solutions exist.

A comment is in order regarding high level representations derived via classification techniques: A multimedia object may belong to multiple classes (or equivalently, be associated with multiple concepts). On the other hand, a classifier typically provides a single decision. There is actually no inconsistency, since: (i) A binary classification can be performed for each concept (present/not present) (ii) If the decision is a matter of degree (as in probabilistic or fuzzy classification approaches) it is possible that the final decision may include multiple classes.

The problem of integrating multiple cues from the same medium to a single representation can actually be examined more generally in the context of information fusion. And while most research papers do not actually focus on this aspect of the document representation, it is convenient for the purposes of this review. A broad classification of the information fusion problems uses two categories [Dasarthy 1994]: (i) Data fusion and (ii) decision fusion.

In the following sections, an attempt is made to identify state of the art techniques for the problem of single-medium, multi-cue document representation using these two categories. Notice that the details of methods used to identify concepts within documents are the subject of Section 5 and are not given here. Instead, this section focuses on techniques used for the fusion of multiple cues provided by low – level features.

3.2 Data Fusion

In data fusion (often denoted “early fusion” as opposed to “late fusion” methods that fall into the decision fusion category), raw data received from multiple sensors are fused into a single dataset. The representation is extracted from this dataset. In a very similar manner, features may be fused instead of raw data. But these are best explained by an example.

Sun [Sun 2003] proposed a method for multi-cue integration based on graph theory. Given the descriptions $\mathbf{h}_i^k$, $i = 1,\ldots,N$ and $k = 1,\ldots,K$ for K different cues of N objects the pair wise distance $d_{ij}^k = D^k\left(\mathbf{h}_i^k, \mathbf{h}_j^k\right)$ based on cue $\mathbf{h}^k$ can be evaluated and collected into a $N \times N$ matrix $\mathbf{D}^k$. There are a total of K distance matrices, one for each cue. Small distance measure d_{ij}^k indicates similar objects. For the same pair of objects, d_{ij}^k may be redundant or inconsistent (consider, for example, the comparison between colour and texture cues).

The author then proceeds to the normalization of the various distance measures. The normalized distance measures $\{f_{ij}^k\}_{k=1}^K$ are integrated into a single affinity measure w_{ij} by exponential decay $w_{ij} = \exp(-\sum_{k=1}^K \lambda_k f_{ij}^k)$ transforming the distance matrices $\{\mathbf{D}^k\}_{k=1}^K$ into an affinity matrix $\mathbf{W}$. The affinity matrix is further normalized as a transition probability matrix $\mathbf{P}$, whose matrix elements become

$$w_{ij} = \frac{1}{Z_i} \exp\left(-\sum_{k=1}^{K} \lambda_k f_{ij}^k \right)$$

The weights $\Lambda = \{\lambda_k\}_{k=1}^K$ control the relative cue importance/expressiveness (application specific). Details for the computation of Z_i are provided in the paper.

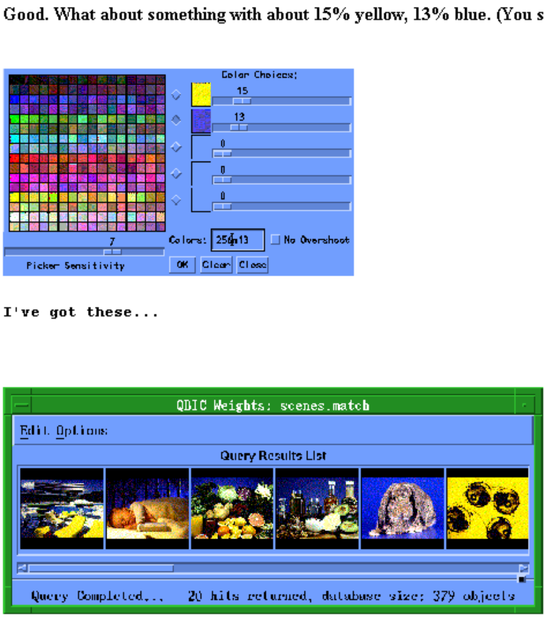
Following this procedure, object similarity is represented by a graph $G(V,E)$ with nodes V (objects) and edges E (similarity). For concept detection purposes, classification is performed via graph partitioning. Additionally, it is shown in the paper how an optimal transition probability matrix $\mathbf{P}^*$ can be computed that can later lead to easy, robust and efficient classification.

In [Yang 2003] two strategies are compared for feature integration: Parallel and serial feature fusion. Assume two feature spaces A and B, defined on a pattern sample space Ω. For an arbitrary sample $\xi \in \Omega$ (e.g., image in an image collection) the corresponding two feature vectors are $a \in A$ and $b \in B$. The serial combined feature of ξ is defined by $\gamma = \begin{pmatrix} a \\ b \end{pmatrix}$ thus combining the two feature vectors into a union-vector. Obviously, if a is n-dimensional and b is m-dimensional then γ is $(m+n)$-dimensional. For the parallel strategy, the two feature vectors a and b are combined into a complex vector $\gamma = a + ib$, which leads to a n-dimensional complex vector space, where n is the maximum number of dimensions of A and B. The authors provide details regarding the utilization of the parallel strategy and provide three experiments: (i) Character recognition based on the NUST603 handwritten Chinese character database, with feature vectors of equal dimensions. (ii) Recognition based on Concordia University CENPARMI handwritten numeral database, with unequal dimensions for the feature vectors. (iii) Face recognition using the ORL database (Olivetti Research Laboratory in Cambridge, UK). In all three experiments the parallel fusion strategy consistently achieved considerable improvement in recognition rates compared to the serial fusion strategy.

The serial fusion strategy, however, has been used in image and video retrieval systems due to its simplicity. One example is [Naphade 2003] where multiple combinations of features fused with the serial strategy were used to train an SVM classifier for learning semantic concepts. Other methods such as [Duygulu 2002], [Lavrenko 2003], [Lavrenko 2004] and [Yavlinsky 2005] match regions of images with keywords to perform image annotation, but despite the complexity of these models, regions are described by features fused using the serial strategy.

Recently Lafon, Keller and Coifman [Lafon 2006] proposed a data fusion approach using diffusion maps. More specifically, the paper addresses the following issues: (i) Proposes the use of Laplace-Beltrami normalization for data fusion by showing that it allows to merge datasets produced by the same source but with different densities. (ii) Suggests a new data fusion scheme by extending spectral embeddings using the geometric harmonics framework. (iii) Presents a novel spectral graph alignment approach to data fusion. Experimental evaluation of the methods is provided with applications to lip-reading and synchronization of head movement data. One important aspect of this work is that fused data are embedded at low-dimensional spaces, while at the same time capture local correlation of the data (contrary to embedding approaches such as Multidimensional Scaling).

Earlier multimedia retrieval systems utilized the low level representations of documents to support the query by example paradigm. The QBIC system [Flickner 1995] supports example queries for each feature separately (Figure 1). When using multiple features (see Figure 2 for example), the distances for each feature are combined. Virage [Bach 1996] proposed an improvement to QBIC for the use of multiple features. It supports queries using an arbitrary combination of colour, composition (colour layout), texture and structure (object boundary information). The user is able to adjust the weight of each feature according to their preference and the result is based on the weighted combination of each feature. In [Iqbal 2003] a more sophisticated similar method is presented, where the distance of each different feature is computed and then a unified distance metric is computed by distance normalization techniques. Then, weighting of distances allow users to fine-tune the retrieval process and increase accuracy.

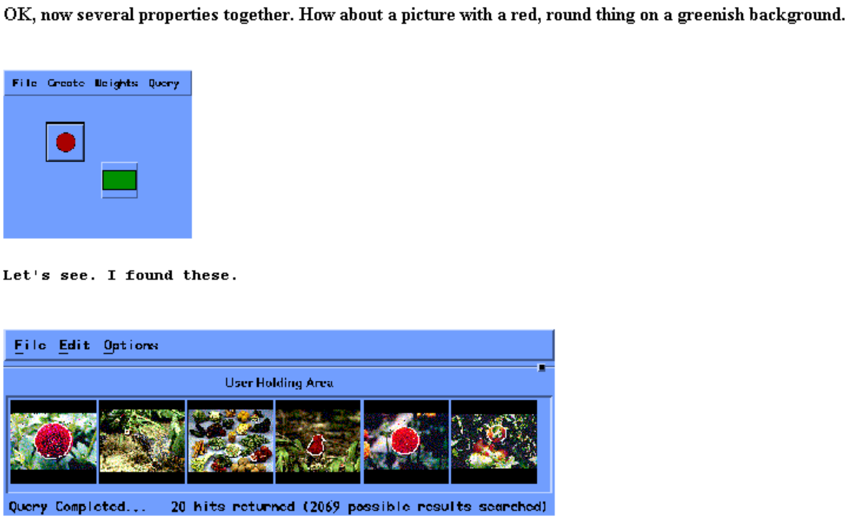


Good. What about something with about 15% yellow, 13% blue. (You say that like this:)

I've got these...

Query Completed... 20 hits returned, database size: 379 objects

Figure 1: Single feature query in QBIC



OK, now several properties together. How about a picture with a red, round thing on a greenish background.

Let's see. I found these.

Query Completed... 20 hits returned (2069 possible results searched)

Figure 2: Combination of features in QBIC

A similar functionality was also present in MARS [Huang 1997, Meh1997, Por1999], but was extended to receive relevance feedback and adjust the weights of features accordingly. Other early systems include Netra [Ma 1997] and VisualSEEK [Smith 1997], but the latter falls into the decision fusion category and is described in the next section.

As seen from the brief presentation of these methods and systems, the main difference of data fusion with most methods used in the decision fusion approach is that features are fused into a single feature (or distance between objects) prior to indexing or classification. Decision fusion methods, on the other hand, use each feature separately with trained classifiers or rule/knowledge based approaches and the final representation depends on their outputs.

Based on the available literature, data fusion techniques are less popular for the purposes of high level multi-cue document representation than decision fusion techniques. Note however, that Kokar et al provided a formalization of information fusion approaches, including data and decision fusion, in [Kokar & Tomasik 2001] and [Kokar 2004] and showed that decision fusion is actually a subclass of data fusion. In the choice between the two approaches, though, computational complexity is often a critical factor, depending on the application. The features' high dimensionality and the massive data produced in multimedia applications often lead to prohibitive computational cost for data fusion.

3.3 Decision Fusion

Smith and Chang [Smith 1997] describe a method for combining colour and spatial queries that was used in the VisualSEEK system. Colour features are combined with spatial locations and spatial extents (represented by bounding rectangles in images) to create region queries and find region matches. Then, Images that contain region matches that satisfy the spatial relationships present in the query image are retrieved. This approach is an example of low-level decision fusion, since each feature is used separately to retrieve relevant images (region matches) and then these are fused to obtain a final decision (based on spatial relations of regions).

A more recent example is provided in [Kushki 2004] for an image retrieval application. The outline of this system's operation is shown in Figure 3. Each feature (colour, texture, etc.) is used separately with its own distance metric. These distances are then used as arguments to membership functions associated with the features and a decision is made regarding the match of an image to the query for each descriptor. Then, aggregation is performed for the decisions of groups of features (e.g., aggregation of colour decisions). An overall aggregation stage provides the final decision.

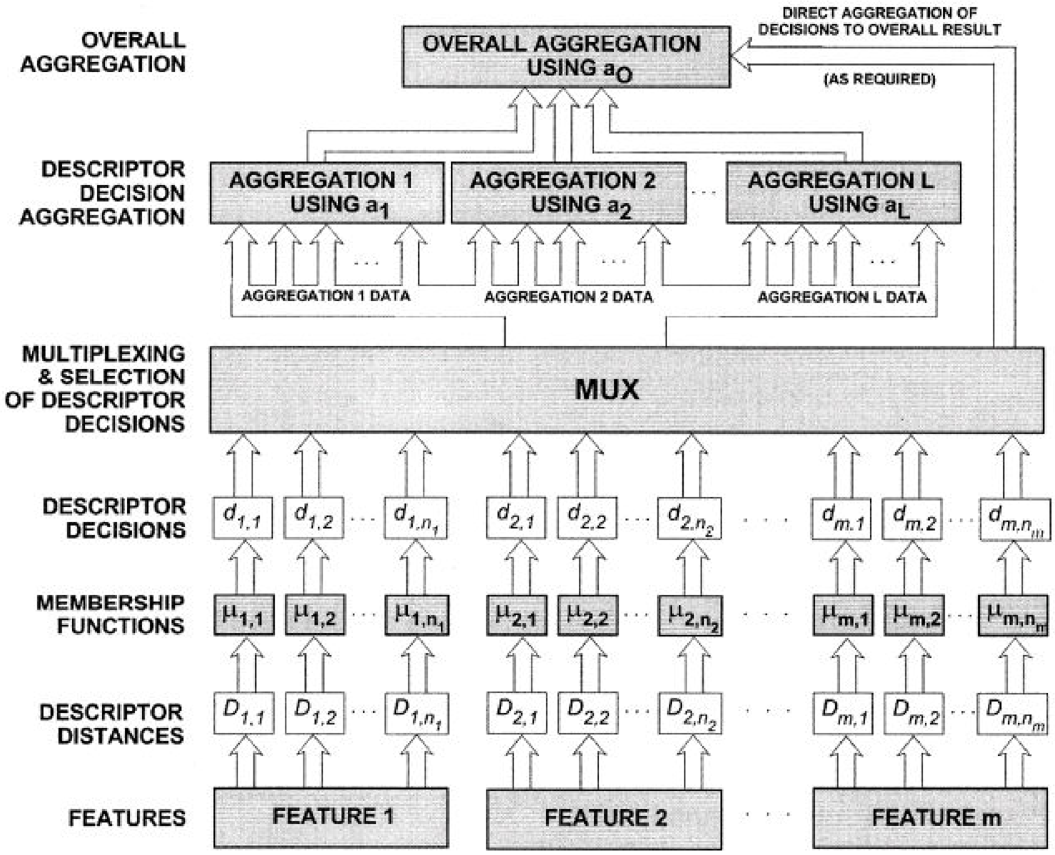

OVERALL AGGREGATION
OVERALL AGGREGATION USING a_O
DIRECT AGGREGATION OF DECISIONS TO OVERALL RESULT
(AS REQUIRED)
DESCRIPTOR DECISION AGGREGATION
AGGREGATION 1 USING a_1
AGGREGATION 2 USING a_2
AGGREGATION L USING a_L
AGGREGATION 1 DATA
AGGREGATION 2 DATA
AGGREGATION L DATA
MULTIPLEXING & SELECTION OF DESCRIPTOR DECISIONS
MUX
DESCRIPTOR DECISIONS
$d_{1,1}$ $d_{1,2}$ d_{1,n_1} $d_{2,1}$ $d_{2,2}$ d_{2,n_2} $d_{m,1}$ $d_{m,2}$ d_{m,n_m}
MEMBERSHIP FUNCTIONS
$\mu_{1,1}$ $\mu_{1,2}$ μ_{1,n_1} $\mu_{2,1}$ $\mu_{2,2}$ μ_{2,n_2} $\mu_{m,1}$ $\mu_{m,2}$ μ_{m,n_m}
DESCRIPTOR DISTANCES
$D_{1,1}$ $D_{1,2}$ D_{1,n_1} $D_{2,1}$ $D_{2,2}$ D_{2,n_2} $D_{m,1}$ $D_{m,2}$ D_{m,n_m}
FEATURES
FEATURE 1
FEATURE 2
FEATURE m

Figure 3: Operation of an image retrieval system based on decision fusion

As many have pointed out (see [Santini 1998] and [Naphade 2002a] for example), query by example systems have some inherent weaknesses, the most noticeable being the fact that often the results retrieved do not match the users' requirements (images with similar features do not always have semantic correspondence). Today, most systems aim at capturing multimedia objects' content via the use of concepts closer to the human perception than low level features. The automatic identification of such concepts requires as much information as possible; hence multiple cues are combined to identify concepts. Decision fusion techniques are often used for concept identification.

Such methods for integration of multiple cues can be classified into two general categories, knowledge-based and machine learning approaches.

3.3.1 Knowledge based decision fusion

The main idea behind knowledge – based methods is the encoding of knowledge related to the domain of interest, either directly or indirectly. Thus, either the features are processed via an algorithm that is suitable for the task at hand (e.g., annotation of soccer game videos) or knowledge is encoded so that concepts are formally defined in terms of low – level features. Then, concept identification is based on the evaluation of the extracted features' values using the defined algorithms, rules, or logic propositions.

In [Miyamori 2000] a method for automatic annotation of tennis video sequences is outlined. Having performed shot detection in the original video data (e.g., [Patel 1997], [Osian 2004]), the authors identify the following objects: Court and Net lines (using Hough transform and a model of the court), the players (using adaptive template matching and tracking), the ball (again using a template matching technique over several frames). Then analysis of players' behaviour is performed based on a model that utilizes the identified players from the previous step. Thus, concepts such as foreside swing, backside swing and over-the-shoulder swing are detected. In this manner, the system can support a wide range of queries regarding the position and behaviour of the players. A similar approach is presented in [Sudhir 1998]. More recently, [Assfalg 2003] presented another method where domain knowledge is expressed by a directed graph (automaton) that encodes the process of highlight detection (e.g., shot on goal). Another approach on the same domain is presented in [Ekin 2003]. For these methods, the extracted features assist in the detection of individual objects or regions in the videos (e.g., the combination of lines extracted via Hough transform fit a tennis court model). Then, their combination via domain-specific rules extracts higher level objects or events. This is a combination of indirect and direct use of domain knowledge. It is worth noting that in such methods there is a strong coupling of the individual decisions: The identification of a concept (such as "shot on goal") requires that the output of previous decisions (e.g., identification of court lines) is correct. [Yuan 2002] is a prime example of strong coupling, where video features are used successively in a decision tree to perform automatic video classification.

Research related to the development of the Semantic Web gave rise to a set of methods that use ontologies to formally and explicitly encode domain knowledge defining high level concepts in terms of low-level features (examples include [Dasiopoulou 2005], [Petridis 2006]). In [Bertini 2005a] and [Bertini 2005b] the incorporation of visual examples within the domain ontology is proposed to enrich expressiveness and performance, while [Holub & Perona 2005] provides an evaluation of the use of domain ontologies for image annotation using medical images as a case study. Hoogs et al [Hoogs 2003] alternatively propose the combination of visual analysis and the WordNet to create a large semantic knowledge base for video content annotation.

An issue that is prevalent to multimedia analysis and is not addressed directly within the context of ontologies is uncertainty and imprecision related to knowledge that best describes multimedia objects. For example, the statements "an apple is red" or "the speaker's voice becomes loud when something interesting happens" are essentially imprecise, since a wide range of colors can be characterized as "red", while "loud" is not much information when dealing with numerical audio samples. Reasoning with fuzzy rules for knowledge representation enables the representation of imprecise concepts and has been proposed in [Dorado 2004, Akrivas 2004, Falelakis 2005, Falelakis 2006]. Other researchers attempt to devise reasoning algorithms for fuzzy description logics [Straccia 1998, Sciascio 2000, Straccia 2000, Stoilos 2005], that would combine the strengths of description languages (ontologies) and fuzzy logic to enhance the expressiveness of knowledge representation for multimedia retrieval systems.

3.3.2 Machine learning and probabilistic decision fusion

In their 1998 paper [Saber 1998] E. Saber and A. M. Tekalp describe a set of methods for the classification of images or image regions based on colour, edge, texture and shape features. Then they propose a set of algorithms for feature integration: Parallel, sequential and Bayesian integration of features. The latter was proposed for colour $\mathbf{w}$ and texture $\mathbf{f}$ features using a single segmentation map $\mathbf{x}$, by using the maximum a posteriori (MAP) probability criterion, by assuming independence between features:

$$p(\mathbf{x} \mid \mathbf{w}, \mathbf{f}) \propto p(\mathbf{w}, \mathbf{f} \mid x) p(x) = p(\mathbf{w} \mid \mathbf{x}) p(\mathbf{f} \mid \mathbf{x}) p(\mathbf{x})$$

The class conditional pdf $p(\mathbf{w} \mid \mathbf{x})$ and $p(\mathbf{f} \mid \mathbf{x})$ are modelled by Gaussians and the a priori probability $p(\mathbf{x})$ is modelled by a Gibbs distribution. Thus, the original equation has the form,

$$p(\mathbf{x} \mid \mathbf{w}, \mathbf{f}) \propto \exp\{e_1 + e_2 + e_3\}$$

The MAP segmentation is performed through an iterative process that computes the mean vectors and covariance matrices used in the models of the above pdfs, estimating $\mathbf{x}$ until a convergence criterion is satisfied. The authors suggest that this process can be used to model and extract meaningful regions from an image and assign keywords to them.

A similar approach was later applied by Naphade and Huang in [Naphade 2000]. The serial data fusion strategy was used to combine features, but integration of various image regions was performed using decision fusion. For detecting sites, Gaussian Mixture Models are used for each feature and temporal flow is not taken into consideration. For objects and events, Hidden Markov Models (HMM) model the feature vector. The EM algorithm is used to estimate means, covariance matrices, mixing proportions (for GMM and HMM) and transition matrix (HMM). Concepts are then defined based on a binary hypothesis testing for each concept. For integrating region cues to identify frame level semantics the following procedure is used: The feature vector of a region is denoted as X_j and a random variable R_{ij} is defined for each region in a video frame, where

$$R_{ij} = \begin{cases} 1, \text{if Concept } i \text{ is present in region } j \\ 0, \text{otherwise} \end{cases}$$

By assuming uniform priors on the presence or absence of any concept in any region, the Bayes' rule gives

$$P(R_{ij} = 1 \mid X_j) = \frac{P(X_j \mid R_{ij} = 1)}{P(X_j \mid R_{ij} = 1) + P(X_j \mid R_{ij} = 0)}$$

To integrate those cues at the frame level, the frame level features F_i are defined

$$F_i = \begin{cases} 1, \text{if Concept } i \text{ is present in the current frame} \\ 0, \text{otherwise} \end{cases}$$

and using the compact notation $\mathbf{X} = \{X_1 \mathrm{K}\ X_M\}$ for all M regions of the current frame,

$$P(F_i = 0 \mid \mathbf{X}) = \prod_{j=1}^{M} P(R_{ij} = 0 \mid X_j)$$

$$P(F_i = 1 \mid \mathbf{X}) = 1 - P(F_i = 0 \mid \mathbf{X})$$

The authors then proceed to model the dependence between the concepts indicated by the features F_i that were ignored in the previous equations.

In [Naphade 2002b] the same approach was used, only this time a model was used for each feature vector (modeled with a GMM consisting of five Gaussians). The difference of log likelihoods for the positive and negative hypotheses was used as a measure of confidence that a concept is detected. A ranked list of all concepts was the produced based on this confidence measure.

The approach used in [Kittler 2001] for sports video annotation is along the same line, with the difference that each low level feature was used to detect a concept separately and these were fused by trained neural networks. In [Li & Wang 2003] multiresolution models with HMMs were used for the automatic annotation of images. Such statistical methods have also been used in more specialized methods, such as [Goldenstein 2003] and [Shet 2004], but some of the ideas presented in these papers can be used in multiple cue integration in general. Note the fact that in all of the above methods the conditional independence between features is assumed. In [Moreno 2005] figure/background estimation is performed using multiple cues with statistical integration. Estimation of pdfs is performed using particle filters; however conditional independence of features is not assumed.

In [Amir 2004] for the IBM TRECVID-04 system a different method was used. SVM classifiers (see Section 5) and maximum entropy models were used to train various models for each concept based on low level features (in some cases combined using the serial fusion strategy). Then, the outputs of classifiers are aggregated using ensemble fusion, described in [Lin 2003]. Each classifier generates an associated confidence score for data in the validation set. These scores are normalized to a range of [0, 1]. The normalization schemes include (i) rank normalization, (ii) range normalization and (iii) Gaussian normalization. After normalization, a combiner function selects a permuted subset of different classifiers and operates their normalized scores, essentially identifying the high-performing and complementary subsets of classifiers. Combination is performed by (i) minimum, (ii) maximum, (iii) average and (iv) product. The best performing normalized ensemble fusion is obtained by evaluating the average precision measure against the ground truth. Another example of the use of SVM classifiers is presented in [Nilsback 2004].

A more thorough and interesting study on classifier integration strategies is presented in [Kuncheya 2002]. Assume that each classifier D_j produces an estimate $d_{ji} \in [0,1]$ of the posterior probability $P(\omega_i \mid x)$ for a feature vector x and a class ω_i (binary classifiers are considered only, as in "concept is" or "is not" identified). The support for ω_i is yielded by multiple classifiers as

$$d_i(x) = F(d_{1i}(x), \mathrm{K}, d_{Li}(x))$$

where F may stand for (i) minimum, (ii) maximum, (iii) average, (iv) median and (v) majority vote. The author then proceeds (under certain assumptions) to model the probability of error for each fusion method. Each classifier gives an output P_j as an estimate of the posterior probability $P(\omega_i \mid x) = p > 0.5$. P_j are i.i.d. coming from a fixed distribution (normal or uniform) with mean p. It was shown that maximum and minimum were identical for 2 classes regardless of the distribution of P_j. The same stands for median and majority vote. Minimum/maximum fusion was found to be the best for uniformly distributed P_j.

Regarding the voting approaches mentioned above, [Brautigam 1998] provides a thorough presentation of various methods. For binary classifiers the Unanimity, Byzantine and Majority thresholding schemes are presented. Another general class of voting schemes, the weighted consensus voting is defined. The paper provides an analysis of how voting schemes can be used for efficient cue integration, along with experimental results. [Hayman & Eklundh 2002] shows another application of voting and probabilistic integration approaches and the inclusion of reliability measures for each cue. Similar techniques were also used in [Spengler 2003].

[Opelt 2006] provides the combined use of some of the techniques presented above for object recognition, with an important difference: The recognition is a two-stage classification process. A weak hypothesis is acquired with multiple features extracted from image regions with weights assigned to the classifiers. Boosting is applied to adjust the weights and provide a final classification result.

A different approach is followed in [Laaksonen 2002] and [Koskela 2006] where Self Organized Maps (SOM) are used to model each visual feature extracted from images and video respectively. The produced maps determine the presence or absence of a concept in the visual scene based on a single visual feature, as in Figure 4. The integration of the different maps is simply performed by summing up the respective map values.

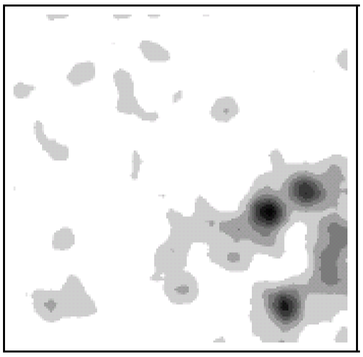


Figure 4: Concept "explosion/fire" on the colour – layout SOM. Areas occupied by objects are shown with gray shades.

3.4 Discussion

The short review in this section presents several techniques for the integration of multiple-cues provided by features extracted automatically from the content of single – medium documents. Researchers in the field of multimedia retrieval are currently approaching the problem of multi – cue integration from a system's perspective (i.e., the integration method forms a part of the overall retrieval system), rather than a "stand alone" theoretical problem. As a consequence, there are not many works that attempt to provide strict formalization as well as evaluation approaches to integration in the context of multimedia retrieval. Despite that, several useful conclusions can be drawn:

Data fusion techniques are appealing due to the fact that the use of multiple classifiers for modelling a single concept is not required, while at the same time local correlations between features that express the same concept can be preserved.

Due to the high dimensionality of the extracted features, similarity queries and concept classification becomes very complex in computational terms. Research is now directed at devising useful low dimensional representations of the fused data, without loss of valuable information.

Decision fusion techniques are more appropriate from a computational complexity viewpoint, since the dimensionality of data used for each intermediate decision is low, compared to data fusion strategies. Additionally, decision fusion systems allow for more flexible design (esp. for the identification of high level concepts) since different classifiers and models may be used for each cue.

Due to the fact that independence of features is often assumed for decision fusion systems, information regarding the local correlation between features is lost, possibly leading to lower concept identification accuracy.

It seems that current state of the art systems usually utilize a combination of data and decision fusion. For different features that may have a strong correlation for the same concept (e.g., colour and texture features) data fusion is used for an initial hypothesis, while these combinations are then fused with decision fusion, often combining encoded prior knowledge and machine learning. Examples include the IBM research team system for the 2006 TRECVID evaluation [Campbell 2006], as well as the methods presented in [Fan 2004].

4 Cross-Media Representation

In this section a *multimedia document* is understood to be a composite group of conceptually or thematically related objects, for example an audio visual program stream consisting of one video, two audio channels, and perhaps a text transcription.

This section covers the process of producing indexing keys or signatures for such composite objects from simpler signatures or other data derived from the separate streams. We call this a *cross-media index.*

4.1 Local and Global Characterization of Multimedia Documents

There are a number of different ways in which one can characterize existing approaches to producing cross-media index representations. In particular the following stand out:

1) whether features from different modalities and media are considered the same or kept separately;

2) whether any attempt is made to reduce the dimensionality of the input feature vectors;

3) since most current successful systems are based on the probabilistic model of information retrieval or the related language modeling approach [Belew, 2000] [Westerveld 2005], whether the approach is generative or discriminative;

4) Whether the approach is based on emergent properties of the data (equivalent to unsupervised machine learning), or assignment of terms or concepts from some previously defined ontology, controlled or restricted vocabulary (equivalent to supervised machine learning).

4.1.1 Modality combination

Most current systems which allow any sort of combined searching keep the different modalities separate [Westerveld 2004]. That is the systems allow searching through text (caption, metadata) or typically low level image features (similarity searching on colour, texture and shape), audio (music similarity), Video, OCR and so on [Hauptmann 2004].

Perhaps the earliest exception was the ImageRover systems [Sclaroff 1997], which combined into a single feature vector a conventional text term vector with colour and texture image features. Latent semantic Indexing and Principal Component Analysis were used to reduce the dimensionality of the feature vector. [Westerveld 2000] adopted a similar approach.

[Blei 2004b] proposed a Latent Dirichlet Allocation (LDA) model that models documents based on exchangeability assumptions and the notion of a hidden semantic index. [Blei & Jordan 2003] extended the LDA model and proposed Correspondence LDA (Corr-LDA): a model that finds conditional relationships between latent variable representations of sets of image regions and sets of words. In other words Corr-LDA relates words and images. LDA is considered a scalable and effective model.

The cross-media relevance model (CMRM) [Jeon 2003] followed earlier machine translation inspired work [Duygulu, et al 2003] to relate textual and visual data, with some considerable success in terms of scale, although there are concerns about their methodology. The CMRM model was followed by the continuous-space relevance model (CRM) [Lavrenko 2003] to enhance scalability. [Lavrenko 2004] explored Gaussian distributions in a relevant setting and adapted CRM to the video domain.

Work on Dempster-Shafer theory for document retrieval has been done by [Ruthven & Lalmas 1997] [Lalmas 1997a] [Lalmas 1997b] but only small scale experiments on limited text test collections were undertaken. The Dempster-Shafer was also used in "EPIC: A Photograph Retrieval System based on evidence combination approach" by [Jose & Harper 1998]. This was also quite small scale.

There are a number of systems (mainly experimental) which have used Dempster-Shafer evidence theory to combine evidence form various modalities. In particular, the use of Dempster-Shafer theory for combining evidence (keywords found for different modalities) in cross-media indexing has been done by the EU-IST Reveal-This project [Yakici & Crestani 2006].

4.1.2 Dimensionality reduction

Conventional textual feature representations tend to have one feature per term (word or lemma), so the dimensionality of the feature vector is of the order of the numbers of terms in the vocabulary Image features are typically of the order of 20 or 30 features per blob or region. These very large vectors are unwieldy computationally and imply also very fine distinctions in indexing, which often cannot be supported by the discriminating power of the multimedia indexing systems.

The main techniques which are in use are:

1) Latent Semantic Analysis (LSA) [Hofmann 1999a] [Hofmann 2001], and the related Singular Value Decomposition (SVD) [Castelli 2003] [Chen 2005];

2) Principal Component Analysis (PCA) [Yang 2004] and Independent Component Analysis (ICA) [Wu 2004].

3) Random Projection where data is projected onto a random lower-dimensional subspace. s. Experiments in [Bingham & Mannila 2001] showed promising results comparable with conventional techniques like PCA.

4.1.3 Generative vs. Discriminative

Most multimedia indexing tasks can be seen as probabilistic decision theory or fundamentally classification tasks (see [Westerveld 2004] for a discussion). For example, how likely is it that this image contains a face, or this key word correctly applies this image or this region, or that the voice on this audio segment is this person. Successful face recognition techniques are investigated in [Yang 2004].

Discriminative models attempt to predict the most likely class or classes given the data space, so the focus is on finding the class boundaries in the data space. Although they are sometimes advocated for the sort of problem we are considering (see for example [Vapnik 1998]), they require consideration of all classes simultaneously: this may be of the order of the number of search or index terms available – thousands in our case – which may render them computational and operational unacceptable. Certainly we could find no examples of their successful use within a strictly probabilistic framework, although [Tsai 2006b] might be regarded as using a related technique using Support Vector Machines. Other models include Multiple Discriminant Analysis (MDA) [Duda & Hart 1973] Fisher Discriminant Analysis (FDA) Biased Discriminant Analysis (BDA) [Zhou & Huang 2001]. Linear Discriminant Analysis (LDA) is also very popular for designing ASR feature vectors; see for example [Eisele 1996].

Generative models estimate the probability density of each of the classes, and typically a Bayesian inversion is used to find the most likely class (see [Ng & Jordan 2002] for a review). The first really successful second generation Content Based Image Retrieval System Blobworld [Carson 2002] used a generative approach – in particular the Expectation-Maximisation (EM) algorithm [Dempster 1977]. Other examples include [Hoiem 2003], [Fergus 2003], [Schmid 2004], [Vasconcelos & Lippman 1998] and [Luo 2003]. A Dynamic Probabilistic Multimedia Retrieval Model implemented in [Ianeva 2004] that uses [Westerveld 2003a] generative model showed good MAP results.

A particular feature of the more successful systems is the use of *k-means* [McQueen 1967] [Kompatsiaris 2001] and Gaussian Mixture models in the more successful systems.

4.1.4 Supervised vs. unsupervised

Fundamentally supervised systems have some data which has been (usually manually) assigned with the required cross-media indexing representation, and the system learns a (generalised) model of this which can subsequently be applied to new, unseen data.

Unsupervised approaches attempt to break the incoming data into clusters or classes of similar data. The best known unsupervised clustering techniques are EM or mean shift [Comaniciu & Mea 1999].

Some systems fall between the two: for example in relevance feedback systems [Rocchio 1971], [Salton & Buckley 1990], searcher selections of relevant items can be used to refine subsequent retrievals, and combined with techniques like collaborative filtering [Hofmann 1999a] can be used to improve training sets which can then be used to train more effective models. Alternatively some systems combine an unsupervised phase with a subsequent labelling of clusters in some form with, for example, annotation key words. Examples of this last include the classic Blobworld work [Carson 2002], [Barnard 2003], and image annotation using Cross-Media Relevance Models [Jeon 2003]. A recent full study of image annotation problems can be found in [Virga & Duygulu 2005].

Examples of pure supervised systems include [Tsai 2006b], and an earlier approach to combine unlabeled data in supervised systems [Wu 2000] [Tsai 2004].

The problem with supervised systems is that large amounts of pre-indexed training data are needed. This is rarely available.

Examples of unsupervised systems include [Chen-Y 2005].

Unsupervised systems, on the other hand, may find patterns in data which are quite unlike human understanding of the data. This may make the cross media indexes unsuitable to support browsing, let alone key word or concept based initial query formulation.

This area is further expanded in Section 5.

4.2 Methods for Evaluation of Cross-Media Representations

There are fundamentally four methods for the evaluation of cross-media index representations:

1) Against a gold standard: a pre-specified set of annotations for a pre-selected collection of multimedia documents. ([Jones 1996], TRECVID [Smeaton 2004a] [Smeaton 2004b], Wang and others on Corel [Marchand-Maillet and Worring 2006]).

2) Human assessment of the accuracy of annotation/indexing terms assigned [Tsai 2006a] [Tsai 2004]

3) Human Centred experiments in which users are given a set of realistic retrieval tasks to do and their performance and satisfaction with the retrieval process is measured by a variety of means ([McDonald & Tait 2003], [Pickering & Rüger 2003]).

4) System Centred experiments in which performance is undertaken on a complete (simulated) retrieval set up with a set of pre-specified queries and known set of relevant documents for each query. This is the usual methodology for text retrieval ([Cleverdon, 1967], [Amir 2003]), but is rarely undertaken for multimedia retrieval due to the expense and difficulty of constructing test collections.

In approaches 1 and 4 assessment is usually via Precision and Recall (i.e. the proportion of the results returned which are relevant to the query, and the proportion of the relevant results which are returned by the system). There is a extensive literature discussing these measures and their interpretation in for example rank retrieval (i.e. where the aim is to return highly relevant documents first), where there are very large numbers of relevant documents (man web queries) and so on ([van Rijsbergen 1979], [Belew 2000], [Smeaton 2004b]).

4.3 Hybrid Visual and Conceptual Content Representation

If we regard the information provided concerning the target images or the possibilities of interaction between the user and the system, keywords and visual content appear to be rather complementary to each other and it is important to rely on both of them for the retrieval of images.

Keywords associated to an image can be divided in two categories: (i) keywords corresponding to *identifiable items* characterizing the *visual content* of the scene and (ii) keywords concerning the *context* and the *interpretation* of the scene. In some cases, it may be possible to automatically obtain keywords in the first category by image analysis, object detection and classification techniques. An explicit use of such keywords during retrieval, to complement descriptors of the visual appearance, will nevertheless be helpful even in cases where the results of image analysis are unreliable. Keywords in the last category are very unlikely to be automatically extracted from the images, unless very strong correlations exist in a specific image database between visual appearance and such keywords. However, the contribution of such keywords to the retrieval of relevant images is undeniable.

Consider for example an image presenting a mass meeting in a square in Lisbon during the Carnation Revolution (April 25th, 1974). Among the keywords associated to this image we may find (i) keywords corresponding to *identifiable items* such as “tank”, “car”, “building”, “people”, (ii) keywords describing rather *visual characteristics* of the scene such as “outdoor”, “daytime”, (iii) keywords concerning the *interpretation* of the scene such as “fraternization” (between the masses and the army) and (iv) keywords related to the *context* of the scene such as “Caetano”, “Salazar”, “Lisbon” or “Carnation Revolution”. In some cases, it may be possible to automatically obtain keywords in the first two categories by image analysis and classification.

Since in a database some images might have no text annotation, or their annotations might be considered incomplete, a significant amount of work in recent years focussed on the (semi)automatic annotation of images with keywords and on the extension of existing text annotations to images that do not have keywords associated with them (see for example [Adams 2003], [Duygulu 2002], [Lu 2000], [Kherfi 2004], [Zhang & Chen 2002], [Zhang 2005a]).

Indeed, indexing and retrieval approaches relying on keywords and visual features together are of great interest for the semantic gap reduction and were heavily investigated in the recent years, even though significant methodological advances are rare. The techniques developed in the CBIR community for query by example and relevance feedback (RF) can be directly applied to the joint use of visual features and annotations if the keywords annotating an image are represented by a vector of fixed dimension (as is usually the case for visual features), which explains why this solution was explored in the literature [Cascia 1998], [Lu 2000], [Smith 2001], [Zhang 2001], [Zhou 2002].

Various methods were proposed to obtain a feature vector representation based on the keywords annotating an image. A direct solution is to associate one dimension to every keyword that appears in the annotation of some image in the database [Lu 2000], [Zhang 2001]. A “soft” representation can also be employed [Zhou 2002], [Kherfi 2004]: the value of a feature is seen as a “degree of association” between an image and the keyword. Since the number of different keywords usually increases with the size of the database, this solution does not scale well. It also has difficulties in taking into account synonymy and homonymy (nevertheless, [Zhou 2002] suggests to use an ontology for initializing the similarities between keywords), as well as similarities between the concepts corresponding to different keywords. These problems are partly solved if latent semantic indexing (LSI) is performed on these sparse representations and the resulting low-dimensional vectors are used instead [Cascia 1998]. LSI can also be applied to the joint visual and keyword-based feature vectors, in order to find a hybrid reduced representation [Zhao 2002] that links sets of keywords and images. Unfortunately, to identify meaningful relations between keywords, LSI needs high amounts of data. This requirement can only be met when a relatively large quantity of text—rather than just a few keywords—is associated to every image.

Since in a database some images might have no annotation, or their annotations might be considered incomplete, a significant amount of work in recent years focussed on the extension of annotations from some images to the others. To establish a close relation between keywords (usually belonging to the first two categories mentioned above) and visual content, some attempts were made to model the visual appearance of images or image regions corresponding to given concepts. In [Barnard 2003], [Duygulu 2002] and [Zhang 2005a], joint statistical models are developed for the occurrence of low-level visual descriptors and keywords related either to image regions or to entire images. Supervised learning is used in [Adams 2003], [Smith 2001] for obtaining models (Markov models or support vector machines) of the "visual content" of "atomic concepts" that can be objects, scenes or events and are associated to keywords. In [Mezaris 2004], descriptions of image regions are directly associated to user provided rough visual descriptions—in terms of colour, position, size, shape—of concepts in an ontology.

Other approaches to the extension of annotations are based on relevance feedback. Indeed, by marking several images as "relevant" during a RF session, the user usually defines a similarity between these images that goes beyond what can be directly obtained from low-level visual features. Considering that this similarity is related to the presence of common keywords in the annotations of some images marked as "relevant", several authors reinforce the links between these keywords and the images top ranked by RF [Lu 2000], [Zhang 2001], [Zhou 2002], [Kherfi 2004]. A relation between the keywords and the images is thus gradually developed. In [Zhang & Chen 2002] the extension of annotations is combined with an active learning method that identifies the most ambiguous non annotated images and asks the user for appropriate annotations. The use of the feature vectors produced by such methods for extending annotations helps dealing with the difficulties created by synonymy and homonymy.

5 Supervised Learning for Document Annotation

5.1 Introduction

Supervised Learning is a machine learning technique for creating a function from training data. The training data are pairs of objects, which are typically *feature* vectors, and the desired outputs. The output of the function can be a continuous value (called regression), or can predict a *class* label of the input object (called classification).

The designer of a supervised learning mechanism for multimedia document annotation has to decide upon three crucial issues:

a. *Feature Selection and Representation*. The designer has to select a set of appropriate features and a consistent representation of them that will be propagated to the training/testing procedure. These features depend on the medium (e.g. colour histograms, edges, corners for images, pitch, formants, spectrum for audio, colour and motion descriptors for video) and the desired characteristics to be learned from the model (e.g. interdependencies among features and/or regions).

b. *Class model representation.* The next task of the designer is to adopt a representation scheme for the models that one-to-one correspond to and identify each *class.* There are two general types of supervised learning methodologies: the *generative* and the *discriminative* methods. These were introduced briefly in Section 4.1.3.

   Generative Methods learn a model of the joint probabilities, *p(x,y)*, of the inputs *x* and the labels *y*, and make their predictions by using Bayes rules to calculate *p(y|x)*, and then picking the most likely label. Such methods reduce to the use of generic classification methods like *Gaussian Mixtures, Naive Bayes, Hidden Markov Models,* and *Bayesian Networks*.

   Discriminative methods on the other hand, model the posterior *p(y|x)* directly, or equivalently, learn a direct map from inputs *x* to class labels. SVMs, k-NN, neural networks, LDA, and K-means are some of the generic classifiers employed by the discriminative methods. The main advantage of discriminative methods is the direct minimization of a classification-based error function, which typically leads to superior classification results [Ng & Jordan 2001]. Additionally since these methods are model-free, they are usually computationally efficient.

c. *Classifier Selection for training and testing*. Once feature and class model representations are selected, it is very essential to choose an appropriate classifier of the discriminative or generative family. The efficiency and effectiveness of those classifiers highly depend on the characteristics of the data to be classified. There is no single classifier that works better than the others on every problem and special care is needed in determining the most appropriate one.

In the remaining part of Section 5, a brief review of widely used generic classifiers is presented. With reference the classification scheme used, a presentation of state of the art supervised document annotation techniques completes the section. These techniques are grouped by (a) the modality under classification (image, video, audio) and (b) the feature selection and representation method.

5.2 Image and Video Classification / Annotation via Supervised Learning Techniques

In this section we review recent state-of-the-art techniques that use supervised learning in order to perform multimedia document classification. The reason for treating classification and annotation as one is that the techniques subsequently described try to semantically annotate a document by assigning to it the label of the corresponding class.

5.2.1 Global models

A framework that hierarchically classifies vacation images is proposed in [Vailaya 2001]; images are first classified as indoor or outdoor; outdoor images are further classified as city or landscape pictures which are then classified into sunset, forest and mountain classes. For the first classification, first and second order moments in the LUV colour space are used. Images are divided into 10x10 blocks from which six features are extracted (three means and three standard deviations). The second classification uses edge direction information and the third uses colour features in the HSV and LSV colour space (histograms and coherence vectors). The system uses Bayesian formalization where distributions are learned with vector quantization. A very similar approach that uses k-NN clustering was proposed in [Vailaya 1998].

In [Yavlinsky 2005] the problem of automatically annotating images with the use of global features is addressed. Images are represented either as real-valued vectors of features or signatures. Global features attempt to model image densities through the distribution of pixel colour in CIE space and a subset of perceptual texture features, resulting in a 24- dimensional feature vector. Density estimation is achieved with kernel smoothing [Parzen 1962], using either Gaussian kernel or a kernel based on the Earth Mover's Distance [Rubner 1998].

Liu et al [Liu 2005] propose a general statistical learning method based on boosting algorithm to perform image classification for photograph annotation and management. The proposed method employs both features extracted from images that are divided into NxN blocks (i.e., colour moment and edge direction histogram) and features from the EXIF metadata recorded by digital cameras. They incorporate linear discriminate analysis (LDA) algorithm to implement linear combinations between selected features and generate new combined features. The combined features are used along with the original features in boosting algorithm for improving classification performance.

5.2.2 Bag-of-words models

A simple approach to classifying images is to treat them as a collection of regions, describing only their appearance and ignoring their spatial structure. Similar models have been successfully used in the text community for analyzing documents and are known as *bag-of-words* models, since each document is represented by a distribution over fixed vocabulary(s).

Early "bag of words" models were used mostly for texture recognition. In [Leung 2001], images from a set of training materials are used to learn a vocabulary which can characterize all materials. The vocabulary consists of 3D textons, which are tiny surface patches with associated local geometric and photometric properties. Textons are then clustered using the k-means (discriminative method). A similar approach is adopted in [Varma 2002]. In this work, texture is modeled as a distribution over textons, but in this case, clustering (k-means) is an extremely low dimensional space and rotationally invariant, as in [Schmid 2001] and texture classification is performed from a single image. Lazebnik et al [Lazebnik 2003] also use texture images that are modeled as sets of regions. Each region is described by an intensity descriptor that is invariant to affine geometric and photometric transformations, based on spin images, introduced by Johnson and Herbert [Johnson & Hebert 1999]. Clustering is performed on these descriptors in a discriminative way, using a standard agglomerative algorithm. The distribution of the descriptors is then summarized in a form of a signature consisting of cluster centres and relative weights indicating the size of the clusters. Applying the Earth Mover's Distance to signatures, a distance matrix that is used for classification and retrieval is constructed.

Recently, "bag of words" models have made great progress in object categorization. A generative bag of words method was proposed in [Csurka 2004]. In this work image patches around SIFT points are located. These patches correspond to a vector of votes on predefined (from training) key point list called "bag of keypoints". Keypoints are actually the centres of the patches clusters obtained by k-means. The model uses either Naïve Bayes or multiple SVMs to classify "bag of keypoints" to categories. SVMs here are used in a generative way, since they select a class according to the best output of the algorithm. Another generative approach to the problem is described in [Blei 2004a], where each document is described by a set of words and the method tries to discover common usage patterns or "topics" in the documents and to organize these topics into a hierarchy. Hierarchies are tree structures in where each node is associated with a topic, which is a distribution across words. A document is then generated by choosing a path from the root to the leaf. Learning is performed by a combination of hierarchical latent Dirichlet allocation with the nested Chinese Restaurant Process (CRP [Aldous 1985]). This method is generic and can be applied to any type of multimedia document. Finally, in [Teh 2006] images are modeled as distributions over a visual dictionary. The model proposed here is an extension of generative "bag of words" that takes advantage of the interdependency of local image patches via a linkage structure that enforces semantic connection between patches was proposed in [Wang 2006]. This is achieved with the use of a variation of the Hierarchical Dirichlet Process (HDP) [Teh 2006] called dependent HDP (DHDP).

5.2.3 Part-based models

Another category of machine learning techniques used for multimedia annotation is part-based models. Such models can capture the essence of most object classes, since they represent both parts' appearance and invariant relations of location and scale between the parts. Part-based models are somewhat resistant to various sources of variability such as within-class variance, partial occlusion and articulation, and they are potentially convenient for indexing in a more complex system [Leibe 2004], [Lowe 2001].

While a part-based model aspires to represent a rigorous geometric relationship among the different parts, it suffers a computational difficulty of having to search among exponentially large number of hypotheses to solve the correspondence problem [Fergus 2005].

Part-based approaches to object class recognition can be crudely divided into following three types:

5.3.3.1 Generative part-based models

A study of the degree to which additional spatial constraints improve recognition performance and the tradeoff between representational power and computational complexity is examined in [Crandall 2005] through the k-fans model. This model represents both appearance and spatial relationships in a graph structure where k denotes the number of spatial dependencies among different parts. Supervised learning is generatively performed via a classical maximum likelihood procedure. This work shows that using more powerful models does not necessary improve classification, as it can lead to over-fitting during learning.

In [Leibe 2004], training is an agglomerative procedure that is performed in two stages: In the first stage, a codebook of local appearance that contains information on which local structures may appear on objects of the target category is learned. In the second stage, learning of an implicit shape model that specifies where on the object the codebook entries may occur defines allowed shapes implicitly in terms of which local appearances are consistent with each other.

Felzenswalb and Huttenlocher in [Feltzenswalb 2005] use pictorial structures introduced in [Fischler & Elschlager 1973] but they use a statistical formulation. Pictorial structures model an object as a collection of parts in a deformable configuration represented as spring like connections among pairs of parts. The best match of such a model to an image is found by minimizing an energy function that measures both a match cost for each part and a deformation cost for each pair of connected parts. The model is learned by a simple maximum likelihood estimation procedure and it is capable of locating multiple instantiations of an object in an image.

In [Fergus 2005-2], an object is represented in a “star graph” in which the location of the model part is conditioned on the location of a landmark part. In the star model any of the leaf (i.e. non-landmark) parts can be occluded, but the landmark part must always be present. The star model also provides benefits in that it has less parameters so that the model can be trained on fewer images without over fitting occurring. Their formulation captures scale and occlusion and the model is learned with EM.

In [Sudderth 2005], features are parts corresponding to SIFT points and their location relative to the object. The model proposed learns conditional probabilities on SIFT values and locations. This is achieved with a combination of Gibbs sampling and EM.

5.3.3.2 Discriminative part-based models

Berg et al [Crandall 2005] proposed an interesting method for shape matching and object recognition across images. Their model first computes correspondences between feature points of different images and then estimates an aligning transform, typically a regularized thin plate spline, resulting in a dense correspondence between the two shapes. Object recognition is then handled in a nearest neighbour framework where the distance between exemplar and query is the matching cost between corresponding points.

In [Opelt 2004] and [Opelt 2004-2], the learning algorithm is provided with a set of labelled images where a positive label indicates that a relevant object appears in the image. Objects are not segmented and pose and location are unknown. The image analysis transforms images to grey values and extracts normalized regions around interest (salient) point to obtain reduced representations of images. As an appropriate representation for the learning procedure, local descriptors of these patches are calculated. Classification is achieved with AdaBoost algorithm.

Vidal et al in [Vidal 2003] apply linear classification rules to informative features to achieve good classification results. Such features are obtained by a two step procedure: (i) an optimal threshold determines the minimum visual similarity so as to locate informative image fragments and (ii) a greedy algorithm adds fragments iteratively to the set of informative features until adding more fragments no longer increases the estimated information content of the set. Features are obtained by wavelet transform that captures frequency and orientation properties, and quantization. For classification, linear SVMs and Tree-Augmented Networks are tested.

An extension of the boosting algorithm that is based in gentleboost [Kohavi & John 1997] is proposed in [Torralba 2004]. Learning is performed on large feature vectors that correspond to regions in the image. The proposed algorithm trains a number of binary classifiers jointly instead of training them jointly, so that they share as many features as possible.

5.3.3.3 Hybrid part-based models

Recently there has been systematic effort to combine generative and discriminative methods into hybrid methods that combine these approaches. In [Holub & Perona 2005], Holub and Perina have developed Fisher kernels based on the constellation model. For every input, Fisher kernel method calculates the Fisher score of the input, and Support vector Machine (SVM) is applied to classification in the Fisher score space. Another path towards generative and discriminative classification is through boosting. Bar-Hillel et al. [Hillel 2005] has presented a boosting-based method based on their own generative model, which, similar to the constellation model, models part relations as a global distribution function. Finally, in [Zhang 2005b], Bayesian Classification is combined with a new method for generative part-based object modelling called Random Attributed Relational Graph (RARG) that captures the advantages of the pictorial structure model [Crandall 2005] and the constellation model since it accommodates part occlusion and the partial relationship among object parts. RARG is a variation of the ARG [Zhang 2005a] model that extends the ordinary graph in graph theory by attaching discrete or continuous feature vectors to the vertices and edges of the graph. Image similarity therefore can be defined by the corresponding ARG similarity.

5.3 Audio Annotation via Supervised Learning Techniques

5.3.1 Supervised learning for structural annotation

After segmenting an audio stream into homogenous segments – e.g. using the model-free Bayesian Information Criterion approach [Chen 1998] – the segments must be further classified to enrich the document annotation. Supervised learning strategies based on Gaussian Mixture Models (GMMs) can be applied to detect the nature of an audio segment. In [Biatov 2006], GMMs are used to discriminate speech from non-speech segments. Gender detection using GMMs has been successfully carried out in rich transcription systems such as [Gauvain 1998]. In [Barras 2006], GMMs are used to detect the transmission channel of a speech segment.

The information extracted by the GMM classifiers can be used to select well suited acoustic models for spoken language annotation, e.g. a model trained only on data from female speakers recorded over a fixed telephone line. This approach has been already applied by the early systems for rich transcription of broadcast data [Gauvain 1998, Woodland 1998].

5.3.2 Supervised learning for spoken language annotation

Most state of the art systems for audio indexing apply a holistic statistical approach to spoken language annotation [Gales 2006, Matsoukas 2006, Chen 2006, Stolcke 2006]. Besides static information such as pronunciation lexica, audio indexing integrates acoustic and linguistic knowledge sources that require supervised training strategies:

1. *Acoustic models* must be trained to model the acoustic properties of the triphones.
2. *Language models* must be trained to model typical word sequences.

5.4.2.1 Acoustic modelling

The context dependent triphones that are used as sub-word units are typically modelled by Hidden Markov Models (HMMs).

Two main paradigms are currently applied for estimating the HMM mean and variance parameters. Maximum Likelihood Estimation (MLE) using Expectation-Maximization [Bilmes 1998] has been the standard for many years. More recently, discriminative criteria such as the Maximum Mutual Information Estimation (MMIE) have been investigated in the area of large vocabulary continuous speech recognition (LVCSR), yielding significant word error rate reductions compared to the maximum likelihood approach [Woodland 2000].

5.4.2.2 Language modelling

M-gram statistical language models are used in both whole word [Gales 2006, Matsoukas 2006, Chen 2006, Stolcke 2006] and sub-word indexing approaches [Larson 2003]. They are often trained using the Maximum Likelihood paradigm, where the ML estimate of an m-gram is estimated on a large textual training corpus. Typically, trigrams or even 4-grams are estimated for language modeling. Due to the high number of unseen m-grams, smoothing methods as reviewed in [Chen 1996] must be applied to redistribute the probability mass.

6 Unsupervised Clustering for (Direct) Feature Indexing

We include this section for completeness only. Although some successes have been reported in combining supervised and unsupervised machine learning techniques for various sorts of cross-media data [Tsai 2003], or straightforward clustering [Chen-Y 2005], we could identify no systems which offer the scale of operation needed for VITALAS.

7 Conclusions

At the present time the production of cross-media document signatures and indexes is a very active area. Significant successes have been reported using various forms of supervised Machine Learning techniques and methods of combining evidence from different sources and modalities to produce single integrated indexes. [Liu 2007] list a sequence of steps that can be applied to narrow down the semantic gap as follows: 1. Using an object ontology to define high level concepts. 2. Supervised or unsupervised ML methods to associate low-level features with query concepts. 3. Relevance feedback for continuous learning of users' intentions. 4. Semantic templates, enabling high level retrieval through low level user-selected icons are not widely used but promising. 5. The web is a large data source, and a combination of the other four techniques can be used to derive training data from the Web. The first three of these steps should prove particularly useful to VITALAS.

7.1 Reasons for the Choice of Supervised Learning in the VITALAS Project

Early approaches aimed to map low level features, such as colour, pitch or term frequency, directly onto high-level semantic concepts. However, this approach does not scale up to the large-scale automatic annotation of video archives. Supervised learning is the "present-day paradigm of choice in generic video indexing" [Snoek 2006]. Machine learning techniques may involve supervised learning, where for example we predict the correct semantic category label from a predetermined set, based on the values of the set of input features describing an image. With unsupervised learning, the set of possible labels is not known beforehand and the aim is to describe how the input data are organised internally or clustered. Supervised learning is the more suited to VITALAS, since through reference to ontologies we will know the range of high level concept categories pertinent to the collection, to which images must be mapped.

A number of supervised learning techniques, such as Support Vector Machines (SVM) have sound mathematical foundations, and have been successfully used for image annotation (Shi et al., 2004). They were originally developed for binary classification, but can be used to learn multiple concepts for image retrieval through a series of "one versus the rest" classifications. Bayesian classifiers are also widely used, such as for the binary classification of indoor versus outdoor scenes. Both SVMs and Bayesian classifiers are presented as viable alternatives by [Adams 2003], whose system will form the basis of the VITALAS state-of-the art cross-media indexing system of D 3.1.2. We will not be using neural networks used for supervised learning, since they require much training data and are computationally expensive, particularly when we have a large set of output nodes corresponding to concept categories.

Supervised learning in general has two main disadvantages: Firstly a large amount of labelled training data is needed, and it is time-consuming, expensive and error prone to produce such data. Secondly, systems are trained for one particular application domain. If we need to change the domain, the system has to relearn using new training data. The first problem is overcome by the availability of publically-available large annotated video training sets such as MediaMill, which is part of the TRECVID data set. For VITALAS, MediaMill will solve the problem of having an initial set of training data before the training data provided by the content partners (Belga and INA) becomes available. It also provides a shared ground truth against which we can compare our approaches with published research. As both training and test data, MediaMill provides both low-level features and their corresponding manually-labelled ground truth semantic concepts (the two extremes of the semantic gap) drawn from a lexicon of 101 predefined semantic concepts based on an analysis of a large number query logs. A requirement for these concepts was that they should be clear from looking at a single still image, and be present throughout the camera shot. As well as visual data, the MediaMill data set contains English text produced from English speech by a speech-to-text transcriber, and also English text translated and transcribed from speech originally in Arabic and Chinese [Snoek 2006]. Anecdotally, previous users of this data report that the annotations are less than perfect, but there is no comparable data set available. The provision of ready-annotated low-level features, thus overcoming the need for expensive multimedia processing, is indeed an advantage of using MediaMill at this early stage, but within VITALAS there are partners capable of providing such feature extraction, which will take place in WP2.

With regard to the problem of changing domains, the VITALAS domain is not a fixed one. The INA data contains an archive of many kinds of television programs. News videos are an important part of this archive, but they are by no means the only ones. Similarly, most of the Belga photographs capture current affairs, but there are many so-called creative pictures. It may be necessary to learn annotations separately for each of these image domains.

7.2 Extending the Range of Concepts and Domains

SML techniques involve the matching of low level image features against high level query concepts. A vocabulary of qualitative definitions of high level concepts is called an object ontology. For small collections with a limited set of concepts in well defined domains, the objects in the ontology may be semantically very simple, and correspond exactly with quantised image descriptors such as "light green" or "medium coarse". However with large collections of images with various contents, more elaborate ontologies are required to reflect the range of concepts represented by the collection (Liu et al., 2006). Deriving an ideal vocabulary representing the rich semantics of an image collection is a difficult task, but in the case of VITALAS, certain partners such as INA already possess hand-built ontologies representing the concepts in their collections. We are also interested in the possibility of creating ontologies automatically, starting for example with an automatically generated vocabulary list derived from the text captions of the images in the Belga collection, later augmented by qualitative descriptors derived from the content of the images. A third possibility, which should prove particularly useful for other work packages involving word sense disambiguation and cross-language information retrieval, is to use WordNet, a humanly-created ontology based on psycholinguistic principles, which organises English words into about 115,000 synonym sets each representing an underlying semantic concept. EuroWordNet, derived from WordNet, contains similar but smaller vocabularies for other languages including French and German (22,745 and 15,132 respectively) .

It is believed that humans can recognise about 5000-30000 object categories [Liu 2007], and our initial target is that the VITALAS search engine should be able to cater for about 1000 cross-media concepts. Depending on how successful we are, we might later aim for about 3000. Object category learning with such large vocabularies has not yet been achieved. To date, the largest vocabularies used in image classification have been the 101 categories achieved by [Fei-Fei 2004] using a Bayesian approach, and at the University of Sunderland by [Tsai 2006b] who used a two-stage approach to classify images into one of 200 categories. The first stage was an SVM classifier based on colour and texture, followed by an inference module based on fuzzy logic which made final decisions on which high-level concept to choose. Currently Wei-Chao Lin, also at the University of Sunderland, is exploring the problem of extending vocabularies for image classification using the statistical measure of Information Gain and the machine learning technique of boosting. The LSCOM (http://www.lscom.org) annotation consists of over 300 keyframe-based labels for visual concepts in video [Hauptmann 2007].

An information retrieval technique often combined with object-ontology and machine learning is Relevance Feedback (RF), which has been described in the context of content-based image retrieval (CBIR) in a review by [Zhu & Huang 2003]. RF enables to user to examine images retrieved in response to an initial query, and provide feedback on these to the system by specifying which of these images are relevant and which are not relevant to the query. The system uses this information to make a second search, this time for images similar to those declared "relevant", and avoiding images declared "not relevant". In its simplest case, RF is available through Google's "More Like This" facility, but research systems such as Okapi use more elaborate RF algorithms. RF provides an additional source of information which assists in the mapping of low-level features to high-level semantic concepts, which will be incorporated into the final VITALAS search engine as detailed in WP 4.2.1.

7.3 Complementing Purely Machine Learning Approaches with Human Knowledge

To overcome some of the challenges of creating a multi-media system which runs to 1000 concepts, we could adopt some sort of human knowledge to complement our purely machine-learning approaches. [Adams 2003] include human knowledge in their system by stipulating manually which media contribute features for the recognition of a high-level concept. For example, the indexer might encode the fact that a rocket launch is characterised by certain non-speech audio sounds and visual features, but not by speech or text in captions. Human knowledge will be needed in any semi-automatic approaches we may take in the production of an object ontology, since purely automatic approaches to ontology generation are still less than perfect. On the other hand, if we find that (Euro) WordNet is adequate for our purposes, human knowledge derived from psycholinguistic experiments will already encoded in the structure. Another form of human-knowledge, not required in the construction of the system but during the execution of image retrieval is relevance feedback, where human judgements on intermediate search results can guide the overall search process.

8 References

[Adams 2003] W. H. Adams and Giridharan Iyengar and Ching-Yung Lin and Milind R. Naphade and Chalapathy Neti and Harriet J. Nock and John R. Smith, Semantic Indexing of Multimedia Content Using Visual, Audio and Text Cues, EURASIP Journal on Applied Signal Processing, Volume 3, pages 170-185, 2003.

[Aizerman 1964] M.A. Aizerman, E.M. Braverman, and L.I. Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 25:821–837, 1964.

[Akrivas 2004] G. Akrivas, G. B. Stamou and S. Kollias. Semantic Association of Multimedia Document Descriptions Through Fuzzy Relational Algebra and Fuzzy Reasoning. IEEE Transactions on Systems, Man and Cybernetics, Vol 34(2), 2004.

[Albiol 2002] Alberto Albiol, Luis Torres, Edward J. Delp, Combining Audio and Video for Video Sequence Indexing Applications, Proc. of Intl Conf. on Multimedia and Expo 2002, , vol.2, pp. 353- 356 2002.

[Aldous 1985] D. Aldous. Exchangeability and related topics. In E´cole d'e´te´ de probabilite´s de Saint-Flour, XIII—1983, pages 1–198. Springer, Berlin, 1985

[Amir 2003] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M. Naphade, A. Natsev, C. Neti, H. Nock, J. R. Smith, B. L. Tseng, Y. Wu, D. Zhang, IBM Research TRECVID-2003 Video Retrieval System, NIST TREC-2003 Text Retrieval Conference, Gaithersburg, MD, November 2003.

[Amir 2004] A. Amir, J. Argillander, M. Berg, S-F. Chang, M. Franz, W. Hsu, G. Iyengar, J. R Kender, L. Kennedy, C-Y. Lin, M. Naphade, A. Natsev, J. R. Smith, J. Tesic, G. Wu, R. Yan, D. Zhang. IBM Research TRECVID-2004 Video Retrieval System. In Proc. of NIST TRECVID 2004, NIST, Gaithersburg, 2004.

[Assfalg 2003] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, W. Nunziati, Semantic Annotation of Soccer Videos: Automatic Highlights Identification, Computer Vision and Image Understanding, Vol. 92(2) pp. 285-305, 2003.

[Bach 1996] J.R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain and C.F. Shu. The Virage image search engine: An open framework for image management. In Proc. SPIE Storage and Retrieval for Still Image and Video Databases, 1996.

[Barras 2006] C. Barras, X. Zhu, S. Meigner, J. L. Gauvain. Multistage speaker diarization of broadcast news, IEEE Transactions on Audio, Speech and Language Processing Vol. 14, No. 5. (2006), pp. 1505-1512.

[Barnard 2003] Barnard Kobus, Duygulu Pinar, de Freitas Nando, Forsyth David Blei, David and Jordan Michael I., Matching Words and Pictures, Journal of Machine Learning Research, Vol 3, pp 1107-1135, 2003.

[Bertini 2005a] M. Bertini, R. Cucchiara, A. Del Bimbo, C. Torniai. Video Annotation with Pictorially Enriched Ontologies. In Proc IEEE Intl. Conf. on Multimedia and Expo (ICME 2005), 2005.

[Bertini 2005b] M. Bertini, A. Del Bimbo, C. Torniai. Enhanced Ontologies for Video Annotation and Retrieval. In Proc 7th ACM SIGMM Workshop on Multimedia Information Retrieval, 2005.

[Belew 2000] Belew R. K. (2000). Finding Out About [Cambridge Univ. Press] ISBN 0-521-63028-2.

[Biatov 2006] Biatov K., Koehler J., Improvement Speaker Clustering Using Global Similarity Features, Intership 2006 – ICSLP, Pittsburgh, PA, USA

[Bilmes 1998] J. Bilmes. A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Technical Report, University of Berkeley, ICSI-TR-97-021, 1997.

[Bingham & Mannila 2001] Bingham E. and Mannila H. Random projection in di-mensionality reduction : applications to image and text data. In Knowledge Discovery and Data Mining, pages 245-250, 2001.

[Blei 2003] David M. Blei, Andrew Y. Ng and Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, pp 993-1022, 2003

[Blei 2004a] DM Blei, TL Griffiths, MI Jordan, JB Tenenbaum: Hierarchical Topic Models and the Nested Chinese Restaurant Process, Advances in Neural Information Processing Systems, 2004.

[Blei 2004b] Blei D. Probabilistic Models of Text and Images. PhD thesis, U.C. Berkeley, Division of Computer Science, 2004.

[Blei & Jordan 2003] Blei D. and Jordan M. Modeling annotated data. In Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 127–134. ACM Press, 2003.

[Boser 1992] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144-152, Pittsburgh, PA, 1992. ACM Press.

[Brautigam 1998] C. G. Brautigam, J. O. Eklundh and H. I. Christensen. A Model-Free Voting Approach for Integrating Multiple Cues. In Proc. 5th European Conference on Computer Vision (ECCV'98), 1998.

[Breiman 1984] Breiman, L., Friedman, J., Olshen, R. and Stone, C: Classification and Regression Trees, Monterey, CA: Wadsworth, 1984

[Brin 1998] Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the Seventh international Conference on World Wide Web 7 (Brisbane, Australia). P. H. Enslow and A. Ellis, Eds. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 107-117. DOI= http://dx.doi.org/10.1016/S0169-7552(98)00110-X

[Burges 1996] C.J.C. Burges: Simplified support vector decision rules, In Lorenza Saitta, editor, Proceedings of the Thirteenth International Conference on Machine Learning, pages 71–77, Bari, Italy, 1996. Morgan Kaufman.

[Campbell 2006] M. Campbell, A. Haubold, S. Ebadollahi, D. Joshi, M. Naphade, A. Natsev, J. Seidl, J. R. Smith, K. Scheinberg, J. Tesic, L. Xie. IBM Research TRECVID-2006 Video Retrieval System. In Proc. of TREC Video Retrieval Evaluation, 2006.

[Carson 2002] Carson C., Belongie S., Greenspan H. and Malik J. Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(8):1026–1038, 2002.

[Cascia 1998] La Cascia, M. and S. Sethi and S. Sclaroff, Combining textual and visual cues for content-based image retrieval on the world wide web, IEEE Workshop on Content-Based Access of Image and Video Libraries, pages 24-28, 1998.

[Castelli 2003] Castelli V., Thomasian A. and Chung-Sheng Li. CSVD: clustering and singular value decomposition for approximate similarity search in high-dimensional spaces. IEEE Transactions on Knowledge and Data Engineering. Vol. 15, no. 3, pp. 671-685. May-June 2003.

[Chang 2005] Shih-Fu Chang, R. Manmatha, and Tat-Seng Chua, Combining Text and Audio-Visual Features in Video Indexing, Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05), Vol. 5, pp. 1005-1008, March 2005.

[Chen 1996] S. F. Chen, J. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling, In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL). 1996

[Chen 1998] Chen S. S., Gopalakrishnan P. S., Clustering via the Bayesian information criterion with applications in speech recognition, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2 (1998), pp. 645-648 vol.2.

[Chen 2005] Xin Chen, Chengcui Zhang, Shu-Ching Chen, Min Chen A Latent Semantic Indexing Based Method for Solving Multiple Instance Learning Problem in Region-Based Image Retrieval. Seventh IEEE International Symposium on Multimedia (ISM'05) pp. 37-45.

[Chen 2006] Chen, S.F. Kingsbury, B. Lidia Mangu Povey, D. Saon, G. Soltau, H. Zweig, G. Advances in speech transcription at IBM under the DARPA EARS program. IEEE Transactions on Audio, Speech and Language Processing,Vol. 14, No. 5. (2006), pp. 1596-1608.

[Chen-Y 2005] Yixin Chen, Wang, J.Z. Krovetz and R. CLUE: cluster-based retrieval of images by unsupervised learning,IEEE Transactions on Image Processing, Volume: 14, Issue: 8, page(s): 1187- 1201, Aug. 2005.

[Cleverdon, 1967] Cleverdon C. W. (1967), The Cranfield tests on index language devices. Aslib Proceedings, 19:173–192.

[Comaniciu & Mea 1999] D. Comaniciu and P. Mea, Distribution free decomposition of multivariate data, Pattern Analysis and Applications, 2, 22. 30,1999.

[Crandall 2005] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In Proc. Computer Vision and Pattern Recognition, 2005.

[Csurka 2004]G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and Bray C. Visual categorization with bags of keypoints. In Proc. European Conference on Computer Vision, 2004.

[Dasarthy 1991] Dasarthy, B: Nearest Neighbor Pattern Classification Techniques, IEEE Computer Society Press, 1991

[Dasarthy 1994] Dasarathy B.V. Decision Fusion, IEEE Computer Society Press, Los Alamitos, CA, ISBN 0-8186-4452-4, 1994

[Dasiopoulou 2005] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis and M. G. Strintzis. Knowledge – assisted semantic video object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2005.

[Demiriz 2002] A. Demiriz and K.P. Bennett and J. Shawe-Taylor. Published 2002 in Kluwer Machine Learning 46, pages 225–254.

[Dempster 1977] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. B. 39, 1, 1–38.

[Domingos 1996] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In Proceedings of the Thirteenth International Conference on Machine Learning, pages 105–112. Morgan Kaufmann Publishers, Inc., 1996.

[Domingos 1997] Domingos, Pedro & Michael Pazzani: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29:103–¬137, 1997

[Dong-Oing 2006] Dong-Qing Zhang Shih-Fu Chang. A Generative-Discriminative Hybrid Method for Multi-View Object Detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, 2006

[Dorado 2004] Andres Dorado, Janco Calic and Ebroul Izquierdo. A Rule – Based Video Annotation System. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14(5), May 2004.

[Duda & Hart 1973] Duda R.O. and Hart P.E. Pattern Classification and Scene Analysis. John Wiley & Sons, Inc., New York, 1973.

[Duda 2000] Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern Classification, Wiley-Interscience, 2nd edition. 2000.

[Duygulu 2002] Pinar Duygulu and Kobus Barnard and de Freitas, J. F. G. and David A. Forsyth, Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, Proceedings of the 7th European Conference on Computer Vision-Part IV, 2002,isbn 3-540-43748-7, Pages 97-112, Springer-Verlag.

[Eisele 1996] Eisele, T. and Umbach, Haeb R. and Langmann, D., A Comparative Study of Linear Feature Transformation Techniques for Automatic Speech Recognition, Philadelphia, PA, Proc. ICSLP 1996.

[Ekin 2003] A. Ekin, A. M. Tekalp and R. Mehrotra. Automatic Soccer Video Analysis and Summarization. IEEE Transactions on Image Processing, Vol 12(7), 2003.

[Falelakis 2005] M. Falelakis, C. Diou, A. Valsamidis and A. Delopoulos. Complexity Control in Semantic Identification. In Proc. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2005.

[Falelakis 2006] M. Falelakis, C. Diou and A. Delopoulos. Semantic Identification: Balancing between Complexity and Validity. EURASIP Journal on Applied Signal Processing, Sp. Issue on Information Mining from Multimedia Databases, Vol 2006, 2006.

[Fan 2004] J. Fan, A. K. Elmagarmid, X. Zhu, W. G. Aref and L. Wu. ClassView: Hierarchical Video Shot Classification, Indexing and Accessing. IEEE Transactions on Multimedia, Vol. 6(1), 2004.

[Fei-Fei 2006] L. Fei-Fei, R. Fergus, and P. Perona. One-Shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence, in press.

[Feltzenswalb 2005] Feltzenswalb P. and Hutenlocher D. Pictorial structures for object recognition. IJCV, 61:55-79, 2005

[Fergus 2003] R. Fergus, P.Perona, and A. Zisserman, Object Class Recognition by Unsupervised Scale-Invariant Learning, CVPR 2003.

[Fergus 2005] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. IJCV(submitted), 2005.

[Fergus 2005-2] Fergus R., Perona P., and Zisserman A. A sparse object category model for efficient learning and exhaustive recognition. In CVPR, 2005.

[Fisher 1936] Fisher, R.A: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7: 179-188, 1936.

[Fischler & Elschlager 1973] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. IEEE Transactions on Computer, 22(1):67{92, January 1973

[Fix 1951] Fix, E. and Hodges, J: Discrimninatory analysis- nonparametric discrimintation: Consistency properties, Techical Report 21-49-004,4, US Air Force, School of Aviation Medicine, Randolph Field, TX, 1951

[Flickner 1995] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafine, D. Lee, D. Petkovic, D. Steele and P. Yanker. Query by image and video content: The QBIC system. IEEE Computer, 1995.

[Freund 1997] Freund, J. B. 1997 Proposed inflow/outflow boundary condition for direct computation of aerodynamic sound. AIAA J. 35, 740-742.

[Freund 2005] Y. Freund and R. E. Schapire. A decision-theoretic generalisation of on-line learning. Computer and System Sciences, 55(1), 1997

[Fukunaga 1990] Fukunaga, K: Introduction to Statistical Pattern Recognition (second edition), Academic Press, 1990.

[Gagnon 2005] L. Gagnon, R&D status of ERIC-7 and MADIS – Two systems for MPEG-7 indexing/search of audio-visual content, Technical Report of R&D Department, Computer Research Institute of Montreal.

[Gales 2006] Gales, M.J.F. Do Yeong Kim Woodland, P.C. Ho Yin Chan Mrva, D. Sinha, R. Tranter, S.E. Progress in the CU-HTK broadcast news transcription system. IEEE Transactions on Audio, Speech and Language Processing Vol. 14, No. 5. (2006), pp. 1513-1525

[Gauvain 1998] J. L. Gauvain, L. Lamel, M. Jardino. The LIMSI 1998 Hub-4E Transcription System. DARPA Broadcast News Workshop, 1998.

[Goldenstein 2003] S. K. Goldenstein, C. Vogler and D. Metaxas. Statistical Cue Integration in DAG Deformable Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 25(7), 2003.

[Hauptmann 2004] Hauptmann, A., and Christel, M. Successful Approaches in the TREC Video Retrieval Evaluations. In Proc. ACM Multimedia (New York, October 2004), 668-675.

[Hauptmann 2007] A. Hauptmann, R. Yan and W-H Lin. How many high-level concepts will fill the semantic gap in news video retrieval? CIVR 2007, ACM International Conference on Image and Video Retrieval, July 9-11 2007, Amsterdam.

[Hayman & Eklundh 2002] E. Hayman and J. Eklundh. Probabilistic and Voting Approaches to Cue Integration for Figure-Ground Segmentation. Lecture Notes in Computer Science, Vol. 2352, Springer, 2002.

[Hellerstein 2000] J. Hellerstein, Jayram Thathachar, and I. Rish. Recognizing end-user transactions in performance management. In Proceedings of AAAI-2000, pages 596-602, Austin, Texas, 2000.

[Hillel 2005] A. B. Hillel, T. Hertz, and D.Weinshall. Efficient learning of relational object class models. In IEEE International Conference on Computer Vision (ICCV), 2005.

[Hofmann 1999a] Hofmann T. Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.

[Hofmann 1999b] Hofmann, T. Latent class models for collaborative filtering. In Proceedings of the 16th International Joint Conference on Articial Intelligence (IJCAI) (1999).

[Hofmann 2001] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. J. 42, 1, 177–196.

[Hoiem 2003] Hoiem, D., Sukhankar, R., Schneiderman, H., and Huston, L. (2003). Objectbased image retrieval using the statistical structure of images. Technical Report IRP-TR-03-13, Intel.

[Holub & Perona 2005] A. Holub and P. Perona. A discriminative framework for modeling object classes. In Int. Conf. on Computer Vision, 2005. [Howland & Park 2003] Howland P. and Park H. Cluster-preserving dimension reduction methods for efficient classification of text data. A comprehensive survey of text mining, Springer-Verlag, pp. 3–23, 2003.

[Hoogs 2003] A. Hoogs, J. Rittscher, G. Stein and J. Schmiederer. Video Content Annotation Using Visual Analysis and a Large Semantic Knowledgebase. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03), 2003.

[Huang 1997] T. S. Huang, S. Mehrotra, and K. Ramchandran, Multimedia analysis and retrieval system (MARS) project. In Proc of 33rd Annual Clinic on Library Application of Data Processing - Digital Image Access and Retrieval, 1996.

[Ianeva 2004] Ianeva T., Boldareva L., Westerveld T., Cornacchia R., Hiemstra D. and de Vries A.P. Probabilistic Approaches to Video Retrieval. In TRECVID 2004 Workshop, Gaithersburg, MD, US, November, 2004.

[Iqbal 2003] Q. Iqbal and J. K. Aggarwal. Feature Integration, Multi-image Queries and Relevance Feedback in Image Retrieval. In Proc. 6th International Conference on Visual Information Systems (VISUAL'03), 2003.

[Jeon 2003] Jeon, J., Lavrenko, V. and Manmatha, R., Automatic Image Annotation and Retrieval using Cross-Media Relevance Models, the Proceedings of SIGIR '03 Conference, pp. 119-126.

[Li & Wang 2003] J. Li and J. Z. Wang. Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 25(9), 2003.

[Joachims 1997] T. Joachims: Text categorization with support vector machines. Technical report, LS VIII Number 23, University of Dortmund, 1997.

[Johnson & Hebert 1999] A. Johnson and M. Hebert, Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes, IEEE Trans. PAMI 21(5), pp. 433-449, 1999.

[Johnson 2000] S.E. Johnson , P. Jourlin, G.L.Moore, K. Sp¨arck Jones & P.C. Woodland Audio Indexing and Retrieval of Complete Broadcast News Shows, RIAO 2000.

[Jones 1996] G.J.F. Jones, J.T. Foote, K. Sparck Jones and S.J. Young, ``Retrieving spoken documents by combining multiple index sources'', /SIGIR 96, Proceedings of 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval/, 1996, 30-38.

[Ng & Jordan 2001] Ng A.Y. and Jordan M.I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In NIPS, 2001.

[Jose & Harper 1998] Jose J.M. and Harper D.J. Epic: A Photograph Retrieval System Based on Evidence Combination Approach. Proceedings of the IFMIP-98 Conference, Anchorage, Alaska. Published as Soft Computing, Multimedia and Image Processing; Trends, Principles and Applications (Jamshidi, M. et al. Eds.) pp 777-782 TSI Press Series.

[Kherfi 2004] M.L. Kherfi and D. Brahmi and D. Ziou, Combining visual features with semantics for a more effective image retrieval, Proc. of the 17th International Conference on Pattern Recognition, 2004.

[Kittler 2001] J. Kittler, K. Messer, W. J. Christmas, B. Levienaise-Obadia and D. Koubaroulis. Generation of Semantic Cues for Sports Video Annotation. In Proc. International Conference on Image Processing (ICIP'01), 2001.

[Kohavi & John 1997] R. Kohavi and G. H. John. Wrappers for feature subset selection. Artificial Intelligence, 1, 1997

[Kokar & Tomasik 2001] Mieczyslaw M. Kokar and Jerzy A. Tomasik. Data vs. Decision Fusion in the Category Theory Framework. FUSION 2001, 2001.

[Kokar 2004] Mieczyslaw M. Kokar, Jerzy A. Tomasik and Jerzy Weyman. Formalizing classes of information fusion systems. Information Fusion, Vol. 5(3), Sept. 2004, pp. 189-202.

[Kompatsiaris 2001] I. Kompatsiaris, E. Triantafillou, and M. G. Strintzis, Regionbased colour image indexing and retrieval, in Proc. 2001 International Conference on Image Processing, vol. 1, pp. 658–661, Thessaloniki, Greece, October 2001.

[Koskela 2006] M. Koskela and J. Laaksonen. Semantic Concept Detection from News Videos with Self - Organizing Maps. In Proc. 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI'06), 2006.

[Kuncheya 2002] L. I. Kuncheva. A Theoretical Study on Six Classifier Fusion Strategies. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 24(2), 2002.

[Kushki 2004] A. Kushki, P. Androutsos, K. Plataniotis and A. N. Venetsanopoulos. Retrieval of Images From Artistic Repositories Using a Decision Fusion Framework. IEEE Transactions on Image Processing, Vol 13(3), 2004.

[Laaksonen 2002] J. Laaksonen, M. Koskela and E. Oja. PicSOM – Self – Organizing Image Retrieval With MPEG-7 Content Descriptors. IEEE Transactions on Nerual Networks, Vol 13(4), 2002.

[Lafon 2006] Stephane Lafon, Yosi Keller and Ronald R. Coifman. Data Fusion and Multicue Data Matching by Diffusion Maps, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 28(11), Nov. 2006, pp. 1784-1797.

[Lalmas 1997a] Lalmas M. Dempster-Shafer's theory of evidence applied to structural documents: modelling uncertainty ACM SIGIR Forum vol. 31, pp110- 118, 1997.

[Lalmas 1997b] Lalmas M., Ruthven I. and Theophylactou M. Structured document retrieval using D-S's theory of Evidence: Implementation and Corrections Technical Report, University of Glasgow, Aug 1997.

[Larson 2003] M. Larson, S. Eickeler. Using syllable-based indexing features and language models to improve german spoken document retrieval. In European Conference on Speech Communication and Technology, 2003

[Lavrenko 2003] Lavrenko, V., Manmatha, R. and Jeon, J., A Model for Learning the Semantics of Pictures, the Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems, vol. 16, pp. 553-560.

[Lavrenko 2004] Lavrenko, V., Feng, S. and Manmatha, R., Statistical Models for Automatic Video Annotation and Retrieval, the Proceedings of the International Conference on Acoustics, Speech and Signal Processing,(ICASSP), Montreal, QC, Canada.

[Lazebnik 2003] S. Lazebnik, C. Schmid, and J. Ponce. Sparse texture representation using affine-invariant neighborhoods. In Proc. CVPR, 2003.

[Leibe 2004] Leibe B., Leonardis A., and Schiele B. Combined object categorization and segmentation with an implicit shape model. In ECCV workshop on statistical learning in computer vision, 2004.

[Leung 2001] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. IJCV, 43(1):29–44, June 2001.

[Li 2000] J. Li, R.M. Gray, and R.A. Olshen, Multiresolution Image Classification by Hierarchical Modeling with Two Dimensional Hidden Markov Models, IEEE Trans. Information Theory, vol. 46, no. 5, pp. 1826-41, Aug. 2000.

[Lin 2003] C. Lin, B. L. Tseng, M. Naphade, A. Natsev and J. R. Smith. VideoAL: A Novel End-to-End MPEG-7 video automatic labeling system. In Proc. International Conference on Image Processing (ICIP’03), 2003.

[Liu 2005] Xuezheng Liu, Lei Zhang, Mingjing Li, HongJiang Zhang & Dingxing Wang. Boosting image classification with LDA-based feature combination for digital photograph management. Pattern Recognition, vol. 38, no. 6, pages 887-901, 2005.

[Liu 2007] Y. Liu, D. Zhang, G. Lu, and W-Y Ma. A survey of content-based image retrieval with high-level semantics. Pattern Recognition 40(1):262-282, 2007.

[Lloyd 1957] Lloyd, S: Least squares quantization in PCM, Technical report, Bell Laboratories. Published in 1982 in IEEE Trans. Inf. Theory 28, 128-137, 1957.

[Lowe 2001] Lowe D. Local feature view clustering for 3d object recognition. In CVPR, pages 682-688, 2001.

[Lu 2000] Ye Lu and Chunhui Hu and Xingquan Zhu and Hong-Jiang Zhang and Qiang Yang, A unified framework for semantics and feature based relevance feedback in image retrieval systems, Proceedings of the 8th ACM International Conference on Multimedia, 2000, isbn 1-58113-198-4, pages 31-37, Marina del Rey, California, USA, ACM Press.

[Luo 2003] Hangzai Luo, Jianping Fan, Jing Xiao, and Xingquan Zhu. Semantic principal video shot classification via mixture gaussian. In IEEE International Conference on Multimedia and Expo (ICME), 2003.

[Ma 1997] W. Y. Ma and B. S. Manjunath. Netra: A toolbox for navigating large image databases. In Proc. IEEE International Conference on Image Processing (ICIP), 1997.

[Marchand-Maillet and Worring 2006] Stéphane Marchand-Maillet and Marcel Worring Benchmarking image and video retrieval: an overview.Proceedings of the 8th ACM international workshop on Multimedia information retrieval, International Multimedia Conference, Pages: 297 - 300,ISBN:1-59593-495-2, 2006.

[Mason 2000] L. Mason, P. L. Bartlett, and J. Baxter. Improved generalization through explicit optimization of margins. Machine Learning, 38(3):243–255, Mar. 2000.

[Matsoukas 2006] Matsoukas, S. and Gauvain, J. L. and Adda, G. and Colthurst, T. and Chia-Lin Kao Kimball, O. and Lamel, L. and Lefevre, F. and Ma, J.Z. and Makhoul, J. and Nguyen, L. and Prasad, R. and Schwartz, R. and Schwenk, H. and Bing Xiang. Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system. IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 5. (2006), pp. 1541-1556.

[McDonald & Tait 2003] Sharon McDonald and John Tait Search Strategies in Content-Based Image Retrieval Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), Toronto, July, 2003. pp 80-87. ISBN 1-58113-646-3.

[McLachlan 1992] McLachlan, G. J: Discriminant Analysis and Statistical Pattern Recognition. John Wiley, New York, 1992.

[McQueen 1967] J. McQueen, Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–296, Berkeley, CA, USA, 1967.

[Mehrotra 1997] S. Mehrotra, Y. Rui, M. Ortega and T. S. Huang. Supporting content-based queries over images in MARS. In Proc. IEEE International Conference on Multimedia Computing and Systems, 1997.

[Mezaris 2004] Vasileios Mezaris and Ioannis Kompatsiaris and Michael G. Strintzis, Region-based image retrieval using an object ontology and relevance feedback, EURASIP Journal on Applied Signal Processing, volume 2004, number 6, pages 886-901, 2004.

[Mitchell 1997] Tom M. Mitchell. Machine Learning, McGraw-Hill,1997.

[Miyamori 2000] H. Miyamori and Shu-ichi Isaku. Video Annotation for Content-based Retrieval using Human Behaviour Analysis and Domain Knowledge. In Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition, 2000.

[Moreno 2005] F. Moreno – Noguer, A. Sanfeliu and D. Samaras. Integration of Conditionally Dependent Object Features for Robust Figure/Background Segmentation. In Proc. 10th IEEE International Conference on Computer Vision (ICCV'05), 2005.

[Naphade 2000] M. Naphade and T. S. Huang. Semantic Video Indexing Using a Probabilistic Framework. In Proc. 15th International Conference on Pattern Recognition (ICPR'00), 2000.

[Naphade 2002a] M. R. Naphade and T. S. Huang, Extracting Semantics from Audiovisual Content: The Final Frontier in Multimedia Retrieval. IEEE Transactions on Neural Networks, Vol 13(4), 2002.

[Naphade 2002b] M. R. Naphade, S. Basu, J. R. Smith, C. Lin and B. Tseng. Modeling Semantic Concepts to Support Query by Keywords in Video. In Proc. International Conference on Image Processing (ICIP'02), 2002.

[Ng 1998] Andrew Y. Ng., On Feature Selection: Learning with Exponentially many Irrelevant Features as Training Examples, In Proceedings of the Fifteenth International Conference on Machine Learning, 1998.

[Ng & Jordan 2002] Andrew Y. Ng and Michael Jordan. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes. In NIPS 14, 2002.

[Naphade 2003] M. R. Naphade and J. R. Smith. Learning Visual Models of Semantic Concepts. In Proc. IEEE International Conference on Image Processing (ICIP'03), 2003.

[Neti 2000] C. Neti, B. Maison, A. Senior, G. Iyengar, P. Decuetos, S. Basu and A. Verma, Joint processing of audio and visual information for multimedia indexing and human-computer interaction, citeseer.ist.psu.edu/412939.html, 2000.

[Nilsback 2004] M. E. Nilsback and B. Caputo. Cue Integration Through Discriminative Accumulation. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), 2004.

[Ohtsuki 2006] Katsutoshi Ohtsuki, Katsuji Bessho, Yoshihiro Matsuo, Shoichi Matsunaga, and Yoshihiko Hayashi, Automatic Multimedia Indexing: Combining audio, speech, and visual information to index broadcast news, IEEE SIGNAL PROCESSING MAGAZINE [69], MARCH 2006.

[Opelt 2004] Opelt A., Fussenegger M., Pinz A., and Auer P. Weak hypotheses and boosting for generic object detection and recognition. In ECCV, 2004.

[Opelt 2004-2] Opelt A., Fussenegger M., Pinz A., and Auer P. Generic Object recognition with boosting. Technical report tr-emt-2004-01. submitted to PAMI.

[Opelt 2006] A. Opelt, A. Pinz, M. Fusseneger and P. Auer. Generic Object Recognition with Boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 28(3), 2006.

[Parzen 1962] E Parzen. On estimation of a probability density and mode. Annals of Mathematical Statistics, 35:1065{1076, 1962.

[Osian 2004] M. Osian and L. J. Van Gool. Video shot characterization. MVA, Vol. 15(3), 2004.

[Patel 1997] N. V. Patel and I. K. Sethi. Video shot detection and characterization for video databases. Pattern Recognition, SPIE, Vol 30(4), 1997.

[Petridis 2006] K. Petridis, S. Bloehdorn, C. Saathoff, C. Simou, N. Dasiopoulou, S. Tzouvaras, V. Handschuh, S. Avrithis, Y. Kompatsiaris and S. Staab. Knowledge Representation and semantic annotation of multimedia content. IEE Proceedings Vision, Image and Signal Processing, 2006.

[Pickering & Rüger 2003] Marcus J. Pickering and Stefan Rüger Evaluation of key frame-based retrieval techniques for video, Computer Vision and Image Understanding, Volume 92, Issues 2-3 , November-December 2003, Pages 217-235.

[Porkaew 1999] K. Porkaew, S. Mehrotra and M. Ortega. Query Formulation for Content Based Multimedia Retrieval in MARS. In Proc. IEEE Conference on Multimedia Computing and Systems (ICMS'99), 1999.

[Quattoni 2004] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In NIPS, 2004.

[Queen 1967] MacQueen, J: Some methods for classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds. L.M. LeCam and J. Neyman, University of California Press, pp 281-297, 1967

[Quinlan 1986] Quinlan, J. Ross: Induction of Decision Trees, Machine Learning, 1:86-106, 1986. Reprinted in Shavlik, J. and Dietterich, T. Readings in Machine Learning, San Francisco: Morgan Kauffmann, 1990, pp. 57-69

[Quinlan 1993] Quinlan, J. Ross: C4.5: Programs for Machine Learning, San Francisco: Morgan Kauffmann 1993.

[Robertson 1995] Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M., and Payne, A. (1995). Okapi at TREC-4, in NIST Special Publication 500-236, the Fourth Text Retrieval Conference (TREC-4), pages 73-96.

[Rocchio 1971] Rocchio J.J. Relevance feedback in information retrieval. In The SMART Retrieval System—Experiments in Automatic Document Processing, pages 313–323, Englewood Cliffs, NJ, 1971. Prentice Hall, Inc.

[Rubner 1998] Y Rubner. The earth-mover's distance as a metric for image retrieval. Technical Report STAN-CS-TN-98-86, Stanford University, 1998.

[Ruthven & Lalmas 1997] Ruthven I. and Lalmas M. Experimenting on D-S's theory of evidence in IR Technical report, Dept. of Comp. Sci., Univ. Glasgow, 1997.

[Saber 1998] E. Saber, A. M. Tekalp. Integration of colour, edge, shape and texture features for automatic region-based image annotation and retrieval. Journal of Electronic Imaging, Vol 7(3), 1998.

[Salton & Buckley 1990] Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41, 288-297.

[Santini 1998] S. Santini and R. Jain. Beyond Query by Example. In Proc. 6th ACM Conference on Multimedia, 1998.

[Schoelkopf 1998] B. Schoelkopf, P. Simard, A. Smola, and V. Vapnik: Prior knowledge in support vector kernels. In M. Jordan, M. Kearns, and S. Solla, editors, Advances in Neural Information Processing Systems 10, Cambridge, MA, 1998. MIT Press.

[Schmid 2001] Cordelia Schmid: Constructing models for content-based image retrieval, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01) - Volume 2, 2001

[Schmid 2004] Schmid C. Weakly supervised learning of visual models and its application to content-based retrieval. International Journal of Computer Vision, Volume 56, Number 1, 2004.

[Sciascio 2000] E. Di Sciascio, F. M. Donini and M. Mongiello. A Description Logic for Image Retrieval. Lecture Notes in Computer Science, Vol. 1792, 2000.

[Sclaroff 1997] Sclaroff S., Taycher L. And Cascia M. L. ImageRover: A content-based image browser for the world wide web. In Proc. IEEE Workshop on Content-based Access of Image and Video Libraries (San Juan, Puerto Rico, June 1997).

[Shet 2004] V.D. Shet, V.S.N. Prasad, A. Elgammal, Y. Yacoob, L.S. Davis. Multi-cue exemplar-based nonparametric model for gesture recognition. In Proc. Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), 2004.

[Smeaton 2004a] A.F. Smeaton, W. Kraaij and Paul Over. TREC Video Retrieval Evaluation: A Case Study and Status Report. In Proceedings of RIAO'2004, coupling approaches, coupling media and coupling languages for information retrieval, Avignon, France, April 2004.

[Smeaton 2004b] A.F. Smeaton, P. Over, and W. Kraaij. TRECVID evaluating the effectiveness of information retrieval tasks on digital video. In ACM Multimedia, New York, NY, Nov 2004.

[Smith 1997] John R. Smith and Shih-Fu Chang. VisualSEEK: A fully automated content based image query system. In Proc. Fourth ACM International Conference on Multimedia, 1997.

[Smith 2001] John R. Smith and Sankar Basu and Ching-Yung Lin and Milind R. Naphade and Belle Tseng, Integrating Features, Models and Semantics for Content-based Retrieval, Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MMCBIR'01), Rocquencourt, France, 2001, pages 95-98.

[Snoek 2006] C.G.M. Snoek, M. Worring, J.C.v.Gemert, J-M Geusebroek and A.M.Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. MM '06, October 23-27, 2006, Santa Barbera, California, ACM, 2006.

[Spengler 2003] M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. Machine Vision and Applications, Vol 14(1), 2003.

[Stoilos 2005] G. Stoilos, G. Stamou, V. Tzouvaras, J. Z. Pan and I. Horrocks. A Fuzzy Description Logic for Multimedia Knowledge Representation, In Proc. International Workshop on Multimedia and the Semantic Web, 2005.

[Stolcke 2006] A. Stolcke, B. Chen, H. Franco, Venkata Ramana Rao Gadde, M. Graciarena, Mei-Yuh Hwang, K. Kirchhoff, A. Mandal, N. Morgan, Xin Lei, Tim Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, Wen Wang, Jing Zheng and Qifeng Zhu. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. IEEE Transactions on Audio, Speech and Language Processing, ol. 14, No. 5. (2006), pp. 1729-1744.

[Straccia 1998] U. Straccia. A Fuzzy Description Logic. In Proc. 15th National Conference on Artificial Intelligence (AAAI-98), 1998.

[Straccia 2000] U. Straccia. A Framework for the Retrieval of Multimedia Objects Based on Four-Valued Fuzzy Description Logics. In Soft Computing Information Retrieval: Techniques and Applications, Fabio Crestani and Gabriella Pasi (eds.) Physica Verlag (Springer Verlag), 2000.

[Sudhir 1998] G. Sudhir, J. C. M. Lee, A. K. Jain. Automatic classification of Tennis Video for High-level Content – based Retrieval. In Proc. International Workshop on Content-based Access of Image and Video Databases (CAIVID '98), 1998.

[Sudderth 2005] EB Sudderth, A Torralba, WT Freeman, AS Willsky. Learning Hierarchical Models of Scenes, Objects, and Parts Tenth IEEE International Conference on Computer Vision, ICCV '05, Volume 2, 2005

[Sun 2003] Zhaohui Sun, Adaptation for multiple cue integration. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03), 2003.

[Teh 2006] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 2006.

[Titterington 1981] Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F., and Gelpke, G.J. 1981. Comparison of discrimination techniques applied to a complex data set of head injured patients. J. Roy. Statist. Soc. A, 144:145–175.

[Titterington 1985] D.M. Titterington, A.F.M. Smith, and U.E. Makov. Statistical Analysis of Finite Mixture Distributions. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 1985.

[Torralba 2004] A. Torralba, K. Murphy, and W.T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In Proc. of the 2004 IEEE CVPR., 2004.

[Tsai 2003] Tsai c., McGarry K. and Tait J.,Image Classification Using Hybrid Neural Networks, SIGIR 2003 Proceedings, Pages 431-432.

[Tsai 2004] Tsai, C.-F., McGarry, K., & Tait, J. (2004) Automatic metadata annotation of images via a two-level learning framework. Proceedings of the 2nd International Workshop on Semantic Web, in conjunction with ACM SIGIR'04.

[Tsai 2006a] Tsai Chih-Fong McGarry, Ken and Tait John Qualitative Evaluation of Automatic Assignment of Keywords to Images. Information Processing and Management, Volume 42 Issue 1, January, 2006. pp 136-154.

[Tsai 2006b] Chih-Fong Tsai, Ken McGarry and John Tait CLAIRE: A Modular Support Vector Image Indexing and Classification Systems, ACM Transactions On Information Systems 24(3) pp 353-379. 2006.

[Vailaya 1998] Aditya Vailaya, Anil Jain & Hong-Jiang Zhang. On Image Classification : City Images vs. Landscapes. Pattern Recognition Journal, 1998.

[Vailaya 2001] Aditya Vailaya, Mario A. T. Figueiredo, Anil K. Jain & Hong-Jiang Zhang. Image Classification for Content-Based Indexing. IEEE Transactions on Image Processing, vol. 10, 2001.

[van Rijsbergen 1979] Rijsbergen C. J. van (1979), Information Retrieval Book, Second Edition, chapter 7 Evaluation , p. 113 Butterworth-Heinemann Ltd, ISBN-13: 978-0408709293.

[Vapnik 1995] Vapnik, V. The Nature of Statistical Learning Theory, John Wiley & Sons, New York, 1995

[Vapnik 1998] Vapnik V. Statistical Learning Theory. Wiley, New York, NY, 1998.

[Varma 2002] M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In Proceedings of the European Conference on Computer Vision, 2002.

[Vasconcelos & Lippman 1998] Nuno Vasconcelos and Andrew Lippman. Embedded mixture modeling for efficient probabilistic content-based indexing and retrieval. In Proceedings 21 of the SPIE Conference on Multimedia Storage and Archiving Systems III, volume 3527, 1998.

[Vidal 2003] Vidal-Naquet M. and Ullman S. Object recognition with informative features and linear classification. In ICCV, 2003.

[Virga & Duygulu 2005] Paola Virga and Pinar Duygulu. Systematic Evaluation of Machine Translation Methods for Image and Video Annotation, The Fourth International Conference on Image and Video Retrieval (CIVR 2005) Proceedings, Pages 174-183, Singapore, July 20-22, 2005.

[Walt 2006] C.M. van der Walt and E. Barnard: Data characteristics that determine classifier performance, in Proceedings of the Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa, pp.160-165, 2006.

[Wang 2000] Yao Wang, Zhu Liu, and Jin-Cheng Huang, Multimedia Content Analysis: Using Both Audio and Visual Clues, IEEE SIGNAL PROCESSING MAGAZINE, 17:6, pp. 12-36, Nov. 2000.

[Wang 2006] G. Wang, Y. Zang and Li Fei-Fei. Using dependent regions for object categorization in a generative framework. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, 2006.

[Warmuth 2006] Warmuth, M. K., Liao, J., and R□ atsch, G. (2006) Totally Corrective Boosting Algorithms that Maximize the Margin, Proceedings of the 23th International Conference for Machine Learning (ICML 06), Carnegie Mellon, Pittsburgh PA. ACM International Conference Proceedings Series, vol. 148, pp. 1001-1008 June 2006.

[Wechsler 1998] H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman - Soulie and T. Huang (Eds.) (1998), FACE RECOGNITION: From THEORY to APPLICATIONS , Springer - Verlag.

[Westerveld 2000] Westerveld Thijs. Image Retrieval: Content versus Context. In Content-Based Multimedia Information Access, RIAO 2000 Conference Proceedings, pp. 276–284, April 2000.

[Westerveld 2003a] Thijs Westerveld, Arjen P. de Vries, Alex R. van Ballegooij, Franciska M.G. de Jong, and Djoerd Hiemstra A Probabilistic Multimedia Retrieval Model and its Evaluation. EURASIP Journal on Applied Signal Processing, 2003(2):186–198, February 2003. Special issue on Unstructured Information Management from Multimedia Data Sources.

[Westerveld 2003b] Westerveld T., Ianeva T., Boldareva L., de Vries A. P. and Hiemstra D. Combining information sources for video retrieval, The lowlands team at TRECVID 2003. In NIST TRECVID-2003, Nov 2003.

[Westerveld 2004] Westerveld Thijs. Using generative probabilistic models for multimedia retrieval. Ph.D. Thesis, University of Twente, Enschede, The Netherlands, 2004.

[Westerveld 2005] Thijs Westerveld and Arjen P. de Vries. Generative probabilistic models for multimedia retrieval: query generation against document generation. IEE Proceedings - Vision, Image, and Signal Processing, 152(6):852–858, IEE, December 2005.

[Woodland 1998] P. C. Woodland, T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk, and S. J. Young, Experiments in Broadcast News transcription, In Proc. ICASSP, 1998

[Woodland 2000] PC Woodland and D Povey. Large Scale Discriminative Training for Speech Recognition. In ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium, pages 7-16, Paris, 2000

[Wu 2000] Ying Wu, Qi Tian and Thomas S. Huang Discriminant-EM Algorithm with Application to Image Retrieval, In Proc. of IEEE Conf. on CVPR'2000, Vol.I, pp.222-227, Hilton Head Island, SC, 2000.

[Wu 2004] Y. Wu, E. Y. Chang, K. C.-C. Chang, and John R. Smith. Optimal Multimodal Fusion for Multimedia Data Analysis. In ACM International Conference on Multimedia, Pages: 572 - 579, October 2004.

[Yakici & Crestani 2006] Yakici M. and Crestani F. Cross-Media Indexing in the Reveal-This System, International Workshop on Crossing media for improved information access, LREC 2006.

[Yang 2003] Jian Yang, Jing-yu Yang, David Zhang and Jian-feng Lu. Feature fusion: parallel strategy vs. serial strategy. Pattern recognition, Elsevier, Vol. 36(6), June 2003, pp 1369-1381.

[Yang 2004] Yang J., Zhang D., Frangi A. and Yang J., Two-dimensional PCA: a new approach to appearance-based face representation and recognition IEEE. Trans. Pattern Analysis and Machine Intelligence, vol. 26, No. 1, 2004.

[Yavlinsky 2005] A. Yavlinsky, E. Schofield and S. Ruger. Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation. In Proc. International Conference on Image and Video Retrieval (CIVR'05), 2005.

[Yuan 2002] Y. Yuan, Q. Song and J. Shen. Automatic Video Classification Using Decision Tree Method. In Proc. First International Conference on Machine Learning and Cybernetics, 2002.

[Yuille 1994] A. L. Yuille and H. H. Bulthoff. Bayesian Decision Theory and Psychophysics. Advances in Neural Information Processing Systems, Vol 6, Morgan-Kaufmann publishers inc., 1994.

[Zhang 2001] Hong-Jiang Zhang and Z. Su, Improving {CBIR} by Semantic Propagation and Cross-Mode Query Expansion, Proceedings of the international workshop on MultiMedia Content-Based Indexing and Retrieval (MMCBIR'01), Rocquencourt, France, 2001, pages 83-86.

[Zhang 2005a] D.Q. Zhang. Statistical Part-Based Models: Theory and Applications in Image Similarity, Object Detection and Region Labeling. PhD Thesis,Graduate School of Arts and Sciences, Columbia University, 2005

[Zhang 2005b] Ruofei Zhang and Zhongfei (Mark) Zhang and Mingjing Li and Wei-Ying Ma and Hong-Jiang Zhang, A Probabilistic Semantic Model for Image Annotation and Multi-Modal Image Retrieval, Proc. of the 2005 IEEE International Conference on Computer Vision (ICCV'05), 2005.

[Zhai & Lafferty 2001] Zhai C. and Lafferty J. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval Proceedings of the 2001 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), 2001.

[Zhang & Chen 2002] C. Zhang and T. Chen, An active learning framework for content-based information retrieval, IEEE Transactions on Multimedia, volume 4, number 2, pages 260-268, 2002.

[Zhao 2002] Rong Zhao and William I. Grosky, Narrowing the semantic gap---improved text based web document retrieval using visual features, IEEE Transactions on Multimedia, volume 4, number 2, pages 189-200, year 2002.

[Zhou, X. S. & Huang 2001] Zhou X. S., and Huang T. S., Comparing Discriminate Transformations and SVM for Learning during Multimedia Retrieval, ACM Multimedia2001, Sept. 30-Oct 5, 2001, Ottawa, Ontario, Canada, 2001.

[Zhou 2002] Xiang Sean Zhou and Thomas S. Huang, Unifying Keywords and Visual Contents in Image Retrieval, IEEE Multimedia, volume 9, number 2, pages 23-33, 2002.

[Zhu & Huang 2003] X.S.Zhu and T.S.Huang. Relevance feedback in image retrieval: A comprehensive review. Multimedia Systems 8(6): 536-544, 2003.

Appendix:

1 Generic Methods for Supervised Learning

1.1 Support Vector Machines

Support Vector Machines (SVMs) [Boser 1992], [Vapnik 1995], [Vapnik 1998] gained widespread use because of successful applications like character recognition and the profound theoretical underpinnings concerning generalization performance.

SVMs map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The hyperplane is the hyperplane that maximizes the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be.

To introduce some notation, suppose we are given *l* observations. Each observation consists of a pair: a vector $\mathbf{x}_i$ in R^n, $i = 1,\ldots,l$ and the associated truth y_i given to us by a trusted source. As a naive example, consider a tree recognition problem where $\mathbf{x}_i$ might be a vector of pixel values with n = 256 for a 16x16 image and y_i would be 1 if the image contains a tree and -1 otherwise (we use -1 instead of 0 to simplify subsequent formulae).

SVMs can be divided into two major classes, the linear and the non-linear, regarding to their ability of performing the classification task to data that is linearly separable or not, respectively.

1.1.1 Linear support vector machines

If the data to be classified are linearly separated, it is the task of SVMs to construct a hyperplane that separates positive from negative examples. The points x which lie on the hyperplane satisfy $w \cdot x + b = 0$, where w is normal to the hyperplane, $|b| / ||w||$ is the perpendicular distance from the hyperplane to the origin and $||w||$ is the Euclidean norm of w.

Let d_+ and d_- be the distances from the hyperplane to the closest positive and negative example respectively. If we define as "margin" the sum $d_+ + d_-$, the support vector algorithm simply looks for the separating hyperplane with the largest margin. This can be formulated as follows:

Suppose that all the training data satisfy the following constraints:

$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1$ for $y_i = +1$ (1)

$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1$ for $y_i = -1$ (2)

These can be combined into one set of inequalities:

$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0$ for each i (3)

Now consider the points for which the equality in Eq. (1) holds. These points lie on the hyperplane H_1: $\mathbf{x}_i \cdot \mathbf{w} + b = 1$ with normal $\mathbf{w}$ and perpendicular distance from the origin $|1 - b| / ||\mathbf{w}||$. Similarly, the points for which the equality in Eq. (2) holds lie on the hyperplane H_2: $\mathbf{x}_i \cdot \mathbf{w} + b = -1$, with normal again $\mathbf{w}$, and perpendicular distance from the origin $| -1 - b| / ||\mathbf{w}||$. Hence $d_+ = d_- = 1 / ||\mathbf{w}||$ and the margin is simply $2 / ||\mathbf{w}||$. Note that H_1 and H_2 are parallel (they have the same normal) and that no training points fall between them. Thus we can find the pair of hyperplanes which gives the maximum margin by minimizing $||w||^2$, subject to constraints (3).

Writing the classification in its dual form by switching to a Lagrangian formulation, and after some computations, Eq. (3) becomes $L_D = \max\left(\sum_i a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j\right)$, where α_i are the Lagrangian multipliers. We can see that the training data only appears (in the actual training and test algorithms) in the form of dot products between vectors. This is a crucial property which will allow us to generalize the procedure to the nonlinear case.

In the case of non linear separable data sets, it can be shown that the aforementioned procedure cannot be applied. In order to deal with this SVMs use the “kernel trick” proposed in [Aizerman 1964]. This method typically maps the observations into a higher-dimensional space; this makes a linear classification in the new space equivalent to non-linear classification in the original space.

1.1.2 Non linear support vector machines

Although the original hyperplane algorithm is a linear classifier, Bernhard Boser, Isabelle Guyon and Vapnik in 1992 [Boser 1992] suggested a way of applying the aforementioned Aizerman’s kernel trick. This allowed the algorithm to fit the maximum-margin hyperplane in a transformed feature space. . The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function. The transformation may be non-linear and the transformed space high dimensional; thus though the classifier is a hyperplane in the high-dimensional feature space it may be non-linear in the original input space.

If the kernel used is a Gaussian radial basis function, the corresponding feature space is a Hilbert space of infinite dimension. Maximum margin classifiers are well regularized, so the infinite dimension does not spoil the results. Some common kernels include, homogeneous and inhomogeneous polynomial, radial basis function, Gaussian, and sigmoid.

1.1.3 Limitations

One of the main drawbacks of SVMs is their high computational demands during the training and the testing phase. While the speed problem in test phase is largely solved in [Burges 1996] this still requires two training passes. Training for very large datasets (millions of support vectors) is an unsolved problem.

Discrete data presents another problem, although with suitable rescaling excellent results have nevertheless been obtained [Joachims 1997]. Finally, although some work has been done on training a multiclass SVM in one step24, the optimal design for multiclass SVM classifiers is a further area for research.

Perhaps the biggest limitation of the support vector approach lies in choice of the kernel. Some work has been done on limiting kernels using prior knowledge [Schoelkopf 1998], but the best choice of kernel for a given problem is still a research issue.

1.2 k-Nearest Neighbors

The k-Nearest Neighbor algorithm (k-NN) is an instance based learning method for classifying objects based on closest training examples in the feature space. In k-NN, the training examples are mapped into multidimensional feature space (prototypes). The space is partitioned into regions by class labels of the training samples. A point in the space is assigned to the class *C* if it is the most frequent class label among the *k* nearest training samples. Usually Euclidean distance is used.

The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the actual classification phase, the same features as before are computed for the test sample (whose class is not known). Distances from the new vector to all stored vectors are computed and *k* closest samples are selected. The new point is predicted to belong to the most numerous class within the set.

The best choice of *k* depends upon the data; generally, larger values of *k* reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good *k* can be selected by parameter optimization using, for example, cross-validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when $k = 1$) is called the nearest neighbour algorithm [Fix 1951].

The accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the features scales are not consistent with their relevance. Van der Walt and Barnard have also shown [Walt 2006] that the optimal value of *k* is influenced by the amount of output noise in the data. They also show that the Achilles heel of the k-NN classifier is the constant distance metric that it uses. Much research effort has been placed into selecting or scaling features to improve classification. A particularly popular approach is the use of evolutionary algorithms to optimize feature scaling. Another popular approach is to scale features by the mutual information of the training data with the training classes.

The algorithm is easy to implement, but it is computationally intensive, especially when the size of the training set grows. Many optimizations have been proposed over the years; these generally seek to reduce the number of distances actually computed. Some optimizations involve partitioning the feature space, and only computing distances within specific nearby volumes. Several different types of nearest neighbour finding algorithms include linear scan, kd-trees, balltrees, metric trees, locality-sensitive hashing (LSH), agglomerative nearest neighbour.

The nearest neighbour algorithm has some strong consistency results. As the amount of data approaches infinity, the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data). k-NN is guaranteed to approach the Bayes error rate, for some value of *k* (where *k* increases as a function of the number of data points).

Despite its simplicity, k-NN has been successful in a large number of classification problems, including handwritten digits, satellite image scenes and EKG patterns. It is often successful where each class has many prototypes and the decision boundary is very irregular. The extensive literature on the topic is reviewed by Dasarthy [Dasarthy 1991].

1.3 Naive Bayes

A naive Bayes classifier [Titterington 1981], [Lan1992], is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers often work much better in many complex real-world situations than one might expect. It has been shown that naive Bayesian learning is remarkably effective in practice and difficult to improve upon systematically [Domingos 1996] and has proven effective in many practical applications including text classification, medical diagnosis and systems performance management [Domingos 1997], [Mitchell 1997], [Hellerstein 2000].

1.3.1 The naive Bayes probabilistic model

Given feature measurements $F_1, F_2, \ldots, F_n$, we want to estimate the probability that our measurement belongs to a class C, that is, define the conditional probability $p(C| F_1, F_2, \ldots, F_n)$. The problem is that if the number of features n is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. Using Bayes' theorem we can rewrite this probability as $p(C \mid F_1, \ldots, F_n) = \frac{p(C)p(F_1, \ldots, F_n \mid C)}{p(F_1, \ldots, F_n)}$. Since the denominator is independent of C and the feature values are known, it is constant and we are only interested in the numerator which is essentially the joint probability p(C, $F_1, \ldots, F_n$). With repeated application of the conditional probability application this can be rewritten as $p(C, F_1, \ldots, F_n) = p(C)p(F_1 \mid C)p(F_2 \mid C, F_1)p(F_3 \mid C, F_1, F_2)\ldots$

The naive Bayes assumption claims that the features are conditionally independent so that $p(F_i \mid C, F_j) = p(F_i \mid C)$, for $i \neq j$. With this assumption the above equation can be written as $p(C, F_1, \ldots, F_n) = p(C)\prod_{i=1}^{n} p(F_i \mid C)$.

In order to estimate the parameters of the probability model, one can use a number of estimation procedures such as maximum likelihood or Bayesian inference.

1.3.2 Classification

The naive Bayes classifier combines this model with the maximum a posteriori (MAP) decision rule. The corresponding classifier is the function that maximizes p(C,Fi,…,Fn):

$$classify(f_1, \ldots, f_n) = \arg\max_c p(C = c)\prod_{i=1}^{n} p(F_i = f_i \mid C = c)$$

.

1.4 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis ([Fisher 1936], [Fukunaga 1990]) searches for those vectors in the underlying space that best discriminate among classes (rather than those that best describe the data). More formally, given a number of independent features relative to which the data is described, LDA creates a linear combination of these which yields the largest mean differences between desired classes. Mathematically speaking, for all the samples of all classes we define two measures: (i) one called within scatter matrix, as given by $S_w = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (\mathbf{x}_i^j - \mu_j)(\mathbf{x}_i^j - \mu_j)^T$, where $\mathbf{x}_i^j$ is the *i*th sample of class *j*, μ_j is the mean of class j, *c* is the number of classes and N_j the number of samples in class *j*; and (ii) the other is called between-class scatter matrix, $S_b = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^{\mathrm{T}}$, where μ represents the mean of all classes.

The goal is to maximize the between-class measure while minimizing the within-class measure. One way to do this is to maximize the ratio $\frac{\det|S_b|}{\det|S_w|}$. The advantage of using this ratio is that it has been proven [Fisher 1936] that if S_w us a non-singular matrix then this ratio is maximized when the column vectors of the projection matrix, **W (W** represents a linear transformation that maps the data space to the lower dimensioned feature space**)**, are the eigenvectors of $S_w^{-1} S_b$.

There are two quite different justifications for using Fisher's linear discriminant rule: the first is that it maximizes the separation between the classes in a least-squares sense; the second is by Maximum Likelihood. For a proof that they arrive at the same solution, we refer the reader to [McLachlan 1992].

1.5 Decision Trees

A decision tree (generally defined) is a tree whose internal nodes are tests (or input patterns) and whose leaf nodes are categories (of patterns). A class number is assigned to an input pattern by filtering the pattern down through the tests in the tree. Each test has mutually exclusive or exhaustive outcomes.

There are several dimensions among which decision trees might differ:

1. The tests might be univariate or multivariate, testing on one or several features of the input at once, respectively

2. The tests might have two outcomes or more than two (if all tests have two outcomes, we have a binary decision tree)

3. The features or attributes might be categorical or numeric

4. We might have two classes or more than two. If we have two classes and binary inputs, the tree implements a Boolean function, and is called a Boolean decision tree

Several systems for learning decision trees have been proposed. Prominent among these are ID3 [Quinlan 1986], C4.5 [Quinlan 1993] and CART [Breiman 1984].

Decision trees are considered to be “flexible” due to their ability of performing multivariate splits and their ability to examine the effects of the predictor variables one at a time, rather than just all at once.

The flexibility of decision trees makes them a very attractive analysis option, but this is not to say that their use is recommended to the exclusion of more traditional methods. Indeed, when the typically more stringent theoretical and distributional assumptions of more traditional methods are met, the traditional methods may be preferable. But as an exploratory technique, or as a technique of last resort when traditional methods fail, decision trees are, in the opinion of many researchers, unsurpassed.

1.6 K-means

K-means clustering ([Lloyd 1957], [Queen 1967]) is a method for finding clusters and cluster centres in a set of unlabeled data. One chooses the desired number of cluster centres, say R, and the K-means procedure iteratively moves the centres to minimize the total within cluster variance. Given an initial set of centres, the K-means algorithm alternates the two steps:

- For each centre we identify the subset of training points (its cluster) that is closer to it than any other centre;
- The means of each feature for the data points in each cluster are computed, and this mean vector becomes the new centre for that cluster.

These two steps are iterated until convergence. Typically the initial centres are R randomly chosen observations from the training data.

To use K-means clustering for classification of labelled data, the steps are:

- Apply K-means clustering to the training data in each class separately, using R prototypes per class;
- Assign a class label to each of the K x R prototypes;
- Classify a new feature x to the class of the closest prototype.

1.7 Learning Vector Quantization (LVQ)

In this technique due to Kohonen (1989), prototypes are placed strategically with respect to the decision boundaries in an ad-hoc way. LVQ is an online algorithm – observations are processed one at a time.

The idea is that the training points attract prototypes of the correct class and repel other prototypes. When the iterations settle down, prototypes should be close to the training points in their class. The learning rate is decreased to zero with each iteration, following the guidelines for stochastic approximation learning rates.

The procedure just described is actually called LVQ1. Modifications (LVQ2, LVQ3, etc) have been proposed, that can sometimes improve performance.

A drawback of learning vector quantization methods is the fact that they are defined by algorithms, rather than optimization of some fixed criteria; this makes it difficult to understand their properties.

1.8 Gaussian Mixture Models (GMMs)

A Gaussian mixture model defines a probability distribution over a feature space [Duda 2000, Titterington 1985]. The features describing the images are assumed to be generated by a mixture of Gaussian sources, where the number of Gaussian components is fixed. A Gaussian mixture model is described by a set of parameters each defining a single component. Each component is described by its prior probability the mean vector and the variance. The process of generating a set of feature vectors is assumed to be the following (see figure 5):

1. Take the Gaussian mixture model $\boldsymbol{\theta}$
2. For each feature vector $\boldsymbol{v}$
    a. Pick a random component c_i from Gaussian mixture model $\boldsymbol{\theta}$ according to the prior distribution over components P(c_i|$\boldsymbol{\theta}$)
    b. Draw a random sample from v according to this component's Gaussian distribution

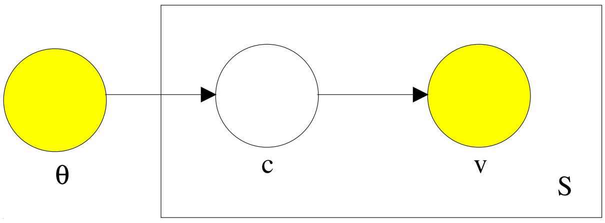


Figure 5: Graphical representation of Gaussian mixture model.

Here, $\boldsymbol{\theta}$ is an observed variable, i.e., the mixture model from which the vectors are drawn, is known. For a given feature vector however, it is unknown which component generated it, thus components are unobserved variables. The probability of drawing a single vector $\boldsymbol{v}$ from a Gaussian mixture model with parameters $\boldsymbol{\theta}$ is thus defined as the marginalization over all possible components:

$$p(v \mid \theta) = \sum_{i=1}^{C} P(c_i \mid \theta) p(v \mid c_i, \theta) = \sum_{i=1}^{C} P(c_i \mid \theta) \frac{1}{\sqrt{(2\pi)^n \mid \Sigma_i \mid}} - e^{\frac{1}{2}(v-\mu_i)^T \Sigma^{-1} (v-\mu_i)}$$

The probability of drawing a set of feature vectors is defined as the joint probability of the individual vectors, where vectors are assumed to be generated independently. Gaussian mixture models can be estimated from data using the Expectation Maximization method. Gaussian mixture models are a form of unsupervised learning that provides a soft clustering of the features in C different subclasses. Still, the models can be useful in a supervised setting, where the classes that need to be distinguished are assumed to consist of a number of unknown subclasses. In such cases, a mixture model may be more suitable to model the underlying class data than a single distribution.

1.9 Latent Dirichlet Analysis

Latent Dirichlet Analysis (LDA)[1] is a generative probabilistic model developed for collections of discrete data [Barnard 2003]. The model describes a probability distribution over a discrete feature space. The model describes, like GMM, a distribution over a set of underlying latent variables (or hidden topics), each of which is assumed to have its own distribution over the data points in the feature space. The generative process underlying LDA is the following:

1. Choose a random multinomial model θ according to the Dirichlet distribution Dir(α)
2. For each of the *N* features f:
    a. Draw a topic z from the multinomial distribution θ
    b. Draw a f from the distribution corresponding to z

An important difference to GMMs (apart from the discrete feature space) is the fact that the model from which the hidden topics are drawn does not correspond directly to a known document, but is drawn from a prior distribution. This means, LDA has the ability to generate a collection of *documents*, each with its own distribution of underlying hidden topics and thus with its own feature distribution. This has a number of advantages. First, the number of parameters needed to estimate does not grow with the number of documents represented. Second LDA can assign probabilities to documents outside the training set and thus LDA is able to give an indication of how likely the new document is to come from the same collection (or class) of documents.

[1] Not to be confused with Linear Discriminant Analysis, which is also abbreviated as LDA.

1.10 Expectation maximization (EM)

Expectation maximization (EM) is a method to obtain maximum likelihood estimates with incomplete or unobserved data [Dempster 1977]. It is useful for estimating the distributions for mixture models like GMM and LDA, where the component of the mixture that generates a sample is unknown. EM alternates between estimating the values of the unobserved or missing data based on the current model distributions, and re-estimating the model distributions based on the current estimates of the unobserved data. We take the GMM case as an example. To accurately describe the different components of a Gaussian mixture model for a given set of features, it is necessary to decide which of the features are generated by which component. The assignments of features to components are unknown, but they can be viewed as hidden variables. The EM algorithm iterates between estimating the a posteriori class probabilities for each feature given the current model settings (the E-step) and re-estimating the components' parameters based on the feature distribution and the current feature to component assignments (M-step).

The EM algorithm first assigns each feature to a random component. Next, the first M-step computes the parameters (prior, mean and covariance) for each component, based on the samples assigned to that component. This assignment of samples to components is a soft clustering; a sample does not belong entirely to one component. In fact, we compute means, covariances and priors on the weighted feature vectors, where the feature vectors are weighted by their proportion of belonging to the class under consideration. In the next E-step, the class assignments are re-estimated, i.e., the posterior probabilities, *P(ci|vj)* are computed. We iterate between estimating class assignments (expectation step) and estimating class parameters (maximization step) until convergence.

1.11 Boosting

Boosting is a machine-learning method that is based on the observation that finding many rough rules of thumb can be a lot easier than finding a single, highly accurate prediction rule. The algorithm trains a sufficient number of weak and rather inaccurate classifiers with a different subset of training examples. The combination of the extracted weak rules into a single prediction rule results in an accurate classifier.

There are several different boosting algorithms, depending on the exact mathematical form of the strength and weight. Adaboost [Freund 1997] is a popular and the historically most significant boosting algorithms, whereas more recent algorithms such as LPBoost [Demiriz 2002] and TotalBoost [Warmuth 2006] have replaced AdaBoost because they converge much faster and produce sparser hypothesis weightings. Most boosting algorithms fit into the AnyBoost [Mason 2000] framework, which shows that boosting performs gradient descent in function space.

From the aforementioned variations of the boosting algorithms, Adaboost is described below:

Given training examples (x_1,y_1),…, (x_m,y_m), where $x_i \in X$, $y_i \in Y = \{-1,+1\}$

Initialize $D_1(i)=1/m$

For $t = 1,\ldots,T$:

- Train "base" or "weak" learner using distribution D_t
- Get base classifier $h_t : X \to \Re$
- Choose $a_t \in \Re$
- Update $D_{t+1} = \frac{D_t(i)\exp(-a_t y_i h_t(x_i))}{Z_t}$, where Z_t is a normalization factor so that D_{t+1} will be a distribution.

The output of the final classifier is:

$$H(x) = sign\left(\sum_{t=1}^{T} a_t h_t(x)\right)$$

What is most important is that the boosting procedure can be combined with any base learning algorithm.

Besides this original view, AdaBoost has been interpreted as a procedure based on functional gradient descent, as an approximation of logistic regression and as a repeated-game playing algorithm. AdaBoost has also been shown to be related to many other topics, such as game theory and linear programming, Bregman distances, support-vector machines, Brownian motion, logistic regression and maximum-entropy methods such as iterative scaling.