User Relevance Feedback, Search and Retrieval of Visual Content

Divna Djordjevic

Submitted for a Degree of

Doctor of Philosophy

Department of Electronic Engineering

Queen Mary University of London

2006

To My Parents

Abstract

The main objective of this work is to study and implement techniques for visual content retrieval using relevance feedback. Relevance feedback approaches make use of interactive learning in order to modify and adapt system behaviour to user's desires by modelling human subjectivity. They allow a more semantic approach based on user's feedback information, while relying on similarity derived from low-level features.

An image relevance feedback framework has been implemented based on support vector machines as a generalisation method. The algorithm for support vector machines solves a convex optimisation problem and the algorithm has been tailored to the relevance feedback scenario.

MPEG-7 standard descriptors and their recommended distance functions have been used to represent low-level visual features as well as several additional descriptors. A multi-feature scenario has been developed in an effort to represent visual content as close as possible to human perceptual experience. A model for feature combination and not just concatenation has been developed and a novel kernel for adaptive similarity matching in support vector machines has been proposed. The new kernel models multi-feature space guaranteeing convergence of the support vector optimisation problem.

To address the problem of visual content representations, a novel approach of building descriptors based on image blocks, their low-level features and their spatial correlation has been proposed as a part of the relevance feedback framework. In accordance to this an accompanying kernel on sets has been proposed that handles both multi-feature space as well as the local spatial information among image blocks.

The relevance feedback module has been applied to a framework for image selection in concept learning. It combines unsupervised learning to organize images based on low-level similarity, and reinforcement learning based on relevance feedback, to refine the classifier model. This research is a part of the EU IST aceMedia project and the described relevance feedback module has been integrated in the project framework.

Acknowledgements

I will always be in dept of gratitude to so many people for their continuous support in my work on this thesis. I owe my deepest gratitude and respect to Professor Ebroul Izquierdo, whose guidance, advices and criticism were crucial for the completion of this thesis.

I would like to thank to the whole Electronic Engineering Department at Queen Mary, University of London. To all people in the Multimedia and Vision Group that I had pleasure to share the last few years with.

To all of them that who warmly greeted me specially Maria, Miguel, Winyu, Andy, Juri, Emilio, Gaban. Over these years, I have enjoyed exchanging ideas and have had the opportunity to work with outstanding people particularly Mr. Andres Dorado and Mrs. Qianni Zhang. I am grateful for the privilege of working with the “aceMedia people” specially Mrs. Marta Mrak, Mr. Nikola Sprljan and Mr. Toni Zgaljic. I am also deeply thankful to Dr. Andrea Cavallaro for his charming ability to listen to all of my worries and of course for numerous advices.

My special gratitude goes to Dr. Ivan Damnjanovic for his advices, continuous support and friendship.

Finally, I am grateful to my beloved parents, my brother and my sister, and to my love Fede, for their ongoing support and patience.

Table of Content

List of Figures

List of Tables

List of Abbreviations

CBIR	Content Based Image Retrieval
RF	Relevance Feedback
MPEG	Moving Picture Expert Group
QbE	Query by Example
QPM	Query Point Movement
ERM	Empirical risk minimization principal
SRM	Structural risk Minimisation
VC	Vapnik-Chervonenkis
KKT	Karush-Kuhn-Tucker
SVs	Support Vectors
SVMs	Support Vector Machines
PD	Positive definite
CPD	Conditionally positive definite
CLD	Colour Layout Descriptor
EHD	Edge Histogram Descriptor,
CSD	Colour Structure Descriptor
DCD	Dominant Colour Descriptor
GF	Gabor Filters feature
GLCM	Grey Level Co-occurrence Matrix descriptor
HSV	Hue-Saturation-Value colour space based descriptor
RBFK	Radio Basis Function Kernel
LK	Laplace Kernel
ACK	Adaptive Convolution Kernel
pdf	probability density function
FCM	Fuzzy C-Means

DCD	Discrete Cosine Transform
IST	Information Society Technologies

CHAPTER 1 : Introduction

1.1. Problem and Motivation

In modern and continuously growing media databases, the inability of accessing accurate and desired content can be as limiting as the lack of content itself. Research in information retrieval whether based on textual description or low-level content is aiming at overcoming the drawbacks of machine limited behaviour and incorporating human understanding into the complex equation of machine responses.

Unlike textual information, which is human defined and precise in meaning, a picture, or audio-video content has a hidden component of creative reasoning of the human brain. This gives the content an overall shape and meaning far beyond capabilities of any language-based representation. In the last decades, the research in information retrieval has moved from strictly text based retrieval systems to multimedia content databases fusing information together. Throughout recent years, the idea of simulating human understanding has been closely related to iterative feedback. This approach incorporates the obtained knowledge into a learning approach that could eventually be able to "think" and "behave" as a human. In this thesis the emphasis is on still image databases and contributions in the domain of interactive retrieval using visual content and user provided feedback.

Content-based image retrieval (CBIR) uses low-level features such as colour, shape and texture to represent visual content and automatically index image databases. However, the search for a particular semantic content based on semantic user defined commonalities can be relatively uncorrelated with low-level feature similarity. A human user usually searches for examples of content with similar semantic meaning however similarity in huge databases can only be provided on feature level automatically extracted from the content. This leads to a major problem, when searching for relevant content, which can be described with both sensory and semantic gap. The sensory gap is a difference in information between objects in the real world and their recorded descriptions as perceived by recording devices. On the other hand, the semantic gap is the lack of consistency in information extracted from the visual content and the user defined interpretation of the same content. In order to bridge these gaps, low-level features need to be consistent representatives of common concepts in the database. They should be invariant to occlusion, illuminationn, noise, clutter and viewpoint. However, at the same time these features should be discriminative enough to distinguish a variety of concepts.

To solve this problem CBIR is coupled with annotations, taxonomies, ontology and especially with user relevance feedback, to emphasize hidden associations between high-level semantic concepts and low-level features extracted from data observations.

1.2. Research Objectives

As means of achieving a step closer to bridging the gap between human and machine driven reasoning, iterative short term and low-effort relevance feedback, has been presented as an unavoidable step. This thesis focuses on CBIR with user defined relevance feedback. Specifically, the following objectives were considered:

- To investigate low-level visual feature extraction and the importance of these features in CBIR systems with relevance feedback.
- To analyze and to select learning approaches for relevance feedback.
- To develop an approach for effective low-level feature combination in conjunction with learning strategies, and the possibility of integrating the overall feature space into the learning method.
- To incorporate localized image information into the learning approach as well as into the low-level similarity.

- To suggest possible application scenarios for effective use of relevance feedback (RF) approaches in a multi-feature space.

1.3. Contribution of the Thesis

Following the guidelines given above a kernel based relevance feedback approach in CBIR has been developed. All necessary steps in enabling an effective RF for natural images have been investigated and novelty introduced in a number of appropriate stages:

- Low-level features were analysed and a discriminative and **descriptive low-level feature combination was proposed**. This combination is able to effectively capture low-level representations of natural images for a retrieval scenario.
- An adaptive **convolution kernel dealing with multi-feature spaces** and guaranteeing convergence of the SVM optimisation problem has been introduced.
- **A set kernel coupled with clustering approaches, defined in structured space has been proposed.** It encloses both multi-feature and spatial information about localized image blocks enabling a higher transparency between low-level image features and semantic concepts.
- Application to classification for **improved image selection in concept learning** has also been introduced. It combines unsupervised learning to organize images based on low-level similarity, and reinforcement learning based on relevance feedback, to refine the classifier model.

The research described in the thesis and improvements of conventional approaches have been presented in a number of author's publications, which are given at the end of this thesis.

This research has also contributed to the EU IST "aceMedia" project and an accompanying relevance feedback module has been integrated and evaluated in the project framework.

As final remark, this thesis is dealing with still image databases, while video databases have been considered in correspondence with key-frame extraction (Djordjevic et al. 2005).

1.4. Structure of the Thesis

In this thesis the necessary elements to understand and develop an effective relevance feedback approach in CBIR are gradually introduced.

Chapter 2 contains description of feature extraction methods, and choice of visual features used to build the multi-feature space as well as similarity measures. It also introduces the Multimedia Content Description Interface, MPEG-7, and gives an evaluation for multi-feature spaces in a content based retrieval and browsing scenario.

Chapter 3 gives a summary of the existent retrieval methods along with state-of-the-art approaches for relevance feedback. The emphasis is put on the description of support vector machines (SVMs) used for learning user preferences, which are introduced through interactive relevance feedback.

Chapter 4 introduces a new kernel for multi-feature spaces. It analyses the necessary mathematical properties of kernels to enable a convex SVM optimization problem and how the proposed kernel fits into these requirements.

Chapter 5 deals with the complex problem of structured descriptor spaces, that connects both low-level and spatial information of localized image blocks. A kernel on sets is introduced. This kernel is based on local kernels defined on local image parts in the multi-feature space, as investigated in the previous chapter.

Chapter 6 describes two application frameworks for the designed method. The first one couples clustering methods and supervised relevance feedback to improve classifier performances. The second application describes the role of the developed relevance feedback module within the “aceMedia” project.

Chapter 7 gives a discussion of the introduced contributions and concluding remarks.

Complementary explanations are organized into three appendixes. Appendix A shows a visual overview of used ground truth image databases. Appendix B introduces several mathematical prerequisites needed for kernel analysis. Appendix C presents a detailed algorithm and the necessary steps used for solving and implementation of the convex SVM optimization problem.

CHAPTER 2 : Low-level Visual Features, and Reliability of Similarity Matching

2.1. Introduction

In this chapter several extensively used methods for visual content descriptions are introduced. Both generic and descriptor specific, similarity measures between visual features are considered. The connection between image descriptors, and similarity measures on one side and the quality of retrieved set in a simple similarity matching CBIR system on the other side, is analyzed. Retrieval performances and reliability of the considered descriptors are further evaluated.

2.2. Low-Level Features

Low-level feature representations aim at capturing low-level visual content similarities. However, these representations are limited to content and cannot infer complex semantic meaning. In an effort to deal with the sensory gap low-level features need to be invariant to distortions in the recordings of objects. Nevertheless, in order to deal with the semantic gap image features also need to be invariant to different instances of the same semantic concept and at the same time, they have to be discriminative enough to be able to differentiate among various concepts.

Image databases either relate to a specific domain or represent a broad variety in content encountered in real world applications. Therefore, a wide variety of approaches and searching scenarios can be encountered. As a consequence the ratio between invariance and discrimination of features cannot be generically defined. Nevertheless some relations can be inferred from the scenario, the type of database and the user requirements. Usually a particular semantic meaning for content can be inferred with a highly representative feature or combination of features. Note that in the rest of the thesis the term descriptor is used to identify a specific syntactic representation of visual feature.

2.2.1 Colour spaces and Colour descriptors

Colour is a very important low-level feature and one of the strongest descriptors for image retrieval. Various colour-based representations have been proposed in the past. However, before defining a colour feature an appropriate colour space has to be chosen.

Colour spaces

In general image pixels can be represented with a three dimensional colour space model. Depending on the application Gevers (2001) distinguished various colour space models for different applications. Though every space model has its advantages, uniformity is the main required characteristic in image retrieval systems. Hence, the colour space needs to be perceptually uniform. This means that the distances between two colours that are equal in the introduced colour space should also imply perceptual equality observed by humans. In general, the chosen colour system for feature representation needs to be independent from the underlying imaging device. The computational transform, from the RGB space in which images are captured to the considered space, should be linear to avoid instability towards noise. Invariance towards a number of changes is also required for image retrieval such as illumination, occlusion, viewpoint, object pose. Hence, when there is little variation in perception of an object or a scene, the RGB colour space is a good choice. A further improvement of colour spaces considers $L^*a^*b^*$space with a relative perceptual uniformity. The HSV colour space considers human intuition, and addresses three of the most important aspects in the perception of color: hue, saturation and value. The hue and saturation components are based on the way human eye perceives colour. Hue corresponds to different colours while saturation varies from unsaturated (shades of grey) to fully saturated (no white component, intense colour). Finally, value or brightness corresponds to colours becoming increasingly brighter.

In this section several commonly used colour descriptors are reported. These were proposed in recent years for image retrieval applications by a number of researchers (Gevers and Stokman 2003). They include colour histograms, colour moments, colour coherency vector, colour correlograms etc.

Colour Histogram

The research in the area of colour histogram representation has been active ever since Swan and Ballard (1991) introduced colour histograms. As mentioned before, an image representation is highly dependant on the application, the colour space used as well as the quantisation level (see Figure 2.1). Finer quantisation results in better colour representation however, the dimensionality of the representative feature is increased. The colour histogram used in this work is a global colour histogram in the HSV space. It is generated by counting the frequency of pixels with colour values that fall into specific colour ranges (bins). The bins are defined based on the colour space and the number of quantisation levels. Colour histograms are robust to translation and rotation of images with slight differences in values when changes in scale, occlusion or viewing angle are introduced. Though histograms are good representatives of colour distributions across the image, they lack spatial colour information. To address this issue local colour descriptors, such as colour layout or region-based descriptors have been developed.

Colour Moments

Colour moments represent a compact feature based on statistical first, second and third order moments for colour components of pixels in an image. These features are usually defined in the $L^*a^*b^*$ and $L^*u^*v^*$ colour spaces, with maximally three values for each of the three colour components.

Colour Coherence Vector

The colour coherence vector incorporates spatial information into the colour histogram. Each bin is replaced by a two dimensional bin with values representing the number of coherent or incoherent pixels belonging to that bin. A pixel is considered coherent if it belongs to a large region with uniform colour, otherwise it is incoherent. Similarly to the case of generalized colour histograms, best performances are achieved for the HSV space.

Colour Correlogram

Colour correlogram was proposed to capture both colour distribution and the spatial correlation of pairs of colours. It consists of a table indexed by three-dimensional entries (i, j, k) that specify the probability of a pixel with colour j being at distance k from a pixel with colour i. If all the possible combinations of color pairs are considered the size of the colour correlogram is very large. Hence, a simplified version that considers only correlation between identical colours, called the colour autocorrelogram, is often used.

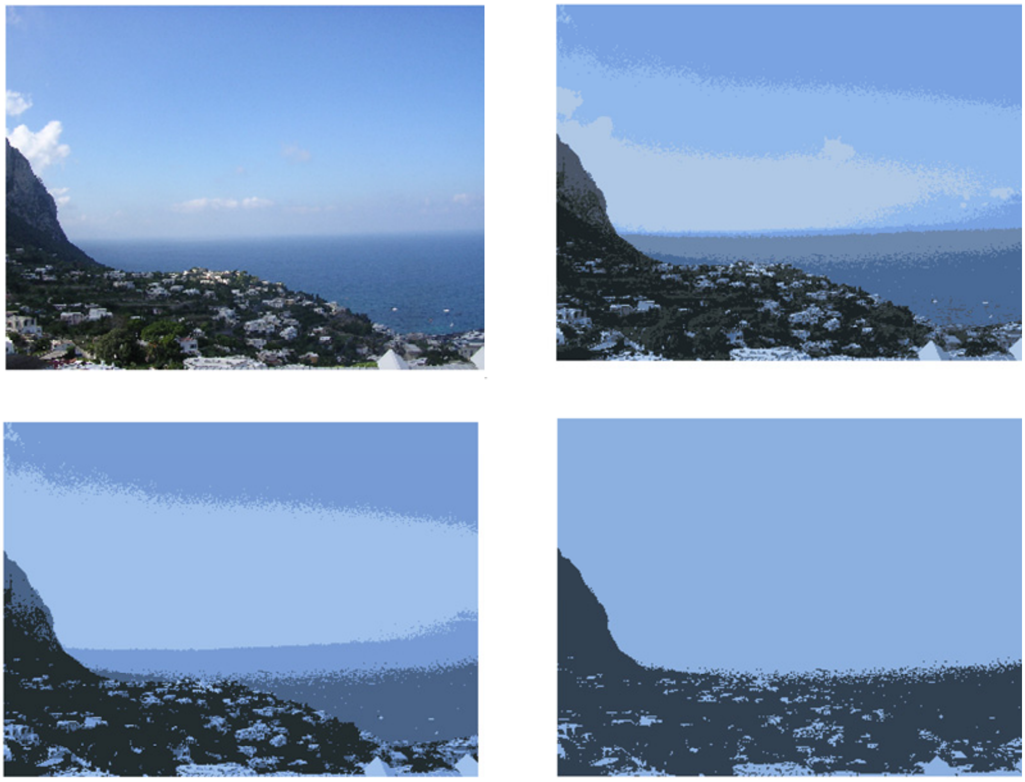


Figure 2.1: Quality of variously quantized images, with different number of histogram bins (top left to bottom right): the original, image quantized into eight, four and two bins.

2.2.2 Texture Descriptors

Texture is commonly identified as visual repeating patterns of varying intensity. It is a function of the spatial variation in pixel intensities and has been the subject of many studies. There are two main perspectives in defining texture: human and computer vision (Tuceryan and Jain 1988). In the human vision perspective various texture descriptions were evaluated against the human visual system. It has been established that second order statistics define a likelihood of observing a pair of gray values in an image at a random location and orientation. Hence, textons are defined as texture pairs with identical second-order statistics and they are used in texture discrimination. Other studies have proposed that the brain performs a multi-channel, frequency and orientation analysis of the visual image. They suggested that the visual system decomposes the image into filtered images of various frequencies and orientations. this was proved by De Valois et al. (1982) who determined that the response of cells in the visual cortex is a sinusoidal grating of various frequencies and orientations. Hence this has been the motivation for vision researchers to apply multi-channel filtering approaches to texture analysis.

Tamura Texture Feature

The Tamura texture features were designed in accordance with philological studies of human visual perception of texture (Tamura et al., 1978). These features include: coarseness, contrast, directionality, linelikeness, regularity, and roughness. The fist three components were used in early image retrieval systems as QBIC (Flickner et al., 1995) and Photobook (Pentland et al., 1996). Coarseness is a measure of granularity of texture, and it is connected with scale and repetition of patterns in an image. Contrast captures the dynamic range of grey levels and polarization of black and white colour distributions. Finally, directionality is a global feature of an image which tries to identify the total degree of directionality.

Grey Level Co-occurrence Matrix

The spatial grey level co-occurrence matrix (GLCM) estimates image properties related to second-order statistics. GLCM is a symmetric matrix of dimensions equal to the number of grey levels N, present in an image. Each element represents the frequency at which two pixels, separated by a certain displacement vector, occur in the image (Haralick 1979). $P_{\mathbf{d}}$ denotes a GLCM with displacement vector $\mathbf{d} = (dx, dy)$, where dx and dy are the distances for each coordinate. The entry (i, j) for the matrix $P_{\mathbf{d}}$

represents the number of occurrences of the grey levels pairs i and j with the displacement vector $\mathbf{d}$.

$$P_{\mathbf{d}}(i,j) = \left|\{((x,y),(p,q)) : I(x,y) = i, I(p,q) = j\}\right|.$$

Where $I(\cdot,\cdot)$ is grey level intensity at specific coordinates, $(p,q) = (x+dx, y+dy)$ and $|\cdot|$ denotes the cardinality of the set. The GLCM reveals certain properties about the spatial distribution of the grey levels for example if most of the entries in the co-occurrence matrix are concentrated along the diagonals, then the texture is coarse with respect to the displacement vector. A total of 14 scalar quantities have been preposed for summarizing the information contained in a co-occurrence matrix. However, typically only a subset of these is used. In Table 2.1 four features are considered: energy, entropy, contrast and homogeneity.

Table 2.1: Four texture features calculated from the grey level co-occurrence matrix.

Texture Feature	Formula
Energy	$\sum_i \sum_j P_{\mathbf{d}}^2(i,j)$
Entropy	$-\sum_i \sum_j P_{\mathbf{d}}(i,j) \cdot \log P_{\mathbf{d}}(i,j)$
Contrast	$\sum_i \sum_j (i-j)^2 P_{\mathbf{d}}(i,j)$
Homogeneity	$\sum_i \sum_j \frac{P_{\mathbf{d}}(i,j)}{1+(i-j)^2}$

The energy feature has a larger value when the co-occurrence frequencies are concentrated only at few locations in the matrix. This can happen along the diagonal for an image with constant grey level values or off- diagonal for structured images. A noisy image with random changes in grey levels will have a low energy value. The entropy feature is larger in images with evenly distributed values in the co-occurrence matrix; Therefore in noisy images the entropy is large. The contrast feature is larger for a co-occurrence matrix with higher off-diagonal values with varying intensity. Finally, the homogeneity feature is large for an image with more constant grey level patches, that is for co-occurrence matrix with large diagonal values.

Gabor Filter Feature

Gabor filters (GF) are very often used to extract texture, since they enable filtering in the spatial and frequency domain. Newsam and Kamath (2005) suggested use of Gabor filters to model the receptive fields of cells in the visual cortex for texture processing. The Gabor transform is a set of shift invariant directional filters. Since it produces much more coefficients than there are pixels in an image, it is redundant and hence, computationally costly. To overcome this disadvantage, Manjunath and Ma (1996) developed an effective feature representation based on first and second moments of transformed coefficients. An attractive mathematical property of Gabor functions is that they minimize the joint uncertainty in space and frequency. A two dimensional Gabor filter $g(x,y)$ corresponds to a sinusoidal wave of a certain frequency and orientation modelled with a Gaussian envelop:

$$g(x,y)=\left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi\ jWx\right],$$

where σ_x,σ_y are sizes of the Gaussian envelope in x,y directions, respectively. Gabor filters are frequency and orientation selective filters, with a corresponding Fourier transform $G(u,v)$:

$$G(u,v)=\exp\left\{-\frac{1}{2}\left[\frac{(u-W)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}\right]\right\}$$

Where $\sigma_u=1/2\pi\sigma_x$, $\sigma_v=1/2\pi\sigma_y$ and W denotes a higher center frequency of interest U_h. A class of Gabor filters is built by dilations and translations of the function $g(x,y)$:

$$g_{mn}(x,y)=a^{-m}g(x',y'),\ a>1,\ m=0,1,...,S-1,\ n=0,1,...,R-1$$

$$x'=a^{-m}(x\cos\theta+y\sin\theta),\ y'=a^{-m}(-x\sin\theta+y\cos\theta)\ \text{and}\ \theta=\frac{n\pi}{R}.$$

Where m,n are integers representing scale and orientation and S,R are total number of scales and orientations in the filter bank. The factor a^{-m} ensures that the energy is independent of scale m. The non-orthogonal property of Gabor filters introduces redundant information. In a design strategy for filters the redundancy is reduced by ensuring that the half-peak magnitude support of the filter responses touches each other in the frequency spectrum (see Figure 2.2). Let U_l, U_h be central lower and upper frequencies of interest. The design strategy models the values for a, σ_u and σ_v as functions of U_l, U_h, scale and orientation. The Gabor transform values of an image are defined as:

$$W_{mn}(x,y) = \int I(x,y) g^*_{mn}(x - x_1, y - y_1) dx_1 dy_1$$

Where $I(x,y)$ denotes grey value in an image, and $(^*)$ denotes a complex conjugate. Mean value μ_{mn} and standard deviation σ_{mn} of transformed coefficients are used to construct the feature vector (Manjunath and Ma, 1996):

$$\mathbf{x} = [\mu_{00}\ \sigma_{00}\ \mu_{01}\ \sigma_{01}\ \ldots\ \mu_{S-1R-1}\ \sigma_{S-1R-1}] \qquad (2.1)$$

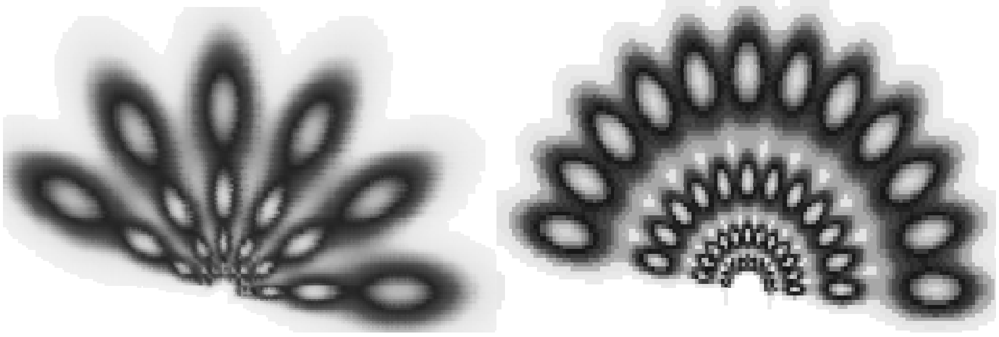


Figure 2.2: Magnitude of Gabor filter responses for scale S=4 and total number of orientations R=6 (left), R=12 (right);

2.2.3 Shape Descriptors

In order to enable extraction of shape features, in many applications a pre-processing step of object segmentation is required. Two main categories for shape features have been defined in literature: contour-based and region-based.

Contour-based shape features describe objects by using only information along the object boundary, e.g. Fourier descriptor for shape (Persoon and Fu, 1977), curvature scale-space representation of a contour (Mokhtarian et al., 1996).

Region-based shape descriptors characterize spatial distribution of both boundary and interior pixels. They can describe complex objects consisting of multiple disconnected regions as well as simple objects with or without holes.

Within the MPEG-7 standard a descriptor based on angular radial transform (ART) is used to decompose an image into a set of orthogonal two-dimensional complex basis functions (Manjunath et al., 2003). The feature vector is composed of magnitudes of the complex ART coefficients for a number of angular and radial basis functions. This feature is scale and orientation invariant. Other region based approaches include shape descriptors based on geometrical moments. A comprehensive overview of shape features is reported by Gevers and Stokman (2003).

2.3. Similarity Measures

Comparing feature sets based on similarity function allows meaningful interpretation of low-level numerical features. This is achieved by providing some level of correlation between low-level similarity and human defined perceptual similarity. In CBIR systems the retrieval result is a list of images ordered by increasing dissimilarity to the query image or images. Retrieval performance is not only influenced by quality of the content representations but also by different similarity measures. A number of similarity measures for image retrieval based on empirical estimates of the distribution of features have been developed in recent years.

In this section, a brief review of the distance measures is given through use of the following notation. Let X be a feature space endowed with a similarity measure d. Observe that d is a distance function, in case d is a metric the feature space (X, d) becomes a metric space (see Appendix B, Definition B.1). Let $\mathbf{x}_i$ be a i-th vector element of X with dimension N. The dimension of the feature space depends on the space itself $X \equiv \mathbb{R}^N$. A feature vector is $\mathbf{x}_i \in X$, $\mathbf{x}_i = [x_{i,1}, x_{i,2}, ..., x_{i,N}]$.

Minkowski form distances

Assuming independence of feature vectors, the Minkowski form distance L_p for a discrete n-dimension feature space $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^N$, is defined as:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{r=1}^{N} \left| x_{i,r} - x_{j,r} \right|^p \right)^{1/p} \tag{2.2}$$

For $p = 1$, the norm in (2.2) represents the Manhattan distance (L_1 distance) and for $p = 2$ it represents the well known Euclidian distance (L_2 distance). For $p = \infty$, the special case of the Minkowski distance leads to the Chebychev distance:

$$d(\mathbf{x}_i,\mathbf{x}_j)=\max_{r\in\{1,\dots,N\}}\left|x_{i,r}-x_{j,r}\right|=\lim_{p\to\infty}\left(\sum_{r=1}^{N}\left|x_{i,r}-x_{j,r}\right|^{p}\right)^{1/p} \tag{2.3}$$

A number of CBIR systems such as MARS (Rui et al., 1997), NeTra (Ma and Manjunath, 1997), Blobworld (Carson et al., 2002) have used Minkowski form distance as the similarity measure.

Histogram Intersection

Swan and Ballard (1991) proposed histogram intersection (HI) distance for colour image retrieval. This distance metric is a version of the L_1 metric that deals with partial matches. The histogram intersection distance of two N -dimensional histograms $\mathbf{x}_i$ and $\mathbf{x}_j$, is defined as:

$$d(\mathbf{x}_i,\mathbf{x}_j)=1-\frac{\sum_{r=0}^{N}\min(x_{i,r},x_{j,r})}{\sum_{r=0}^{N}\max(x_{i,r},x_{j,r})} \tag{2.4}$$

Colours which are not present in one of the histograms do not contribute to the intersection value, consequently background colours do not influence to the overall distance. In case two histograms are identical the intersection is 1 and distance 0.

Weighted-Mean-Variance

Manjunath and Ma (1996) proposed this empirical distance for Gabor filter features (2.1). Empirically, the values for means μ_{mn} and standard deviation σ_{mn} of transformed coefficients in an image of size $m\times n$ are normalized by a standard deviation $\sigma(\cdot)$ of appropriate values.

$$d(\mathbf{x}_i,\mathbf{x}_j)=\sum_m\sum_n\left|\frac{\mu_{mn}(x_i)-\mu_{mn}(x_j)}{\sigma(\mu_{mn})}\right|+\left|\frac{\sigma_{mn}(x_i)-\sigma_{mn}(x_j)}{\sigma(\sigma_{mn})}\right|. \tag{2.5}$$

Quadratic form distance

Very often, feature components are not independent with different levels of importants. To integrate this into the distance function the quadratic distance is defined as:

$$d(\mathbf{x}_i,\mathbf{x}_j)=((\mathbf{x}_i-\mathbf{x}_j)^T A(\mathbf{x}_i-\mathbf{x}_j))^{1/2} \tag{2.6}$$

Where $A=[a_{ij}]$ is a cross-correlation matrix, denoting similarity between the i*-th* and j*-th* feature components. This metric has been used for colour histogram retrieval, since it incorporates cross-bin similarity and leads to better results than simple Minkowski or histogram intersection distance.

Mahalanobis distance

The Mahalanobis distance takes in account various levels of correlation between components of feature vectors (Wilson and Martnez, 1997):

$$d(\mathbf{x}_i,\mathbf{x}_j)=((\mathbf{x}_i-\mathbf{x}_j)^T C^{-1}(\mathbf{x}_i-\mathbf{x}_j))^{1/2} \tag{2.7}$$

Where C is the covariance matrix of feature elements. The matrix entry at postion (i,j) corresponds to cross-covariance value between two n-dimensional feature vectors $Cov(i,j)=E[(\mathbf{x}_i-E[\mathbf{x}_i])^T(\mathbf{x}_j-E[\mathbf{x}_j])]$ where E is the expected value.

Kullback-Leibler Divergence and Jeffrey-Divergence

The Kullback-Leibler (KL) divergence measures the difference between two probability distributions $p(\mathbf{x}_i)$and $q(\mathbf{x}_j)$ for feature vectors $\mathbf{x}_i,\mathbf{x}_j$. In image retrieval, histograms are frequently used to obtain nonparametric estimators of empirical feature distributions. The histogram value $f(r,\mathbf{x}_i)$ corresponds to the number of image pixels in bin r for feature $\mathbf{x}_i$.

$$d_{KL}(\mathbf{x}_i,\mathbf{x}_j)=p(\mathbf{x}_i)\log\frac{p(\mathbf{x}_i)}{q(\mathbf{x}_j)}=\sum_r f(r,\mathbf{x}_i)\log\frac{f(r,\mathbf{x}_i)}{f(r,\mathbf{x}_j)}. \tag{2.8}$$

This distance is not a metric as it is not symmetric and does not satisfy the triangle inequality. Hence, its symmetric version the Jeffrey-divergence (JD) is often used:

$$d_{JD}(\mathbf{x}_i,\mathbf{x}_j)=\sum_r\left(f(r,\mathbf{x}_i)\log\frac{f(r,\mathbf{x}_i)}{\hat{f}(r)}+f(r,\mathbf{x}_j)\log\frac{f(r,\mathbf{x}_j)}{\hat{f}(r)}\right),$$

where $\hat{f}(r)=\dfrac{f(r,\mathbf{x}_i)+f(r,\mathbf{x}_j)}{2}$ is the mean histogram (Rubner et al., 2001).Similarly another frequently used statistical distances is the Chi-square distance (Michalski et al., 1981)

Earth Mover's Distance

The Earth Mover's Distance (EMD) is a flexible method for calculating similarity between multidimensional distributions in a feature space (Rubner et al., 1998). EMD defines the minimum effort to be made for transferring one feature signature to another. It represents a transportation problem that can be solved by linear optimisation algorithms. Given two distributions, one can be interpreted as the mass of "earth" spread in space and the other as collection of "holes" in the same space. The EMD gives a measure of least amount of work needed to transfer the earth into the holes. For two feature signatures $\mathbf{x}_i = \{(\overline{\mathbf{x}}_{i,r}, w_{\overline{\mathbf{x}}_{i,r}})\}, \mathbf{x}_j = \{(\overline{\mathbf{x}}_{j,s}, w_{\overline{\mathbf{x}}_{j,s}})\}, r = 1,..,m, s = 1,..,n$, the feature representatives and weights are denoted as $\overline{\mathbf{x}}_{i,r}, \overline{\mathbf{x}}_{j,s}$ and $w_{\overline{\mathbf{x}}_{i,r}}, w_{\overline{\mathbf{x}}_{j,s}}$, respectively. In case of feature signatures representing histograms the weights are histogram bin values. The cost of moving a unit of a single feature ("earth") representative in the feature space is defined with some ground distance d_{rs} (e.g., L_2 distance). In this case the distance between two signatures is the sum of minimal costs needed to move individual features:

$$d_{EMD}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{r,s} g_{rs} d_{rs}}{\sum_{r,s} g_{rs}}, \qquad (2.9)$$

Where d_{rs} is the ground distance between the r-th and s-th feature representative of feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. The optimal flow $g_{rs} \geq 0$ between two features is defined such that the total cost $\sum_{r,s} g_{rs} d_{rs}$ is minimized, under constraints:

$$\sum_s g_{rs} \leq w_{\overline{\mathbf{x}}_{i,r}}, \sum_r g_{rs} \leq w_{\overline{\mathbf{x}}_{j,s}},$$
$$\sum_{r,s} g_{rs} = \min(\sum_r w_{\overline{\mathbf{x}}_{i,r}}, \sum_s w_{\overline{\mathbf{x}}_{j,s}}).$$

The fist constraint limits the amount of "earth" that could be transferred from $\mathbf{x}_i$ to $\mathbf{x}_j$ to its weight. The second constraint limits the amount of "earth" that could be transferred into the holes, represented with feature $\mathbf{x}_j$. The last constraint defines the total flow, and maximizes the amount of "earth" to be moved. The total flow is also used as a normalisation factor in (2.9), so that matching of parts with different total flows could be done. EMD can be applied to the more general variable-size features that are present, for instance, in images segmented into different number of regions.

Aksoy and Haralick (2000) assume two models of similarity measures: probabilistic and geometric. The probabilistic similarity measure is defined as the likelihood of ratio for two conditional probabilities. The probability that a particular distance between two feature vectors occurs when images belong to the relevant class, and accordingly the same probability when images belong to the irrelevant class. To estimate the conditional distributions, multivariate normal distribution or less restrictive fitted distributions were used. Results show that L_1 distance generally performs better than L_2 in a retrieval scenario, and that metrics taking into consideration sample distribution perform better than purely geometric measures.

Li et al. (2003) defined a perceptual dissimilarity function based on human perceptual similarity. Though not a metric, this distance finds the dissimilarity between two vectors based on vectors of reduced dimensionality. This is achieved by keeping only a certain number of the smallest absolute differences among feature vector components. Improved performance were also shown over the fractional measure as suggested by Aggarwal et al. (2001), this measure is based on Minkowski distance for parameter values $0 < p < 1$.

Rubner et al. (2001) present quantitative performance evaluations for a variety of dissimilarity measure presented in this section and different scenarios such as classification, image retrieval, and segmentation. Comprehensive overviews of existing similarity measure can also be found in Santini and Jain (1999) and Long et al. (2002).

2.4. The MPEG-7 Framework

The MPEG-7 Standard is defined as standard multimedia content description interface offering a set of audio-visual descriptions in an effort to provide standardized tools for describing multimedia content (Martinez, 2001; Manjunath et al., 2001). In order to provide an efficient and human compliant visual content representation, semantic high-level description of content is needed. The main initiative for standardisation of image descriptions originated from the Moving Picture Expert Group (MPEG) that developed the MPEG-7 standard. Chang et al. (2001) stated that the main goal of MPEG-7 is to enable interoperability among systems and applications used in generation, management, distribution and consumption of audio-visual content.

MPEG-7 has a number of normative elements, including audio–visual Descriptors, Description Schemes and Description Definition Language (Sikora, 2001). The Description Definition Language is a standardized language for the definition of additional Descriptors and Description Schemes. Descriptors define syntax and semantics of features for audio-visual content with different levels of abstraction. They may include features of low-level abstraction such as colour, texture, shape, motion or high-level abstractions such as events, concepts etc. Descriptor representations, as definitions of low-level content characteristics, belong to the normative part of the MPEG-7. Though extraction of these descriptors and similarity matching is not normative, a detailed description of the recommended methods for extracting and matching are presented in the visual XM document (MPEG, 2001) as a non-normative part of the MPEG-7 standard. Low-level characters of image content are the basis for generating visual descriptors in the MPEG-7 standard. Several commonly used basic visual MPEG-7 descriptors are given in Table 2.2.

Table 2.2: Basic MPEG-7 visual descriptors for still image characterisation.

Colour	Texture	Shape
Colour Layout Scalable Colour Colour Structure Dominant Colour	Texture Browsing Homogeneous Texture Edge Histogram	Contour Shape Region Shape

In several approaches, statistical properties of the MPEG-7 descriptors and their performances with different metrics were analysed. Eidenberger (2003) investigated statistical properties of MPEG-7 descriptors such as redundancies, sensitivity to changes in content and completeness (overall coverage of content). However, not all of visual media objects can be fully captured by MPEG-7 descriptors. Therefore it was suggested that additional descriptors to MPEG-7 ones should be used for content-based retrieval and browsing applications. Ojala et al. (2002) compared MPEG-7 colour descriptors with colour autocorrelogram features and obtained the best performances for MPEG-7 colour structure descriptor. In a similar manner Eidenberger (2004) also tested MPEG-7 descriptors with their recommended distances and with several other measures in line with human psychological factors. He showed that MPEG-7 recommended distances, specified for retrieval and browsing scenarios give in general better performances. Hence, when using any MPEG-7 descriptor in this thesis the appropriate distance exploiting syntactical meaning of the descriptor is used. The descriptors used for experiments in this thesis are presented in more details in the following subsection.

2.4.1 Colour and Texture Descriptors

The following descriptors are a part of the visual descriptors in the MPEG-7 Standard, they are designed for specific retrieval and browsing purposes. These descriptors represent variations of classical features described in section 2.2.1. Since they are part of the set of descriptors used in this thesis, they are presented in more details. Accompanying similarity measures tuned to the particular implementation of the descriptor within the MPEG-7 standard are also presented.

Colour Layout Descriptor

Colour Layout Descriptor (CLD) is a compact descriptor designed to capture the representative colours in an image or an arbitrary-shaped region (Manjunath et al., 2003). Even though, it is derived from global colour histograms, it also incorporates a number of localized histograms. Every image is partitioned into 8x8 blocks to achieve resolution and scale invariance. Average colour of each block is calculated and the discrete cosine transform (DCT) applied on this set of colours. For each of the three components in YCrCb colour space 64 coefficients are obtained. The coefficients are scanned and only the first few are non-linearly quantized. Scalable representation of the descriptor is enabled by controlling the number of coefficients. The number of coefficients is chosen from the following set {3, 6, 10, 15, 21, 28, 64}. This descriptor can be presented with the following vector:

$$\mathbf{x} = \{\mathbf{DY}, \mathbf{DCr}, \mathbf{DCb}\} \tag{2.10}$$

with each sub vector representing coefficients of a particular colour component from the YCrCb colour space. For the CLD the MPEG-7 standard recommended a similarity measure which is a weighted version of the Euclidian L_2 distance (2.2) :

$$d_{cld}(\mathbf{x}_i,\mathbf{x}_j)=\sqrt{\sum_k \omega_{yk}(DY_{i,k}-DY_{j,k})^2}+\sqrt{\sum_k \omega_{rk}(DCr_{i,k}-DCr_{j,k})^2}+\sqrt{\sum_k \omega_{bk}(DCb_{i,k}-DCb_{j,k})^2} \tag{2.11}$$

This is not a very complex descriptor and as a consequence the speed of extracting is rather high. The weights of DCT coefficients for each component are designed to better simulate human visual system, with larger weight for lower frequencies.

Colour Structure Descriptor

The colour structure descriptor (CSD) describes colour distribution and local colour structure in an image. This descriptor is constructed by scanning a colour quantized image with 8x8 structure window and counting the number of times a particular colour appears in the structuring window. A colour histogram of 256 bins is then generated in Hue-Max-Min-Diff (HMMD) colour space defined by MPEG-7 standard. Additional bin unification may be necessary in case the number of desired bins is less that 256. This descriptor represents a one-dimensional array of eight bit quantized values:

$$\mathbf{x}=h_s(m),\, m\in\{1,...,M\} \tag{2.12}$$

h denotes a histogram, M can take values 256, 128, 64 and 32, and s is the scale of the structuring window (Manjunath et al., 2003). CSD is a generalized case of global colour histogram since it reduces to it when structure window of size 1x1 is used. The distance metric recommended for CSD is a normalized version of Manhattan L_1 distance (2.2).

Dominant Colour Descriptor

The dominant colour descriptor (DCD) specifies a set of representative colours in an image or a region. Colours are clustered into a small number of representative colours and then quantized. Since the number of clusters varies depending on the image content, the dimensionality of the descriptor is not fixed. The feature vector can be presented with a number of optional coefficients (Manjunath et al., 2003). It consists of sets of elements describing each dominant colour with the overall dimension of the feature vector depending on the image content itself. The feature vector is made up of 4-touples of elements:

$$\mathbf{x}_i=\{(x_{i,r},\mathbf{c}_{i,r})\},\ r=1,...,N(\mathbf{x}_i), \tag{2.13}$$

where $\mathbf{c}_{i,r}$ is the r-th 3-dimensional colour component in RGB colour system, $x_{i,r}$ is the percentage of pixels that have corresponding colour values for the r-th dominant colour. Notice that this is an example of a feature space with variable dimension and the similarity measure is not a metric since $d_{dcd}(\mathbf{x}_i, \mathbf{x}_i)$ is not necessarily zero. That is, this feature space is not a conventional vector or even metric space. Furthermore, the dimension of each feature $N(\mathbf{x}_i)$ is variable, which makes this feature space more difficult to handle. The distance measure for the DCD is the quadratic form distance (2.6):

$$d_{dcd}(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{r=1}^{N(\mathbf{x}_i)} \sum_{s=1}^{N(\mathbf{x}_j)} ({x_{i,r}}^2 a_{rr} + {x_{j,r}}^2 a_{ss} - 2 x_{i,r} x_{j,s} a_{rs}) \right)^{1/2} . \qquad (2.14)$$

Here, a_{rs} is the similarity coefficient between two colours $a_{rs} = 1 - \|\mathbf{c}_r - \mathbf{c}_s\| / d_{max}$ for $\|\mathbf{c}_r - \mathbf{c}_s\| \le T_d$, otherwise zero. T_d is the maximal distance for the two colours to be considered similar and $d_{max} = \alpha \cdot T_d$. As recommended in MPEG-7 standard, T_d takes values between 10 and 20 and α between 1.0 and 1.4 (Manjunath et al. 2003).

Edge Histogram Descriptor

The Edge Histogram Descriptor (EHD) describes local edge distribution of an image. After dividing an image into 4x4 sub-images and detection of edges, local edge histograms are calculated. Five types of edges are defined (horizontal, vertical, diagonal 45 degrees, diagonal 135 degrees and non-directional). A histogram with 80 bins is calculated by dividing each sub-image into image-blocks and using edge detectors to classify each block into one of the five categories (Martinez, 2001).

For matching purposes a version of the Manhattan L_1 distance (2.2) and local, semi-global and global edge histograms of the input features are used:

$$d_{ehd}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=0}^{79} \left| x_{i,r} - x_{j,r} \right| + 5 \times \sum_{r=0}^{4} \left| x_{i,r}^{g} - x_{j,r}^{g} \right| + \sum_{r=0}^{64} \left| x_{i,r}^{s} - x_{j,r}^{s} \right| \qquad (2.15)$$

The appropriate coefficients from left to right in (2.15) represent normalized histogram values $\mathbf{x}_i, \mathbf{x}_j$ of the two images, as well as their global and semi-global versions. In order to emphasize the meaning of global edge histogram the weight factor 5 is added.

Bin values for the global edge histogram are computed from 80 local bins by obtaining cumulative distribution over the whole image for each type of edge. As for the semi-global histogram, sub-images are grouped into 13 different segments as shown in Figure 2.3. To form semi-global histograms edge histograms, for each type of edge, are summed over these segments.

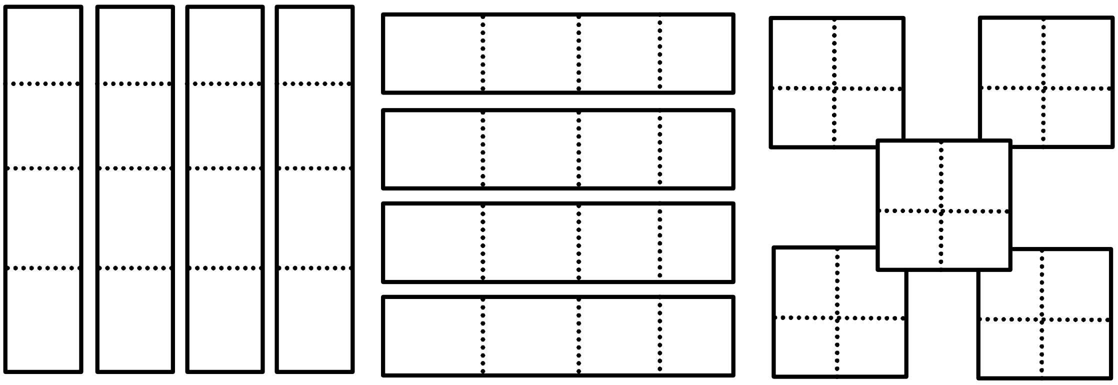


Figure 2.3: Thirteen segments used to generate semi-global histograms (Manjunath et al., 2003).

Homogenous Texture Descriptors

Homogenous Texture Descriptors (HTD) is a derivation of Gabor filter features. The extraction procedure is a result of standardisation activity, tests for retrieval and browsing applications and is limited to a filter size of 128x128. At first, the image is filtered with a bank of orientation and scale sensitive filters (Bober, 2001). Then the frequency domain is partitioned into 30 channels, and modelled by two-dimensional Gabor functions. Finally, the energy and the energy deviation of each channel are computed, and logarithmically scaled to obtain values for the mean energy, e_i, and the mean deviation, d_i, of the i^{th} channel. Mean and standards deviation of the whole image are denoted as f_{DC}, f_{SD}. The descriptor is formed as follows:

$$\mathbf{x} = [f_{DC}, f_{SD}, e_1, e_2, ..., e_{30}, d_1, d_2, ..., d_{30}] \tag{2.16}$$

The distance function is the weighted absolute L_1 distance between two sets of feature vectors:

$$d_{htd}(\mathbf{x}_i, \mathbf{x}_j) = \sum_r \left| x_{i,r} - x_{j,r} \right| / a(r) \tag{2.17}$$

Where $a(r)$ is the standard deviation of appropriate features components.

2.5. Feature Evaluation

In a CBIR scenario, each relevant class was taken in turn to perform a content-based query by example (QbE). Different feature vectors were used as well as different similarity measures, for a number of ground truth databases with varying sizes of the relevant class.

2.5.1 Feature Selection

Descriptors used for analysis consist of subsets of the considered set of visual descriptors. The decision about the choice was made based on the extraction method, desired properties and discrimination power. They include CLD, CSD, DCD, EHD, HTD, HSV histogram and GLCM. Note that the first five descriptors are MPEG-7 descriptors while the remaining two descriptors are used in an effort to achieve better representation and discrimination of visual features. More details for the dimension, parameters and the similarity measures of used features are given in the following paragraph.

- CLD: the feature is given in (2.10), and with the following combinations of dimensions for each colour space component $\{28, 15, 15\}$. The distance function is the recommended MPEG-7 distance (2.11).
- CSD : the feature is given in (2.12) , here M equals 32 and 64 with a scanning window of size 8x8.
- DCD: The feature is given in (2.13). The distance function used is given in (2.14) with following parameter values for $T_d \approx 15$ and $\alpha = 1.2$.
- EHD : the feature is a 80 bin histogram and (2.15) is used as a version of L_1 distance with transformations of the original input feature into global and semi-global histograms.
- HTD: this feature was considered in part of this thesis when retrieval on whole images was performed, i.e. in case images larger than 128x128, since this is the minimal accepted filter size. Additionally the descriptor dimension is set to 32 or 62.
- HSV histogram: for this feature 32 bins are used for H component and 30 bins for S component. The histogram intersection distance (2.4) was used as the similarity measure.

- Finally, GLCM uses four normalised values from Table 2.1. The distance function being the L_2 distance.

Furthermore, a combination of descriptors has the potential to capture various levels of relevant information.:

- 'CLEH': combination of CLD and EHD, the dimension of this feature space is 138.
- 'CLCSEHHT': combination of CLD, CSD, EHD and HTD, the dimension of this feature space is 202.
- 'CONC': combination of CLD, CSD, HSV, EHD, HTD and GLCM, the dimension of this feature space is 268.

This combinations were chosen based on experimental results as best performing combinations. Since the DCD had worse results in retrieval with individual descriptors and appropriate similarity measures, it has been excluded from the joint combination.

In all of these cases the distance function used for retrieving similar images, is a linear combination of normalized distances per each feature (see Appendix B, Definition B.2).

Having a number of descriptors per image can lead to "the curse of dimensionality" (Bellman, 1961). This refers to the exponential growth of the quantity of training points required to describe data, depending on the dimensionality of the input variables. However, the amount of training data is limited; hence increasing the dimensionality of the input space can lead to poor generalization capabilities. Though there are a number of methods to decrease dimensionality this is not the aim of this thesis. Hence the dimension of the feature space was kept limited and the focus was redirected to learning algorithms behind the relevance feedback method in CBIR scenarios.

2.5.2 Ground Truth Image Databases

Ground truth image databases that reflect on a wide range of contexts are considered in this subsection. Since there are countless possibilities in defining ground truth classes based on different subjective criteria, adapting the retrieval approach to a specific context does not necessarily guarantee improvement of performances. Several ground truth databases (depicted in Appendix A) are selected for evaluation. They range in image type, image size, application domain and category size:

- *DColour* database has 5 colour distinctive classes from the Corel dataset (Corel Corporation, 1990). This database contains "easy" classes used for effective colour descriptor evaluation (see Appendix A, Figure A. 1).
- *VisTex* database (VisTex, 2002), collection of high quality texture images. Two main components are available in the database: reference textures of more than 100 images of homogeneous textures and texture scenes for images with multiple textures in real world scenes (see Appendix A, and Figure A. 2 ; Table A. 1).
- *D25-1800* database is composed of 25 distinct classes from the Columbia color database (Nene et al., 1996). Each image is a 128 x 128 pixel representation of 72 views for an object. The images were taken for objects on a turn table against black background with view angle every 5° of a 360° rotation (see Appendix A, Figure A. 3). The membership of each object to a class is not ambiguous. Hence this image database is often used in image recognition. However this database also allows for effective estimation of relevance feedback approaches.
- *D8* database is a subset of the ETH-80 database (Leibe and Schiele, 2003). It consists of 80 objects from 8 different categories (apple, tomato, pear, toy-cows, toy-horses, toy-dogs, toy-cars and cups). In the original ETH-80 image database each object is represented by 41 images from different viewpoints. A subset of the available views, 7 views per object, was used in this evaluation database (see Appendix A, Figure A. 4).
- *D7-700* database has 700 images with a variety of simple and complex classes from the Corel database (Corel Corporation, 1990). The categories overlapping in meaning with different numbers of images per class, which are: buildings, clouds, cars, elephants, grass, lions and tigers. The number of ground truth images per class is 141, 264, 100, 100, 279, 99 and 100 respectively. There is a wide range of low-level visual diversities within each class (see Appendix, Figure A. 5). This makes this database a good candidate for search with relevance feedback.
- *Caltech 101* image database (Fei-Fei et al., 2004) consists of images for objects belonging to 101 categories including a background category. There are about 40 to 800 images per category. The size of each image is approximately 300 x 200 pixels (see Appendix A, Figure A. 6; Table A. 2).

2.5.3 Performance Measures

The evaluation of image retrieval is a necessary step for the successful use of retrieval systems and their practical applications. Evaluation is based on retrieval of the most similar neighboring samples to the query sample. The measures most frequently used are precision and recall. Each relevant class item is taken and a content-based query by example search. Evaluations were based various scopes with different feature vectors, different similarity measures, and for various size of relevant classes. The following set of values is often used to characterize CBIR system:

- the number of images relevant for a particular query A,
- the number of images irrelevant for a particular query B,
- the number of retrieved images D,
- the number of relevant retrieved images, E.

Through the use of human subjects, the unlabelled database $C = A + B$ can be turned into the ground truth.

Precision is the ratio of the number of relevant retrieved images to the total number of retrieved images:

$$P = \frac{E}{D} \tag{2.18}$$

Recall is the ratio of the number of retrieved relevant images to the total number of relevant images in the database:

$$R = \frac{E}{A} \tag{2.19}$$

High recalls correspond to a better answer of the system to the query. However, this measure alone is not enough to qualify the quality of the system. Best results correspond to high values of both precision and recall. As it can be noticed from the equations above precision and recall are not independent measures, for a retrieved set D and total relevant set A:

$$\frac{P}{R} = \frac{A}{D} \tag{2.20}$$

These values depend on the size of the retrieved set, if D is considerably less than the size of the relevant set A the values for recall can never be high, similarly if D is much larger than A the precision can never be high. By additionally normalizing the retrieved set of images to the total relevant set, focus of attention is restricted along the diagonal of precision-recall graphs where $P = R$.

Huijmans and Sebe (2003) showed that precision-recall diagrams depend on the size of the ground truth classes with respect to the size of the database. They showed that the performances degrade for large image databases. A new measure, generality, representing the amount of relevant images in the whole database is used in conjunction with precision and recall. The same effect can be achieved for two-dimensional precision-recall curves by keeping parameters constant or performing experiments for different sizes of databases.

In this work precision-recall approach was used, with several ground truth databases of different size and content. Precision values are averaged by using constant "relevant scope" values, rather than using constant recall values. Relevant scope is defined as the ratio between relevant images retrieved and the relevant class size. Averaging by constant relevant scope allows equal possibilities for precision-recall performances for each concept, regardless of the relevant class size (Huijmans and Sebe, 2003).

2.5.4 Normalization of Image Features

Though it can be stated that any value stored in a computer is discrete at some level, different attribute types can be distinguished within a feature, forming a descriptor with syntactical meaning.

In case one of the input features has a relatively large range, then it can overpower other features within a joint combination of features. If common L_1 and L_2 distances are used, features having a larger number of dimensions will have more weight in the overall distance compared to features with less dimensions. Usually either feature vectors or distances are normalized.

Aksoy and Haralick (2000) investigated several approaches for feature normalization. For example ddistances are often normalized by linear scaling to unit range [0, 1] by using upper and lower bounds of the value in question. Alternatively, in an effort to avoid outliers, scaling to unit variance is often used. This is performed by shifting and rescaling a random variable to a normal distribution with zero mean and unit variance. However since a strong Gaussian assumption has to be made and this is often not the reality, prior domain knowledge of the feature and similarity measure type is valuable information. Feature normalization by fitting feature histograms with well-known distributions as Normal, Lognormal, Exponential and Gamma densities was also investigated.

The aim of both feature and distance normalization is to create conditions such that, without prior knowledge, all image features become equally important when contributing to the overall distance between two images. This relation is based on the assumption that the distance between two image descriptors is a random variable with a Gaussian distribution.

In Figure 2.4 estimated probability density functions for each distance measure for features are depicted. The experiments were performed on random sets of images with more than 244650 image pairs. The histogram bins are normalized with the overall number of samples times the range of each bin. This normalisation results into a relative histogram corresponding to the probability density function for each feature. Each of the distribution is closely fitted with a mixture of Gaussian (Djordjevic, and Izquierdo, 2006a). Distances for each distribution are normalize to the unitary range.

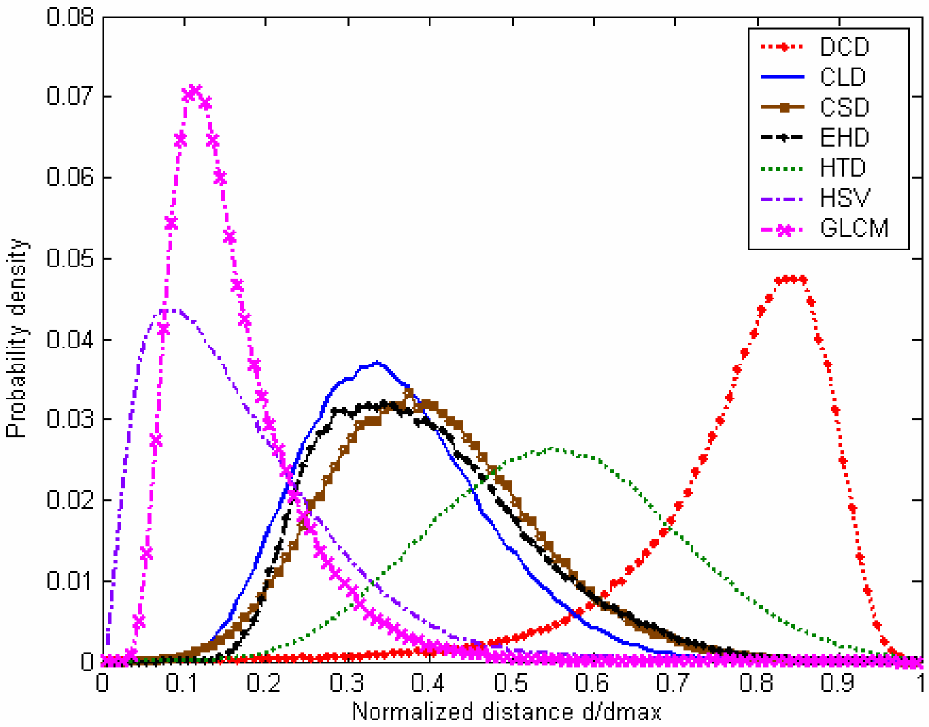


Figure 2.4: Probability density functions for distances between pair of images for different descriptors.

2.5.5 Feature Evaluation

The following precision-recall curves were calculated for individual descriptors and their combinations described in section 2.5.1, and for the six databases described in section 2.5.2. In all cases appropriate similarity measures and their linear combinations are used.

In Figure 2.5 precision–recall curves which are averaged per relevant scope, are shown for the DColour database. Each element of a class is considered as a query image and the most similar images are retrieved. The precision-recall curves show average retrieval performances for different descriptors and similarity measured over all five classes for the DColour database. Since the considered database is based on colour, individually CLD and CSD have the highest precision values for constant values of recall. These combinations of individual descriptors show slight improvement over the best performing individual descriptors, however the computational complexity and the dimensionality of the feature space is considerably increased. This is an example of a database where low-level similarity correlates to categories, and while as this is not a real life scenario it enables estimation of low-level descriptors. However, DCD gives the worst performances not being able to discriminate among different classes.

In case of the VisTex database, the highest precision is obtained as expected for combination of colour and texture descriptors, see Figure 2.6. Since this is a mainly texture database with both relevant colour ad texture information (see Figure A. 2).

For the D8 database, representing an object database with uniform background, combinations of both colour and texture descriptors lead to higher precision accuracy. Furthermore, clear distinction of several individual descriptors CLD, CSD and EHD, can be noticed see Figure 2.7.

For the D25-1800 database, the images have homogeneous background but with difference in scale and viewing angles per object. The highest performances were detected for several individual descriptors: CSD, HSV histogram and CLD, in descending order. The three feature combinations show improvement comparing to results for CLD, they also give worse results than HSV histogram, see Figure 2.8. Hence the colour is the dominant feature used here, with best performing CSD that incorporates a level of spatial information. The additional texture descriptors add noise since this database is composed of single object images with different camera angles and illumination. Since the background is uniform and the texture depends on the two mentioned conditions, colour descriptors are more resilient to these types of changes and thus give better performances.

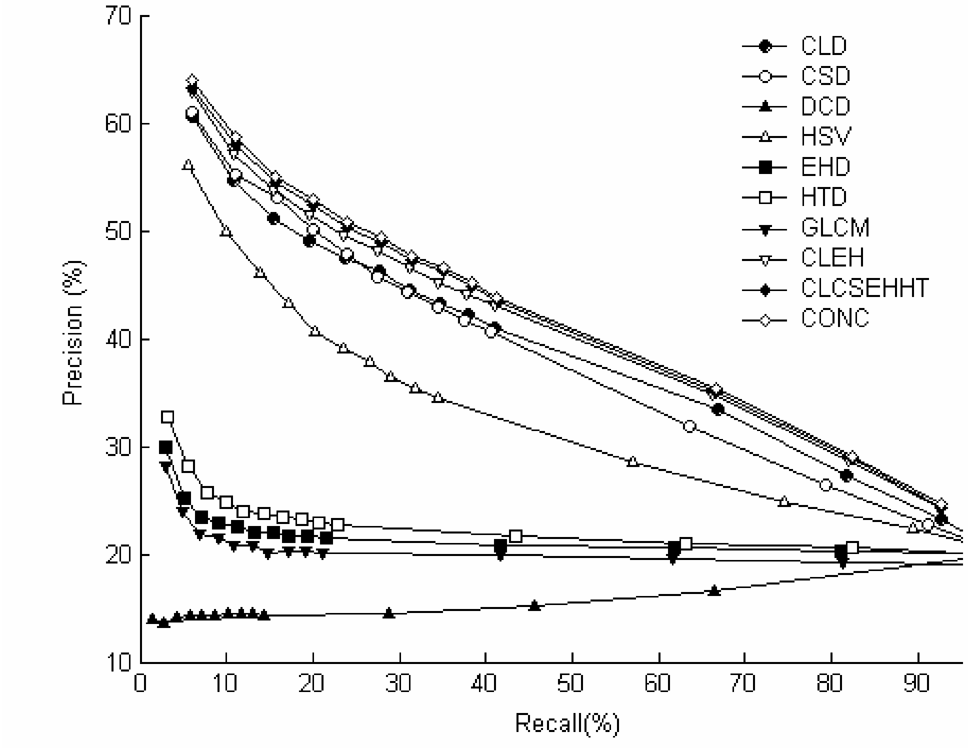


Figure 2.5: Average Precision-recall over all classes, for selected individual descriptors and three feature combinations (DColour database).

Figure 2.9 and Figure 2.10 show results for two real life databases, with very complex classes, these are D7-700 and Caltech 101 (see Figure A. 5, Figure A. 6). For the Caltech 101 database all three descriptor combinations have very similar results, therefore only one combination was depicted for clarity of presentation, CLEH.

For Figure 2.9, the D7-700 database, the backgrounds as well as the content within the same category is very diverse. In this case the DCD shows considerably better performance than the other individual descriptors as well as descriptor combinations. Hence the type of the database is of extreme importance and influences the quality of results for each descriptor.

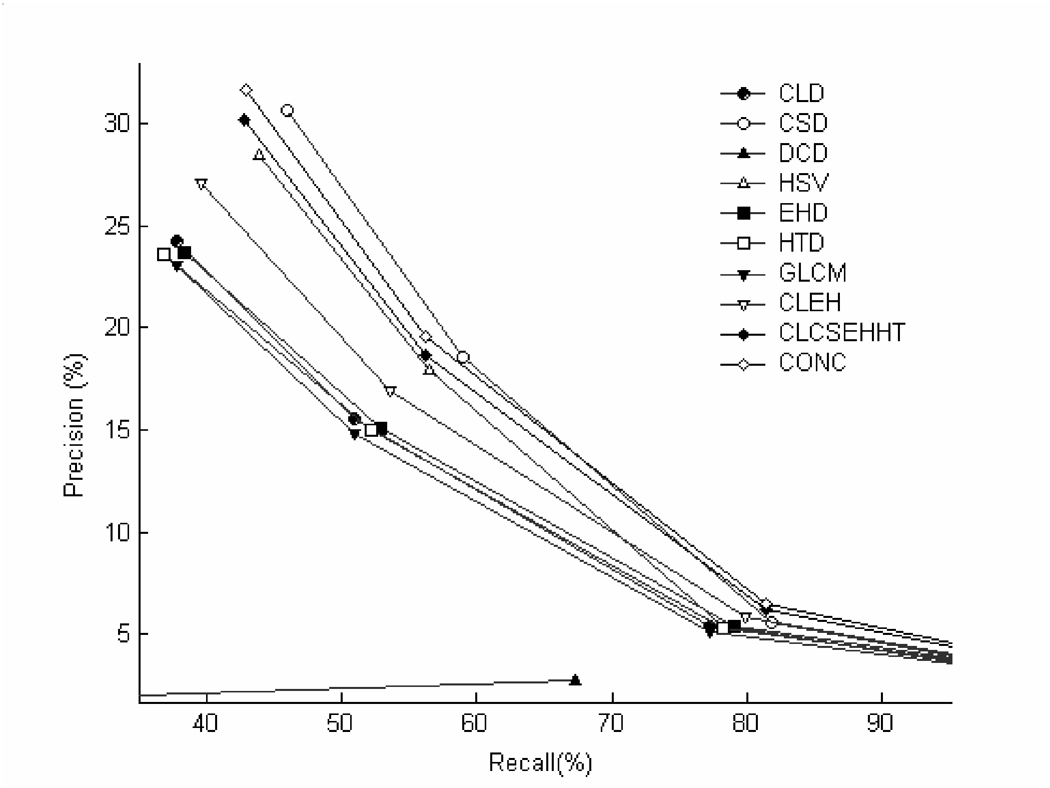


Figure 2.6: Average precision-recall for selected individual descriptors, and three of three feature combinations (VisTex database).

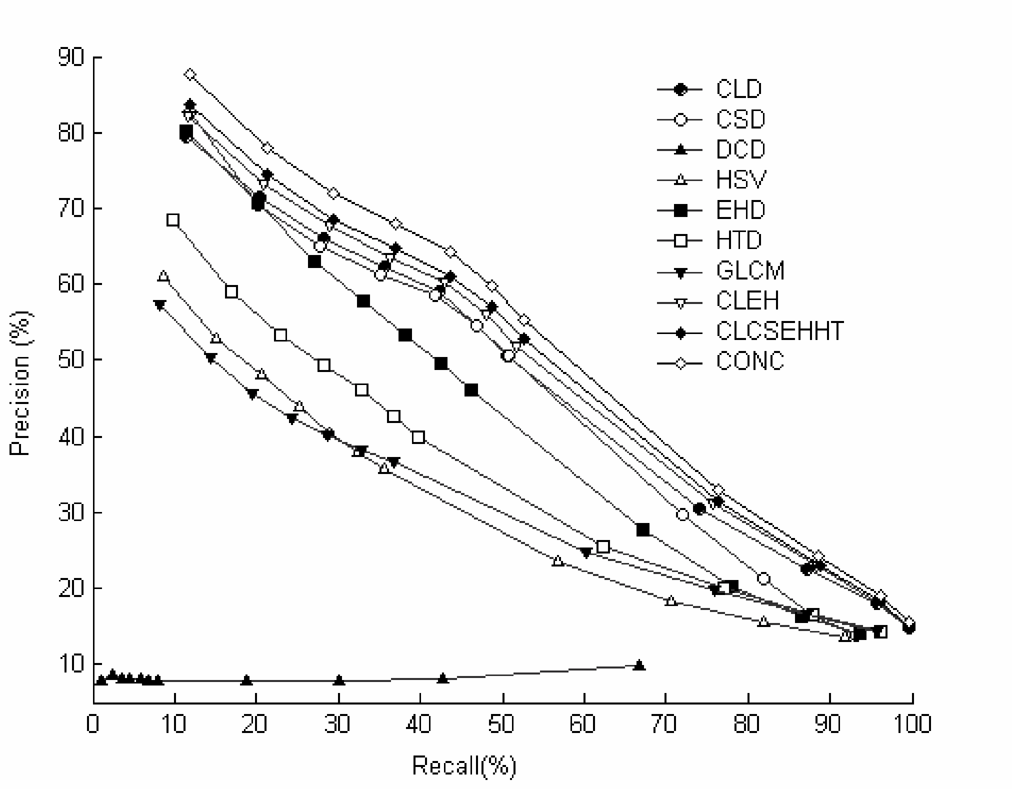


Figure 2.7: Average precision-recall for selected individual descriptors, and three fetaure combinations (D8 database).

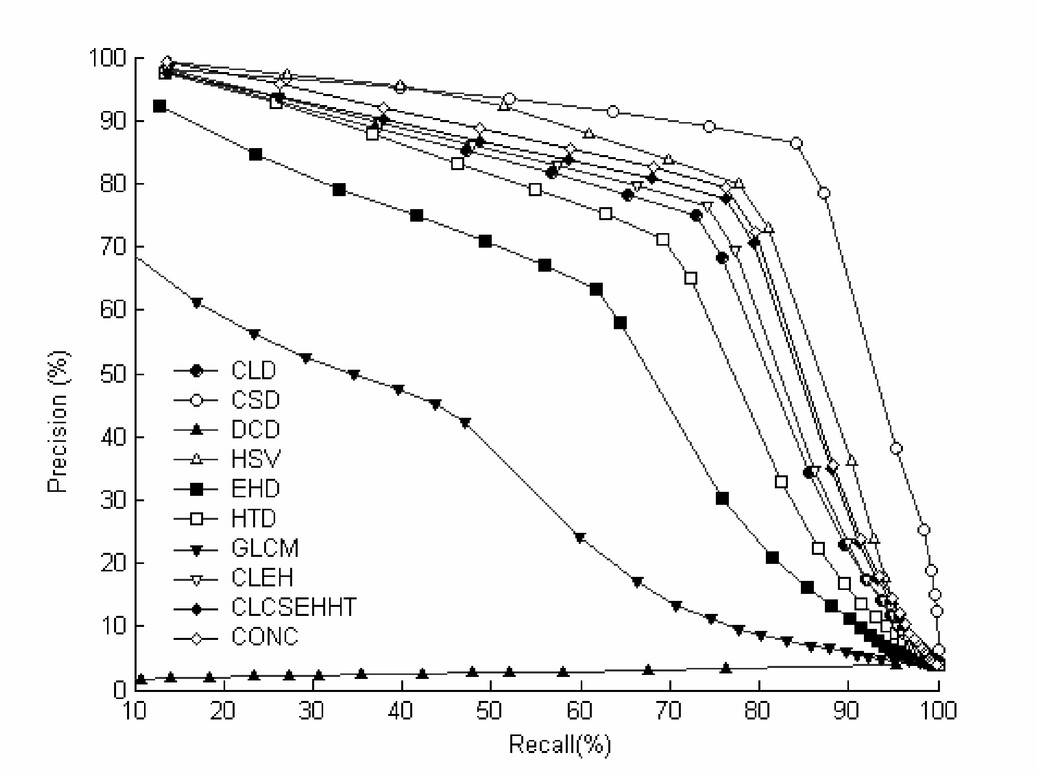


Figure 2.8: Average precision-recall for selected individual descriptors, and three feature combinations (D25-1800 database).

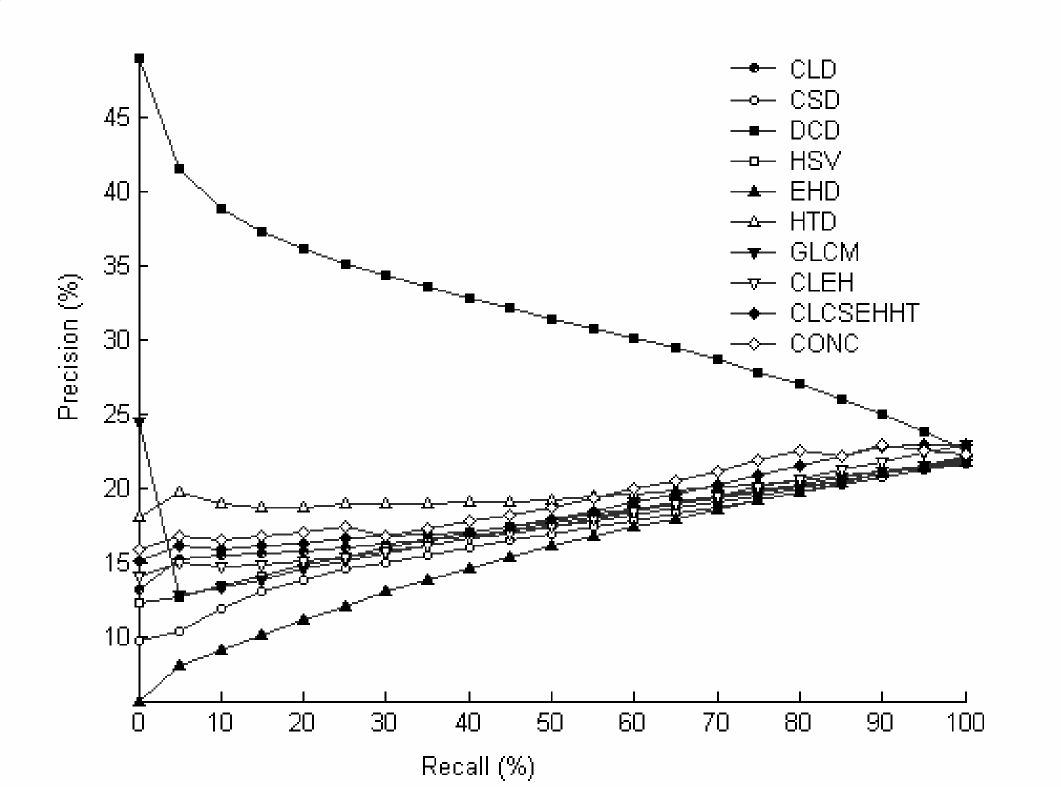


Figure 2.9: Average Precision-recall for selected individual descriptors, and three feature combinations (D7-700 database).

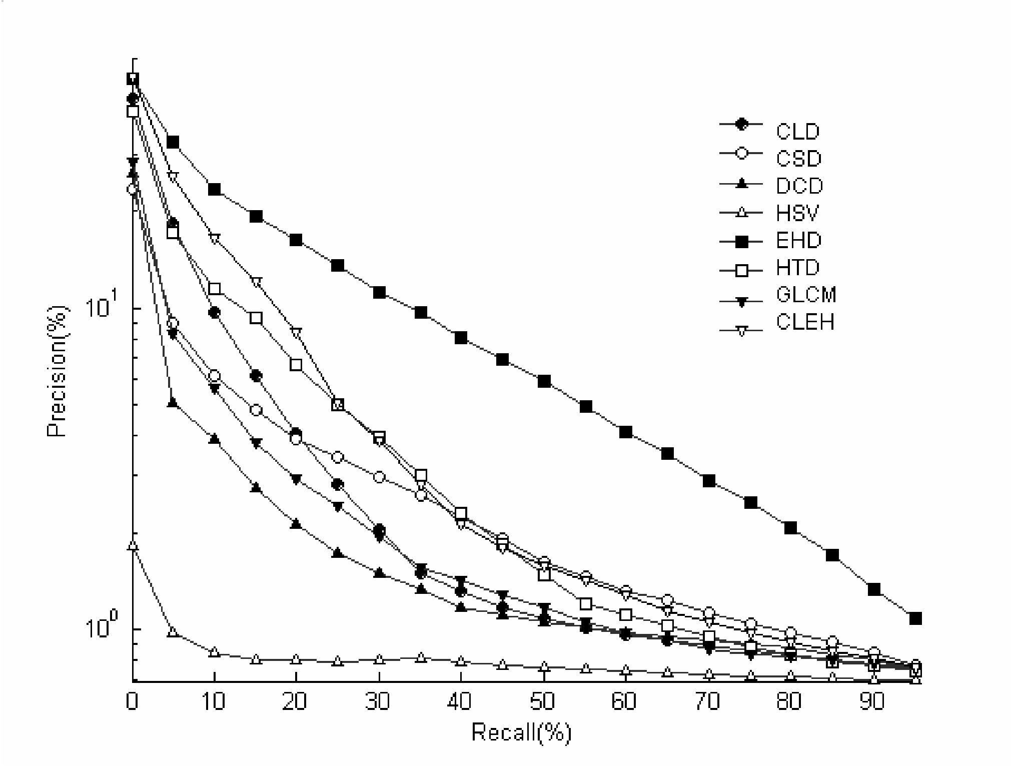


Figure 2.10: Average precision-recall for selected individual descriptors, and their joint feature combinations (Caltech 101 database).

For Figure 2.10, the Caltech 101 database, one individual descriptor EHD outperforms other descriptors and descriptor combination, over a variety of categories. Since there is a big variety in colours within each category (Figure A. 6) , texture descriptors as EHD and HTD outperform colour descriptors as well as the combinations.

As expected depending on the type of the database individual descriptors can perform better than descriptor combinations. This is due to the fact that for some databases the additional descriptors add also noise as well as valuable information.

However, in a generalized case, if an adaptable approach can be found, it would take out of consideration the need to have feature selection in advance. This is the strategy taken in this thesis since specific focus was given to the learning approach and to the interactive adaptation of relevance for individual features. This is achieved by incorporating subjective user defined information into the learning approach.

CHAPTER 3 : Relevance Feedback for Interactive CBIR

In this chapter issues related to relevance feedback for image retrieval are presented. and a review of developments in content based systems with relevance feedback is given. Then the evolution of learning machines used for relevance feedback is presented with a focus on the SVM framework. The learning theory behind this framework is discussed as well as the advantages and disadvantages that make it the learning method of choice used in this thesis. Issues of parameter adaptations in SVMs are addressed, and comparison of several kernels for RF approach considered.

3.1. Introduction

Early approaches for image retrieval were based on keywords, manually annotated images inspired by information retrieval in text documents (Rijsbergen, 1979). Though manual annotations were developed to preserve knowledge they are burdensome, and dependent on subjective interpretations of the professional annotator. In CBIR a search session is usually initialized by either visual example or a small set of previously annotated images related to a given semantic concept. Visual similarity search is than performed on the whole database. The search engine returns images that are similar to the query image based on low-level feature representations extracted from the content. However, low-level similarity relies only on machine’s interpretations of the content, and does not reflect conceptual meaning for the visual content.

Alternately, for specific application scenarios limited and well-defined ontologies can be used (Simou et al., 2005). In a specific domain ontology the aim is to propagate "words" to the whole content database using relationships and rules defined over the domain knowledge. Though very promising this approach places a heavy burden on the designer of high-level relations among concepts.

The most natural way of getting user's subjective information and preferences into the system is by using models that incorporate online learning from the user interactions with the search engine. The idea behind this model is to integrate a "relevance feedback" loop into the system with the user at the centre and the machine learning from user's feedback. This loop is illustrated in Figure 3.1. The relevance feedback concept is based on the analysis of relevant and irrelevant information fed back into the system by the user. This analysis predicts and learns user's preferences in order to iteratively improve retrieval results. Semi-automatic adaptive learning strategies based on relevance feedback are aimed at learning relations between high-level semantic concepts used by humans to identify objects in an image and low-level descriptions of the same visual information.

In every iteration the user provides feedback about the relevance of previously retrieved content. A crucial part of a relevance feedback loop is a learning step which utilizes user preferences by adapting the learning approach in the system, based on provided knowledge accumulated in time. The goal is to minimize necessary interaction between the user and retrieval engine in order to obtain the targeted image or to retrieve as much as possible desired images with a certain concept. A generic CBIR system with RF is presented in Figure 3.1.

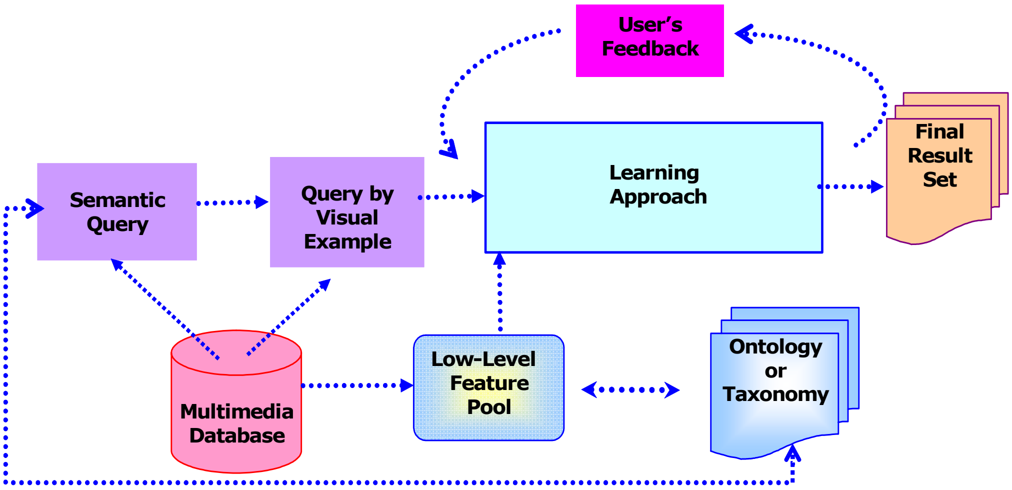


Figure 3.1: Generic architecture of a CBIR engine with RF.

Early works in CBIR dealt with subjectivity of visual impressions for different users which does not easily relate to the low-level feature space (Kurita and Kato, 1993; Picard et al., 1996). Hence this led to a number of general-purpose image search engines that have been developed in the last decade.

One of the first available CBIR systems was Query by Image Content (QBIC) initially based on colour retrieval (Flickner et al., 1995). QBIC was later extended to enable retrieval based on images, sketches and drawings for colour, texture, and motion features.

Next, the Virage Image Search Engine (Virage) allows querying of general image primitives as colour, shape, texture or domain specific as face recognition. It focuses on query refinement by changing the relative weights of visual features (Gupta and Jain, 1997).

In the Multimedia Analysis and Retrieval System (MARS) an image is represented as weighted multi-feature object (Rui et al., 1997). The relevance given to an image in a relevance feedback scenario is modelled with numerous levels of importance and a probabilistic Boolean retrieval.

The ImageRover system (Sclaroff et al., 1997) was based on k-d trees search and dimensionality reduction with principal component analyses.

The VisualSEEK system (Smith and Chang, 1996) uses a feature back-projection scheme to extract salient image regions. The system enables joint content-based and spatial search. Images with similar arrangement of regions as the queried image are retrieved. A sketch tool enables the user to sketch and position regions on a grid as well as to allocate colours. WebSEEK, is a similar system for web applications (Smith and Chang, 1997), which automatically gathers content from the Internet and saves it into a database with extendible subject taxonomy.

The NeTra retrieval system (Ma and Manjunath, 1997) uses a robust segmentation algorithm and features as colour, texture, shape as well as spatial location information for image regions in a region based-search and retrieval engine.

The Photobook was developed in the MIT Media Laboratory (Pentland et al., 1996) and it consists of a number of tools for image browsing and search. A user feedback region annotation tool is used to infer image labels. Photobook uses both content-based features and text annotations for querying.

The SIMPLIcity (Wang et al., 2001) is an image retrieval system for integrated region matching based on image segmentation. It uses semantics classification methods to reduce the search space by defining categories.

Blobworld (Carson et al., 2002) allows querying based on limited number of regions by merging single-region query results. Joint distribution of colour, texture, and position features is modelled with Gaussian mixture models and parameters are estimated with Expectation Maximization (EM).

Finally, Cortina (Quack et al., 2004) uses combinations of standardized MPEG-7 descriptors and scales over large databases.

However in image retrieval the individuality of various users can not be ignored since different users have different interpretations of visual content. Therefore including the user into the retrieval loop is the only way to identify the target of user's search and capture different interpretations or intended usage for the same multimedia content.

Early learning approaches exploiting user relevance feedback were developed by Sclaroff et al. (1997), Rui et al. (1997), Ishikawa et al. (1998), Nastar et al. (1998), Cox et al. (1998) etc.

3.1.1 Application Scenarios for the RF Problem

Visual impressions of content are not only subjective to a particular user but also differ based on prior knowledge, as well as current circumstances of the search and retrieval session. Therefore an automated annotation of images is sensible only in well-defined image databases with clear distinctions between classes and when the application scenario requires such categorization. The application scenario and domain knowledge are crucial for defining different problem variations in relevance feedback approaches. Several directions can be taken starting with very different assumptions.

Cox et al. (1998, 2000) used a *targeted search* scenario when the user has an already envisaged concept and selects semantically similar images to the desired one. In a different scenario search by *association or browsing* was considered. This is a non-professional user scenario when the user is not certain what he is searching for. The search is very subjective and difficult to evaluate since the system does not need to retrieve all relevant images just some based on users preferences. Alternately, in a *category search* scenari*o* the user has a semantic concept in mind that corresponds to a class of similar images. In this scenario a user is searching for most of the images from the relevant class. The measure of effectives is the ability to retrieve all relevant images to a concept before the irrelevant ones , in as less iterations as possible. This is the scenario used for evaluation in this thesis.

The targeted search correspond to a “greedy user model” defined by Zhou and Huang (2003). The user is impatient and wants to view the most relevant images to the concept even though they might not be the most informative for the RF learning approach. They also defined a “cooperative user model”, where the user is patient enough to provide more feedback iterations and view examples which are most informative for the system., with highest levels of uncertainty. Similarly to this model Cox et al. (2000) searched for sets of images that would minimize the expected number of future iterations. In an alternative approach (Tong and Koller, 2000; Tong and Chang, 2001) proposed active learning with SVM for text classification and image retrieval, as means for guiding the learning approach to minimize the number of iterations.

What is relevant in RF approaches is to rank relevant images before irrelevant ones, hence the method behind the approach does not need to be a classifiers with clear dictions among classes, but a ranking module with probabilistic levels of importance. However, RF approaches are often used as a part of a larger framework to facilitate automatic annotation by propagating textual annotations based on low-level similarity of user provided feedback (Lu et al., 2000; Dorado et al., 2006; Djordjevic et al., 2005). Hence in this case the problem is classification, with a crisp decision about membership to a particular class. Accuracy of classification and rate of reducing the error through relevance feedback iterations, are measures of effectiveness.

Kurita and Kato (1993) described the problem of having variable relevance feedback information in time even for the same user. Hence in its basic form relevance feedback is a short term refinement strategy, as opposed to long term feedback based on many users that can be considered to accumulate knowledge (Bartolini et al., 2001; Fournier and Cord, 2002). Other key issues in RF approaches are:

- How to appropriately define and select a feature space in order to present not only visual image content but also the interpretational characteristics of the human visual system. Some examples of possible RF systems deal with features based on: whole images, regions, tiles, patches interest points.

- How to choose the similarity measure between representative feature vectors to reflect high-level semantic similarity between media items observed by humans.
- How to exploit various user feedback information: relevant samples, irrelevant samples, fuzzy levels of relevance, relative judgement etc.
- How to enable real time interaction between the user and the machine while dealing with huge databases.

3.1.2 Overview of Learning Approaches for Relevance Feedback

Various relevance feedback algorithms have been proposed in the last 10 years as an integral part of content-based image retrieval systems. The learning process can relays on only positive examples (Sclaroff et al., 1997; Ishikawa et al. 1998; Heesch and Ruger 2003; Jing et al. 2004a, 2004b), both positive and negative examples (Nastar et al., 1998; Porkaew et al., 1999), multiple levels of relevance (Rui et al. 1998), or fuzzy membership functions (Yap and Wu, 2003).

A wide class of approaches is based on learning object structure via training on segmented regions.

For Instance, Ratan et al. (1999) identified an image as a set (a bag) of regions with every region represented by its own low-level features. Entire bags are labelled as relevant or irrelevant and the learning approach for RF tries to learn important regions for the current search session. The approach searches for areas in the feature space that are close to all positive and far from all negative bags of regions.

Forsyth and Fleck (1997), defined “body plans” for objects (e.g. horses) based on colour, texture, shape information and geometrical relations among parts of a scene.

Vasconcelos and Lippman (2000) used Bayesian inference on block based image local features for relevance feedback learning.

Hong and Huang (2001) used attribute relational graphs for context of objects and scene. EM was exploited to learn parameters of a probabilistic model based on multiple graph examples.

The following section gives a review and analysis on several investigated techniques and approaches for relevance feedback. In general based on the learning method the relevance feedback models can be classified into:

- **Descriptive learning models**. The relevant class is modelled with a parametric or non-parametric model, based on training samples in the feature space (e.g. Gaussians, Gaussian Mixture Models, Parzan windows).
- **Discriminative learning models.** These models do not describe classes but the boundaries separating the classes (e.g. SVMs, Biased Discriminative Analyses). This class of models also includes **neural networks**, since they implicitly infer ranking and classification.

Several approaches representing examples of the above mentioned classification of supervised learning techniques have been described in the following sub-section in an effort to give an insight into the development of this area (Djodjevic and Izquierdo, 2005).

Descriptive Learning Models

As described in Huang and Zhou (2001) early image relevance feedback methods were based on heuristic techniques with empiric parameter adaptation. The initial idea was based on independent weighting of various features in the joint feature space. The reasoning behind this idea is that the feature providing the most compact clustering of relevant images and separation of relevant and irrelevant image was weighted the most.

In many systems, a relevance feedback module is a part of query by example (QbE) paradigm. One of the first CBIR systems using relevance feedback is MARS (Rui et al. 1998). It introduced re-weighting techniques and query point movement (QPM), for exploitation of user's feedback. The re-weighting scheme assumed independency of all features, hence positive samples were modelled with a single Gausssian distribution. The new query vector was a mean of all positive examples in the feature space, weighted with inverse variance. The alternative approach based on QPM, shifted the new query feature vector towards positive examples and away from negative examples. This approach is based on a modification of the classical Rocchio's method used in text based document retrieval. A new query vector for each feature $\mathbf{q}_i'$ is formulated out of the old query vector $\mathbf{q}_i$ and relevance feedback information from the user:

$$\mathbf{q}_i' = \alpha \cdot \mathbf{q}_i + \beta \left(\frac{1}{N_R} \sum_{j \in D_R} \mathbf{x}_{ji} \right) - \gamma \left(\frac{1}{N_N} \sum_{j \in D_N} \mathbf{x}_{ji} \right), i = 1, \ldots, m$$

Where m is the number of features, α,β,γ are constants, N_R and N_N are numbers of images in the relevant D_R and irrelevant set D_N, respectively. And $\mathbf{x}_{ji}$ is j^{th} training sample in the appropriate set.

Even some of the latest approaches (Jing et al., 2004a, 2004b) considered information retrieval techniques for RF. The new query vector is combined with emphasized importance on the latest labelled positive examples, and decreased importance of prior positive examples, again based on a modification of Rocchio's method.

Heesch and Ruger (2003) combined QPM and weight update for their RF approach. The new query is rendered based on user's feedback. The visual interface plots thumbnails of retrieved images with respective distance from the centre of the screen proportional to the dissimilarity of each image to the query. By moving the images on the screen the user provides RF to the system as a real-valued vector of distances. These distances differ from values provided by the system using low–level features. Minimizing the sum of square errors for these two different distances leads to a solution of the weight updates for the next iteration.

Newsam et al. (2001) described a search engine based on combination of keywords and low-level features. Content-based retrieval is performed in semantic categories using multiple image features. The updated query vector is obtained through QPM.

In Cox et al. (1998), the PicHunter system uses relevance feedback in a form of relative judgment as opposed to the stronger classification into relevant/irrelevant. An extension of k-d trees to stochastic settings leads to the proposed "stochastic-comparison search".

Porkaew and Chakrabarti (1999) introduced a query extension scheme based on multipoint query opposed to QPM approach. The relevant samples are clustered and nearest-neighbour decision approach is used to define membership to a cluster, for a multipoint query.

Ishikawa et al. (1998) generated a new optimal query approach by minimizing the total distances of positive examples from the new query, based on Mahalanobis distance. In this case parameter estimation lead to a solution for the new query vector that resulted in a weighted average of all positive images. In principle this approach is based on singular Gaussian distribution assuming that all positive examples are clustered together. However it is the first non heuristic technique solving an optimisation problem and considering correlation among features.

While early probabilistic RF systems assumed that positive images follow a single Gaussian distribution, on the other hand later systems were based on more complex distributions such as Gausssian mixture models (GMMs). Qian et al. (2002) used Gaussian Mixture Model (GMM) to represent the distribution of relevant images using both relevant and irrelevant examples as well as unlabelled data. Estimation of parameters for the mixture of Gaussians is based on positive relevance feedback and the assumption that positive examples may be grouped in the feature space and mutually separated by negative examples.

In a similar manner, Schettini et al. (1999) considered negative examples and updated weights for each separate feature, by comparing the variance of positive examples to the variance of the overall positive and negative examples.

In the effort to avoid assumptions for shape of a density model, Meilhac and Nastar (1999) used a non-parametric distribution for targeted samples based on Parzen window density. The problem is defined as a difference of densities for relevant and irrelevant samples. Every labelled image is the centre of a Gaussian with a high probability that the neighbouring elements are similar. The approach is incremental and user's feedback is used to more precisely estimate the model. i.e. the Gaussian smoothing function become narrower with increasing number of labelled images.

Discriminative Classification Models

A straightforward way for dealing with non-linearity of visual feature space was introduced by discriminative approaches and kernel based algorithms. As mentioned before, discriminative classification models do not primarily concentrate on estimating the correct distribution of relevant and irrelevant data but rather on estimating the boundaries between classes. Kernel based approaches were used with non-linear distributions based on extension of linear discriminative analysis in Zhou and Huang (2001a, 2001b). Neural networks implicitly enable classification and ranking and thus can be effectively used for relevance feedback (Laaksonen et al., 1999; Koskela et al., 2004; Yap and Wu, 2003). Statistical methods based on kernel SVMs and their good generalization capabilities as well as active learning were considered for RF by Hong et al. (2000), Tian et al. (2000a), Jing et al. (2004a, b), Tong and Chang (2001).

One of the directions modern relevance feedback techniques are heading considers solving a two class (relevant/irrelevant) problem by use of Discriminative Analysis (DA). Conventional linear discriminative analysis (LDA) approach tries to cluster all irrelevant images into one class, whereas it is known that negative examples belong to many different classes. Muller et al. (2001) tried to find linear projections such that different classes are well separated. Separability is measured by how far apart are the projected means of two classes and how big is the variance of data in the projected direction. Therefore, in the following sub-section variations of DA approaches and their application in relevance feedback systems, are depicted.

Wu et al. (2000) considered image retrieval to be a transductive learning problem, using both labelled and unlabelled data samples. A linear transformation based on the labelled data set that could be generalize onto the unlabelled dataset was found through multiple discriminative analyses (MDA). The transformation maximizes the ratio of inter-class vs. intra-class scattering, with all negative examples scattered among each other in the feature space. MDA is a generalization of LDA for multiple classes, a supervised statistical method that requires large number of labelled samples. By combining MDA with EM into discriminate expectations maximization algorithm (D-EM), the MDA approach is provided with enough labelled data. D-EM initially starts with a weak Bayesian classifier, hence the unlabeled dataset is provided with probabilistic labels and the new formed training set projected to a different feature space with MDA. The main disadvantage of MDA is the assumptions that both positive and negative examples cluster into distinctive classes, that is, each negative example is treated as from a different class. However several negative examples can come from the same class and splitting negative examples into different classes can mislead the resulting discriminating subspace.

Hence, Zhou and Huang (2001a, 2001b) suggested a variation of the above approach in a form of Biased Discriminative Analysis (BDA). They assumed that all the positive examples are clustered in one class while negative examples negative examples can come from an uncertain number of classes. In this case a biased classification problem is defined with multiple classes but with the user being interested only in one, the positive class. The goal is to determine a function, mapping, that would enable maximum clustering of positive examples moving them away from the negative examples. In other words maximizing scattering of negative examples to a positive centroid and minimizing within class scattering of positive examples. As a modification of the method above, an extension to non-linear space is proposed based on kernel methods. Kernel BDA is formulated through expressing the BDA problem using inner product in the transformed space. Zhou and Hung (2001*a*) showed that kernel BDA produces better results than both LDA and SVMs. However, kernel BDA is sensitive to imbalance between positive and negative samples and needs a considerable amount of training to perform well. Nakazato et al. (2003) further broadened the approach by combining a group oriented user interface with BDA. Group Biased Discriminative analyses deals with image retrieval using classification of multiple positive and multiple negative classes. Again this is done by maximizing scattering for negative classes and clustering for positive classes as well as moving away negative clusters from positive ones.

Indisputably there are number techniques integrating neural network learning approaches and relevance feedback systems.

Laaksonen et al. (1999) developed the PicSOM CBIR system with tree-structured self-organizing maps (TS-SOM). This is an unsupervised topologically ordered neural network for image indexing along low-level features. High-dimensional input feature space for positive and negative examples is mapped to a low-dimensional lattice and its neighbouring neurons. After low-pass Gaussian filtering appropriate labels are spread to corresponding neighbourhoods under assumption that a good mapping will keep positive examples clustered while negative examples will scatter. This approach considers QbE and iterative refinement using relevance feedback (Koskela et al., 2004). The PicSOM system supports multiple features by employing parallel self-organizing map for each feature. The disadvantage of this approach is the high computational requirement for obtaining subsets of relevant images for each feature and the assumption of feature independence.

Yap and Wu (2003) presented a fuzzy RF approach. A continuous fuzzy membership function models user's fuzzy feedback by weighting differently different images based on subjective levels of relevance. Radial bases function (RBF) neural network is used as a learning approach. Three hidden layers of the RBF neural network correspond to relevant, irrelevant or fuzzy labels. The output layer combines responses from the modules as a liner combination of the three sub-networks. Fuzzy weights are determined using a fuzzy membership function and the closeness of a sample to the centre of a relevant cluster.

In the last couple of years there has been a huge impact of SVMs (Vapnik, 1995; Schokopf and Smola, 2002) on classification, recognition and retrieval problems, specially RF. The input data is mapped into higher-dimensional feature space using a non-linear transform, kernel. In the newly defined feature space linear discrimination boundary between classes can be easily found. A subset of feature vector, which are closest to the decision surface and therefore define it, are called support vectors (SVs).

Hong et al. (2000) and Tian et al. (2000a) used SVMs to estimating the weight of relevant images in the covariance matrix of Mahalanobis distance, used as a similarity measure. This approach is a combination of already exploited techniques and the statistical learning algorithm for SVM. Different weights are assigned based on the distance of positive examples from the SVM based hyperplane, the larger the distance the more distinguishable the relevant examples from the negative ones and the larger the weights.

Jing et al. (2004a, b) used multiple regions to generate image signatures of different dimensions and match them with Earth Mover's distance (see Chapter 2). A new kernel for SVMs based on EMD is introduced to better accommodate the region-based approach. It incorporates features of all the regions and allows many-to-many relationship among the regions. Thus enabling higher robustness to inaccurate segmentations. However there is no theoretical explanation and proof that the introduced kernel leads to an optimal solution of the SVMs problem (see Chapter 4).

Chen et al. (2001) used in their RF approach a statistical learning machines called one-class SVM. A kernel based one-class SVM acts as a density estimator for positive examples. Though the approach shows promising results the negative examples when they exist, are a source of information completely ignored in this approach.

Similarly, Guo et al. (2002) developed a constrained similarity measure for image retrieval, based on SVMs. This measure learns the enclosing boundary for similar images in a non Euclidean space. The relevant images inside the boundary are ranked based on the Euclidean distances from it and images outside of the boundary are treated with less relevance.

Learning approaches play an important role in RF algorithms, however they have two main drawbacks: scarce training data, usually there is much less training data available than required by dimensionality of the feature space; and imbalance in size of different classes (Chang et al., 2003; Zhou and Huang, 2003). The choice of using SVMs for the learner in this thesis was made on the following arguments:

- SVMs show good generalization and learning properties even based on limited number of training examples.
- There is no need for prior assumption for the shape of the distribution of the relevant class, even though some assumptions have to be made when the non-linear kernels are used for mapping.
- SVMs enable ways of adapting the kernel to the data by integration of prior knowledge, unlabelled samples and active learning.
- Though this is mainly a hard binary classification approach, appropriate ranking, necessary for RF methods, can be inferred based on the closeness of samples to the decision surface.

3.2. The Learning Algorithm, Support Vector Machines

Relevance feedback systems deal with an explicit need for generalizing the behaviour of labelled training set data over the unseen and unlabelled testing data. Therefore effectiveness of a relevance feedback system does not only depend on the features and the metric used, but also on the supervised learning approach chosen to learn behaviour of different classes. A generalized, nonlinear and high-dimensional feature space is considered in this thesis, while offering additional multi-feature information it also makes the problem more challenging.

3.2.1 The Learning Theory

The beginnings of learning theory date from 1960's when Rosenblatt introduced a new kind of learning machines, perceptrons or neural networks. They were designed as connected neurons where each neuron defines a different separating hyperplane. The overall learning approach defines a piecewise linear separating surface (Cortes and Vapnik, 1995). The input space is a N-dimensional feature space X, with m training patterns $\mathbf{x}_i$ and their binary labels y_i. Since this is a supervised approach the input can be denoted as follows:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_m, y_m) \in \mathbb{R}^N \times \{-1, +1\} \tag{3.1}$$

The neurons divide the space X into two regions, the part where the output y_i takes value -1 and the part where it takes value 1. The output is generate based on the following functional dependence to an input pattern x_i:

$$y_i = sign\{(\mathbf{w} \cdot \mathbf{x}_i) - b\}, i = 1, ..., m$$

Where vector $\mathbf{w}$ and bias b define the position of the separating hyperplane $(\mathbf{w} \cdot \mathbf{x}_i) - b = 0$. In 1962 Novikoff (cited Vapnik, 1995) introduced the first theorem for perceptrons making the connection between, the generalization capabilities required by learning approaches, with the principle of minimizing the number of training errors.

3.2.2 The Learning Problem

The goal of a learning task is to generalize the dependency established on limited number of training samples onto a larger testing set coming from the same underlying distribution as the training set. The input is a binary labelled supervised training dataset, as in (3.1). The learning approach aims at finding a set of functions $f(\mathbf{x}, \alpha)$, that depend on the input samples $\mathbf{x}$ and a set of parameters Λ, where $\alpha \in \Lambda$. The goal of a learning machine is to choose a function that best models the supervisors input. A way to solve this task is to measure the error between a response given by the supervisor and an approximation returned by the machine. This error can be represented by a loss function $L(y, f(\mathbf{x}, \alpha))$. Minimizing the expected value of the loss function leads to the risk function:

$$R(\alpha) = \int L(y, f(\mathbf{x}, \alpha)) dP(\mathbf{x}, y) \tag{3.2}$$

Specific formulations of such learning problems depend on the form of the loss function and are frequently used in pattern recognition, regression estimation and density estimation learning problems (Vapnik, 1999). In the case of binary classification the most common loss function is the so-called 0/1 function:

$$L(y, f(\mathbf{x}, \alpha)) = \frac{1}{2} \cdot \begin{cases} 0, \text{ if } (y \cdot f(\mathbf{x}, \alpha)) < 0 \\ 1, \qquad \text{otherwise} \end{cases} \tag{3.3}$$

The actual problem consists in minimizing the risk function using empirical training data. Previous considerations provide the basis for empirical risk minimization (ERM) principle. In ERM the expected risk function from (3.2) is replaced with the empirical risk based on m training samples:

$$R_{emp}(\alpha)=\frac{1}{m}\sum_{i=1}^{m}Q(\mathbf{x},y,\alpha), \tag{3.4}$$

where $Q(\mathbf{x},y,\alpha)$ denotes a specific loss function.

Vapnik (1995) defined the *theory of consistency* of a learning process which provides necessary and sufficient conditions for convergence of the ERM principle. These lead to a uniform convergence of the empirical risk $R_{emp}(\alpha)$ towards the expected risk $R(\alpha)$:

$$\lim_{m->\infty} Probability\left\{\sup_{\alpha\in\Lambda}\left|R(\alpha)-R_{emp}(\alpha)\right|>\varepsilon\right\}=0,\quad \forall\varepsilon \tag{3.5}$$

Vapnik (1998) also defined the entropy on a set of indicator functions. In a classification approach a set of indicator functions $Q(\mathbf{x},y,\alpha)$, takes only the values 0 or 1. For different values of $\alpha\in\Lambda$ the set of indicator functions $Q(\mathbf{x},y,\alpha)$ can generate N^{Λ} different separations of m independent and identically distributed data samples, producing binary loss vectors:

$$(Q(\mathbf{x}_1,y_1,\alpha),...,Q(\mathbf{x}_m,y_m,\alpha)) \tag{3.6}$$

The necessary and sufficient conditions for consistency of the ERM principle, given in (3.5), are fulfilled for:

$$\lim_{m\to\infty}\frac{H^{\Lambda}(m)}{m}=0, \forall\varepsilon>0\,,$$

where $H^{\Lambda}(m)=\mathrm{E}\ln N^{\Lambda}$ defines random entropy or diversity on the set of indicator functions. Tighter non-asymptotic bounds on uniform convergence given in (3.5), can be expressed through annealed VC-entropy denoted with $H_{ann}^{\Lambda}(m)$ and growth function $G^{\Lambda}(m)$. Both these values lead to quality estimation of the ERM principle, and they are connected with random entropy $H^{\Lambda}(m)$ through the following inequality (Vapnik, 1998):

$$H^{\Lambda}(m)\leq H_{ann}^{\Lambda}(m)\leq G^{\Lambda}(m)\,. \tag{3.7}$$

The annealed VC-entropy $H_{ann}^{\Lambda}(m)$ is used to formulate a fast rate convergence of ERM principle. Whilst the growth function $G^{\Lambda}(m)$ enables both consistency of ERM principle and fast asymptotic rate of convergence independently on the underlying probability.

The *VC-dimension* on a set of indicator functions $Q(\mathbf{x}, y, \alpha), \alpha \in \Lambda$ is described as a capacity of a set of functions. It is the maximum cardinality h of the input set which can be separated into all possible 2^h ways, by using the binary loss vectors from (3.6). And at the same time there is no set of higher cardinality satisfying the above mentioned property. The tightest bound can be defined through the growth function, which is either linear $G^{\Lambda}(m) = m \ln 2$ with infinity VC-dimension or it is bounded by a logarithmic function:

$$G^{\Lambda}(m) < h\left(\ln\frac{m}{h} + 1\right) \tag{3.8}$$

where h is a finite VC-dimension and $h < m$.

Based on (3.7) and (3.8) the VC dimension h provides a constructive upper bound on the growth function, and enables asymptotic high rate of convergence independently of the problem.

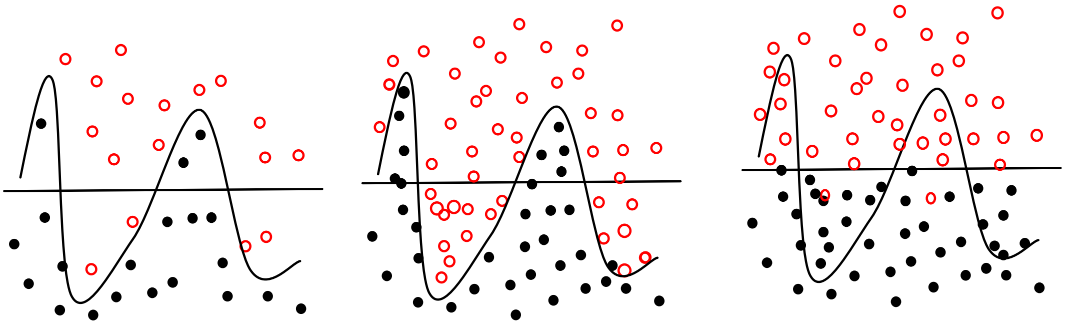


Figure 3.2: Small number of training samples (left), the under-fitted problem (middle) and the over-fitted problem (right).

For large training sets, minimizing the empirical training risk would lead to good minimization of the expected risk (3.5), however for small training sets the problem of over-fitting and under-fitting appears. If the number of samples is small both linear and nonlinear separation bounds are correct, though the linear approach is simpler with a slightly larger error, Figure 3.2 (left). However, when more training data is available there is considerable difference in error and the correct plane can be easily identified. In

Figure 3.2 (middle) the nonlinear bound is correct while the linear bound under-fits the data. In case the linear bound is correct as in Figure 3.2 (right) the nonlinear bound over-fits the data. With a small training set, minimizing the empirical risk does not guarantee a small expected risk and good generalization properties, since the case of over-fitting can often occur.

Vapnik-Chervonenkis theory (Vapnik, 1995) restricts the complexity of a set of function $f(\mathbf{x},\alpha)$, implemented by a learning machine, to a class of functions with a capacity corresponding to the amount of training data. For a loss function defined with $\frac{1}{2}\left|f\left(\mathbf{x}_i,\alpha\right)-y_i\right|$ and some η, $0\leq\eta\leq 1$ the following holds true with probability of at least $1-\eta$, and for VC dimension $h<m$:

$$R(\alpha)\leq R_{emp}(\alpha)+\sqrt{\frac{h\left(\log\frac{2m}{h}+1\right)-\log\left(\eta/4\right)}{m}} \tag{3.9}$$

The ERM principle is intended for large training sets where m/h is large and the second term in (3.9), *the VC confidence*, is small thus the expected risk is close to the empirical risk. In this case small value of the empirical error leads to small expected error as defined with (3.5). However when there is not enough training data the value for m/h is small, in this case because the VC confidence might be large and even if the empirical risk is small there is no guarantee that the expected risk $R(\alpha)$ will be small. In that case the ERM principle does not work and a new approach called Structural Risk Minimization (SRM) principle is used. The new principle simultaneously minimize both the empirical risk and the VC confidence form (3.9).

The VC confidence term depends on the chosen set of functions $f\left(\mathbf{x},\alpha\right)$ and is a monotonically increasing function of h. However, empirical and expected risk depend on a particular function from the set of functions chosen for the training phase. The goal is to find a subset of the chosen set of functions, so that the upper bound of expected risk is minimized (see Figure 3.3). The entire set of functions is divided into nested subsets and for each subset the exact or bounded value of h is found. SRM finds the subset of functions that minimize the expected risk and the solution are those functions for which the overall sum in (3.9) is minimal.

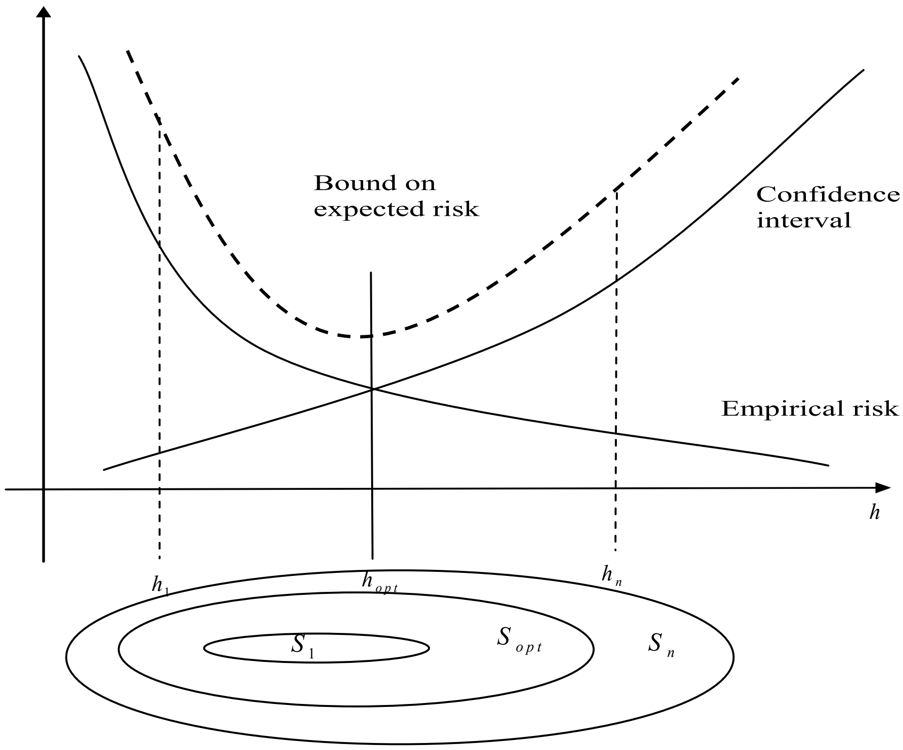


Figure 3.3: The smallest bound of the expected risk is achieved for an optimal subset of the set of functions, with minimized sum of empirical risk and VC confidence bound (Vapnik, 1998).

Generally speaking, separating hyperplanes in an N dimensional feature space have a VC-dimension of $N+1$, as proven by Burges (1998). In a high-dimensional feature spaces the VC confidence, as a monotonically increasing function of h, for small training sets with small ratio m/h, might also have a large value. Hence the expected risk and generalisation properties of the learning machine might lead to large expected risks and bad generalisation over testing sets. However, as it will be explained in the next subsection, *margin* based hyperplanes can have small training sets and still generalize well.

3.3. Support Vector Machines, Optimal Hyperplane Classifier

SVMs are based on SRM principle, i.e. they minimize the upper bound of expected risk. SRM gives better results than ERM principle, which minimizes the error of training data, and is used in classical neural networks. This difference in approaches is what enables SVM's better generalization properties over the unseen data, which is at the end goal of learning methods.

For a binary separable learning problem the set of indicator functions, defining separating hyperplanes can be denoted as :

$$(\mathbf{w}\cdot\mathbf{x}_i)+b=0\,,\ \mathbf{w}\in\mathbb{R}^N, b\in\mathbb{R}\,, i=1,\ldots,m$$

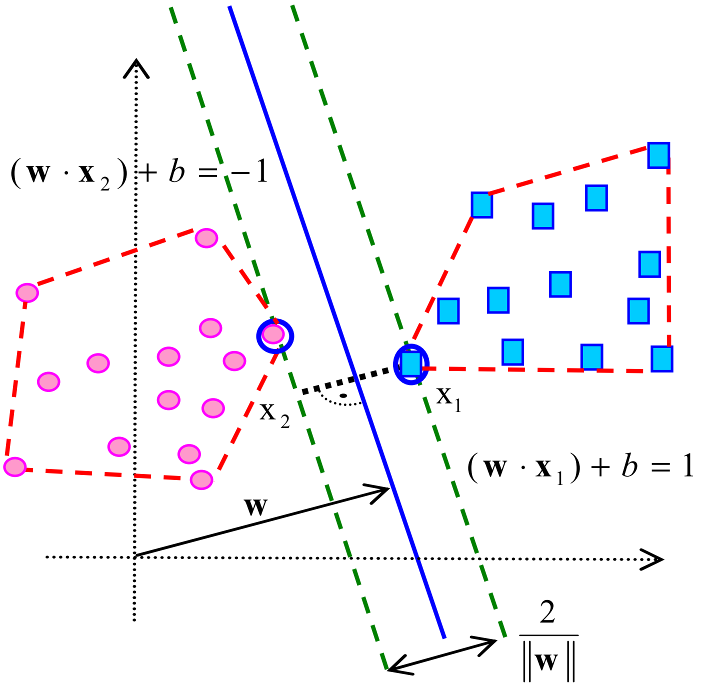
Where vector $\mathbf{w}$ and scalar bias b define the position of the hyperplane. For set of functions denoting hyperplanes, the VC-dimension can be bounded in terms of another quantity, the margin. The margin is the minimal distance of samples for different classes from the decision surface. In case the training dataset is separable, the weight vector $\mathbf{w}$ and bias b are rescaled in such a way that the closest points to the hyperplane satisfy the following condition:

$$\left|(\mathbf{w}\cdot\mathbf{x}_i)+b\right|=1. \tag{3.10}$$

In this case a canonical hyperplane representation is obtained in the following form:

$$y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)\geq 1,\quad i=1,..,m.$$

The hyperplane is optimal if it separates a set of vectors without error and maximizes the distance between vectors from different classes that are closest to it.



$(\mathbf{w}\cdot\mathbf{x}_2)+b=-1$
x_2
x_1
$(\mathbf{w}\cdot\mathbf{x}_1)+b=1$
$\mathbf{w}$
$\frac{2}{\|\mathbf{w}\|}$

Figure 3.4: Separating margin, in a binary classification problem.

From Figure 3.4 it can be seen that the optimal hyperplane is orthogonal and at half-way of the shortest line connecting two convex hulls, where each convex full encloses points from different classes. If two samples $\mathbf{x}_1$ and $\mathbf{x}_2$ from different classes are considered, and projected onto the hyperplane, the margin will be the distance of these two points perpendicular to the hyperplane:

$$\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x}_1\text{-}\mathbf{x}_2) = \frac{2}{\|\mathbf{w}\|}. \tag{3.11}$$

Hence the distance of a sample point to the margin is:

$$d(\mathbf{w}, b, \mathbf{x}_i) = \frac{|\mathbf{w} \cdot \mathbf{x}_i + \mathrm{b}|}{\|\mathbf{w}\|}. \tag{3.12}$$

The VC-dimension of a class of separating hyperplane is linked to the margin, hence to the length of vector $\mathbf{w}$, introducing a possibility to control it for SRM principle (Vaprink, 1995):

Let R be the sphere of the smallest radius, enclosing all training points, $R = \{\mathbf{x}_i \in X, \|\mathbf{x}_i - \mathbf{a}\| < R\}$, and let $\mathbf{a} \in X$ be the centre of the sphere. If $f(\mathbf{w},\mathrm{b},\mathbf{x}_i) = \mathrm{sgn}((\mathbf{w} \cdot \mathbf{x}_i) + \mathrm{b})$ is the hyperplane decision function on training points, then the set of hyperplanes $f(\mathbf{w},\mathrm{b},\mathbf{x}_i)$ has the following VC-dimension:

$$h \leq R^2 A^2 + 1 \text{ , for } \|\mathbf{w}\| \leq A \tag{3.13}$$

Hence, from (3.11) and (3.13) the margin, of a set of hyperplane functions, is bounded from below $\frac{1}{\|\mathbf{w}\|} \geq \frac{1}{A}$, which enables control of the VC-dimension h. If we have a small margin, then a much larger class of problems can be separated. Only a hyperplane that is farther from any sample than $1/A$ can be a potential optimal hyperplane. Therefore, the number of possible planes and the capacity of a class of hyperplanes decreases as the margin increases (see Figure 3.5).

The condition (3.13) on a set of hyperplanes defines a set of functions with VC-dimension $N+1$, for an N dimensional input feature space X. Since the margin is bounded from below, the VC-dimension can be much smaller than N hence allowing high dimensional feature space. Based on (3.9) the expected risk does not depend on the dimensionality of the feature space but on the VC-dimension.

Hence, maximizing the margin is equivalent to minimizing the upper bound on the VC-dimension, and here lies the connection between SRM principle and the optimizing approach in SVMs.

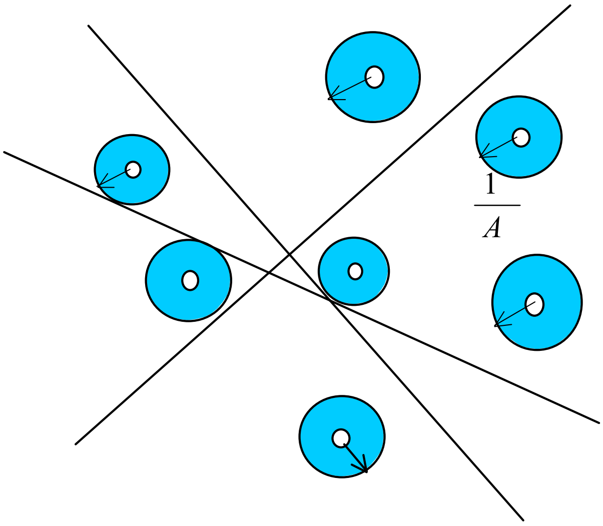


Figure 3.5: Constraints on the number of possible hyper-planes. Capacity of a class of hyperplanes decreases as the margin (blue circle around a sample) increases.

The optimal hyperplane is a solution of a following optimisation problem, obtained by maximizing the margin $2/\|\mathbf{w}\|$:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.14}$$

Subject to the constraint:

$$y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)\geq 1, \quad i=1,..,m. \tag{3.15}$$

The solution of this minimization problem finds the training points for that lie on the borders of the margin, circled points in Figure 3.4. It is difficult to explicitly obtain $\mathbf{w}$, therefore the above minimization problem is solved by introducing Lagrangian multipliers, one for each inequality in (3.15), $\alpha_i \geq 0, i=1,...,m$:

$$L(\mathbf{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{m}\alpha_i(y_i\cdot((\mathbf{w}\cdot\mathbf{x}_i)+b)-1) \tag{3.16}$$

Since the objective function is convex, and points satisfying constraints form a convex set, this is a convex quadratic programming problem (see Appendix B, Definition B.5-B.8). Instead of solving a primal problem (3.14), the dual problem can be equivalently solved by maximizing the Lagrangian so that the gradient with respect to primal variables $\mathbf{w},b$ vanishes subject to constraints on dual variable α_i:

$$\max_{\alpha\geq 0}(\min_{\mathbf{w},b} L(\mathbf{w},b,\boldsymbol{\alpha})) \tag{3.17}$$

Since the problem at hand is a constrained optimization problem, the Karush-Kuhn-Tucker (KKT) condition play a vital role (see Appendix B, Definition B.9). The problem of SVMs is convex, and for convex problems the KKT conditions are necessary and sufficient for $\mathbf{w}, b, \boldsymbol{\alpha}$ to be a solution:

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3.18}$$

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \tag{3.19}$$

$$y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) - 1 \geq 0, \quad i = 1,..,m \tag{3.20}$$

$$\alpha_i \geq 0, \forall i \tag{3.21}$$

$$\alpha_i \cdot (y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) - 1) = 0, \forall i \tag{3.22}$$

The dual problem becomes:

$$\max_{\alpha > \mathbf{0}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \,,\; \alpha_i \geq 0, \; i = 1,...,m \,.$$

The Lagrangian multipliers are non zero values only in those saddle points where constraint (3.15) is exactly met, hence the condition in (3.22) is fulfilled. Only these points satisfying $y_i \left|(\mathbf{x}_\mathrm{i} \cdot \mathbf{w}) + b\right| = 1$ have non-zero α_i, they define the margin and are called Support Vectors. The weight vector $\mathbf{w}$ defines the hyperplane and can be explicitly obtained from the training set based on (3.19). Bias b can be determined from (3.22) for any α_i different from zero. Thus, the hyperplane decision function can be written as:

$$f(\mathbf{x}) = \mathrm{sgn}(\sum_{i=1}^{m} \alpha_i y_i \cdot (\mathbf{x} \cdot \mathbf{x}_i) + b) \,.$$

3.3.1 Generalization of the Non-separable Case

If a classification problem is not separable but noisy with large class overlap, then a separable hyperplane cannot be easily constructed. In this case, classical hard margin classification might not lead to the minimum of the expected risk and might lead to over-fitting of data. As a consequence, a slack variable is introduced to relax the hard margin constraint (Cortes and Vapnik, 1995):

$$y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)\geq 1-\xi_i,\ \ \xi_i\geq 0, i=1,..,m \tag{3.23}$$

One implementation of soft margin classifiers C-SVMs, considers minimizing the following function:

$$\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{m}\xi_i \tag{3.24}$$

By minimizing (3.24) the SRM principle is achieved by keeping the upper bound of the VC-dimension small i.e. capacity of the classifier is controlled with $\|\mathbf{w}\|$as in (3.13). At the same time the upper bound on the number of misclassified training samples $\sum_{i=1}^{m}\xi_i$ is minimized. Hence parameter C defines the trade-off between the complexity term and the empirical error. When $C\to\infty$, the problem boil down to the hard-margin SVM problem in which all $\sum_{i=1}^{m}\xi_i$ would be forced to zero.

The dual problem in soft margin classifiers has the following form:

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i\cdot\mathbf{x}_j) \tag{3.25}$$

$$0\leq\alpha_i\leq C,\ i=1,...,m$$

$$\sum_{i=1}^{m}\alpha_i y_i=0$$

The only difference between this dual problem and the same in a separable case, is the upper bound C on Lagrangian multipliers. In this way a potential influence of outliers can be limited. Based on the KKT optimality conditions, three different cases can be distinguished for different values of α_i:

$$(1)\ \alpha_i=0\Rightarrow y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)\geq 1$$

$$(2)\ 0<\alpha_i<C,\ \ \Rightarrow y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)=1$$

$$(3)\ \alpha_i=C,\ y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)\leq 1$$

One of the most important properties of SVMs are sparse solutions in Lagrangian multipliers. Many patterns are outside the margin area and the optimal Lagrangian coefficients are set to zero. KKT conditions show that only patterns on the margin, case (2), or inside the margin area, case (3), are non-zero. In case that Lagrangian values are zeros, the patterns are classified correctly. They do not lie on the margin and therefore they are not SVs. The non-zero Lagrangian values α_i are those satisfying constraint (3.23) with the equality sign. For cases (2) and (3) with non-zero α_i, the input vectors $\mathbf{x}_i$ are called SVs. However in case (3) when the corresponding SVs do not satisfy (3.15), these SVs are considered to be errors.

3.3.2 Non-Linear Support Vector Machines

In case of complex non separable data, an appropriate non-linear mapping of the input space into a higher dimensional feature space, may enable linear separation. In a newly defined non-linear space the separation boundary might by more achievable. The input patterns are mapped into a high dimensional feature space, Hilbert space H (see Appendix B, Definition B.10, B.11 and Example B.12):

$$\Phi : \mathbb{R}^N \to H \text{ , } \mathbf{x}_i \to \Phi(\mathbf{x}_i)$$

The same algorithm as in the linear case is consider with the input data now represented as:

$$(\Phi(\mathbf{x}_1), y_1), (\Phi(\mathbf{x}_2), y_2), ..., (\Phi(\mathbf{x}_m), y_m) \in H \times \{-1, +1\}$$

The curse of dimensionality implies that as the dimension of the space N increases so does the difficulty of the estimation problem. In general the required number of samples increases as an exponential function of N. However, learning a simpler class of functions in a higher dimensional feature space is the key, since the complexity of learned functions has more influence than the dimensionality of the space.

In the SVM algorithm (3.25) the data appears only in a form of inner product $\left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle = \mathbf{x}_i \cdot \mathbf{x}_j$, where $\langle \cdot, \cdot \rangle$ is a inner product operator in X. The inner product in the input space is replaced with inner product in Hilbert space:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \right\rangle = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \tag{3.26}$$

$K(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel, a generalized non-linear similarity measure between two feature vectors. The inner product can be evaluated directly in the input space, by applying the non-linear function Φ. This is often referred to as the *kernel trick (*Scholkopf, 2000). The goal is to embed data into a Hilbert space and then seek linear relations in this space.

Mercer's condition 1909 defines the general form of inner products in Hilbert spaces (Vapnik, 1995): If $K : X \times X \to \mathbb{R}$ is a continuous and symmetric real valued function on Hilbert space with a squire integral function $f \neq 0$, $\int_X f^2(\mathbf{x})d\mathbf{x} < \infty$, then:

$$\iint K(\mathbf{x}_i, \mathbf{x}_j) f(\mathbf{x}_i) f(\mathbf{x}_j) d\mathbf{x}_i d\mathbf{x}_j \geq 0$$

In that case these are the necessary and sufficient conditions to expand $K(\mathbf{x}_i, \mathbf{x}_j)$ (an inner product in some feature space) as a uniformly convergent series on $X \times X$:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{\infty} \lambda_r \cdot \Phi_r(\mathbf{x}_i) \cdot \Phi_r(\mathbf{x}_j), \quad \lambda_r > 0 .$$

These conditions are equivalent to $K(\mathbf{x}_i, \mathbf{x}_j)$ being a positive definite kernel (see Chapter 4). If K is a continuous kernel of a positive integral operator as defined by Mercer's condition, there exists a mapping $\mathbf{\Phi}$ of an input space into a space where the kernel can be represented as a inner product (3.26).

The corresponding quadratic programming problem to (3.25) is now:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{3.27}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 , \; 0 \leq \alpha_i \leq C, \; i = 1,...,m$$

Accordingly the decision function in the higher dimension feature space is:

$$f(\mathbf{x}) = \mathrm{sgn}(\sum_{i=1}^{m} \alpha_i y_i \cdot K(\mathbf{x}, \mathbf{x}_i) + \mathrm{b}).$$

If the kernel function satisfies Mercer's condition the solution of a convex optimisation problem given in (3.27) converges and is optimal. Several kernel functions that satisfy mentioned conditions are presented in Table 3.1.

Table 3.1: Examples of most common kernel functions.

Linear kernel	$K(\mathbf{x}_i,\mathbf{x}_j)=\left\langle \mathbf{x}_i,\mathbf{x}_j \right\rangle$	
Radial Basis Function kernel (RBFK)	$K(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\gamma\left\Vert \mathbf{x}_i-\mathbf{x}_j\right\Vert^2\right),$ $\gamma=\frac{1}{2\sigma^2},\sigma>0$	(3.28)
Laplace kernel (LK)	$K(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\gamma\left\vert \mathbf{x}_i-\mathbf{x}_j\right\vert\right),\gamma>0$	(3.29)
Polynomial kernel	$K(\mathbf{x}_i,\mathbf{x}_j)=(\gamma\left\langle \mathbf{x}_i,\mathbf{x}_j \right\rangle+r)^d,\gamma>0$	
Sigmoid kernel	$K(\mathbf{x}_i,\mathbf{x}_j)=\tanh(\gamma\left\langle \mathbf{x}_i,\mathbf{x}_j \right\rangle+r)$	

3.3.3 Sensitivity to Scale

Several ground truth databases, analyzed in Chapter 2, have various class size with different scales. In a retrieval applications there is a significant change in scale across different databases, as well as across classes within one database. Specifically in a RF scenario there is a strong disparity is scale between the class of interest and the rest of the database.

For a SVM classifier the scale sensitivity can be tuned up with the bound C on Lagrangian coefficients or the kernel scale parameter γ. If the bound C is set to a high value the allowed error is very small, and in case of a RF scenario when the most important samples are those first retrieved it does not significantly influence performances. However the kernel scale parameter γ depends on the size of classes and it is difficult to tune up. Figure 3.6 and Figure 3.7 show behavior of the RBF kernel for the “chessboard” and “spiral” classification problem for different boundary values and kernel scale parameters, respectively. Lighter areas correspond to values of the decision function closer to zero; the black line is the separation boundary, and the area between the hyphenated lines is the margin area. SVs are those points which have black borders;

In Figure 3.6 it is shown how increase in the boundary C reduces the number of misclassification errors, in case the boundary is infinity the soft classification problem reduces to hard binary SVM classification. On the other hand images on the top left side for both datasets, have the lowest boundary value and the highest misclassification rate.

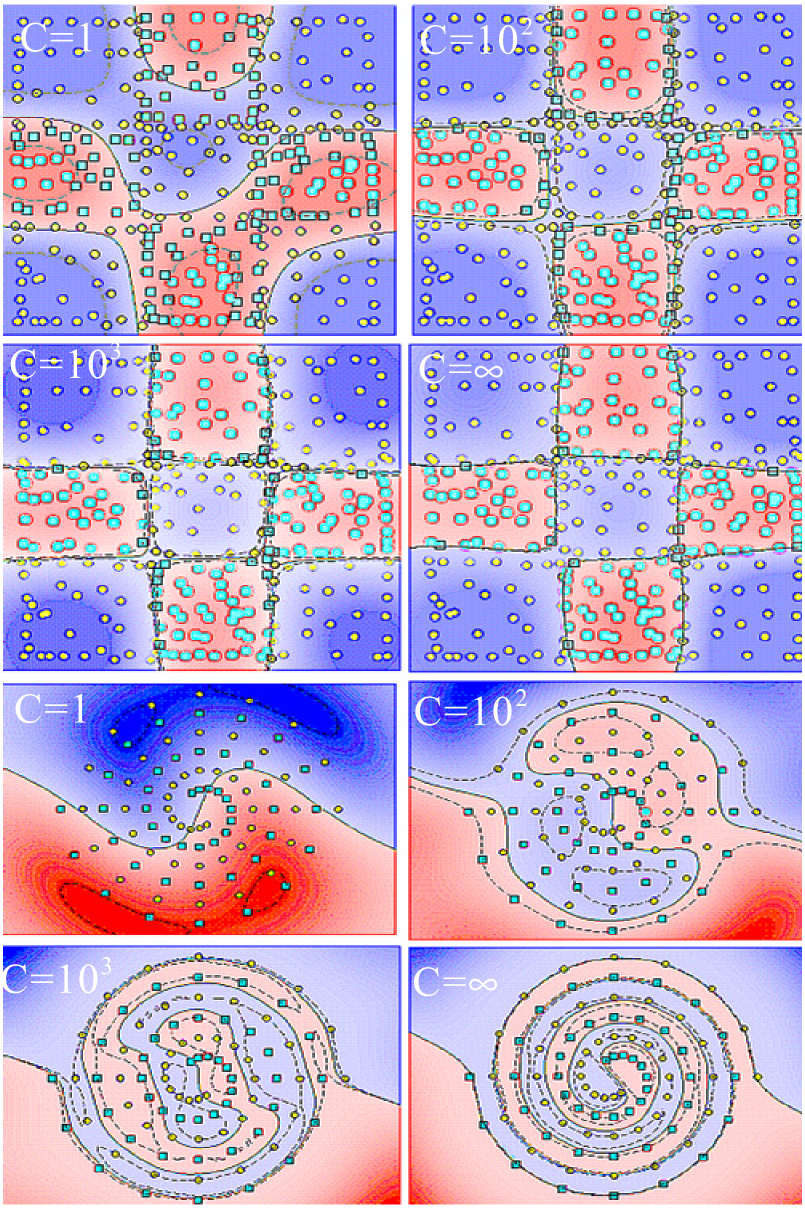


Figure 3.6: The chessboard (top four images) and the spiral (bottom four images) classification problem, RBF kernel for a fixed scale parameter $\gamma = 2, (\sigma = 0.5)$ *and variable bounds* $C = \{1, 10^2, 10^3, \infty\}$. *Smaller* C *leads to more misclassification errors.*

In Figure 3.7 it is shown how the scale parameter influences separation properties of RBF kernel. For $\sigma = 10$ the data in both chessboard and spiral dataset is under-fit and for $\sigma = 0.5$ the data is over-fit (see Figure 3.2)

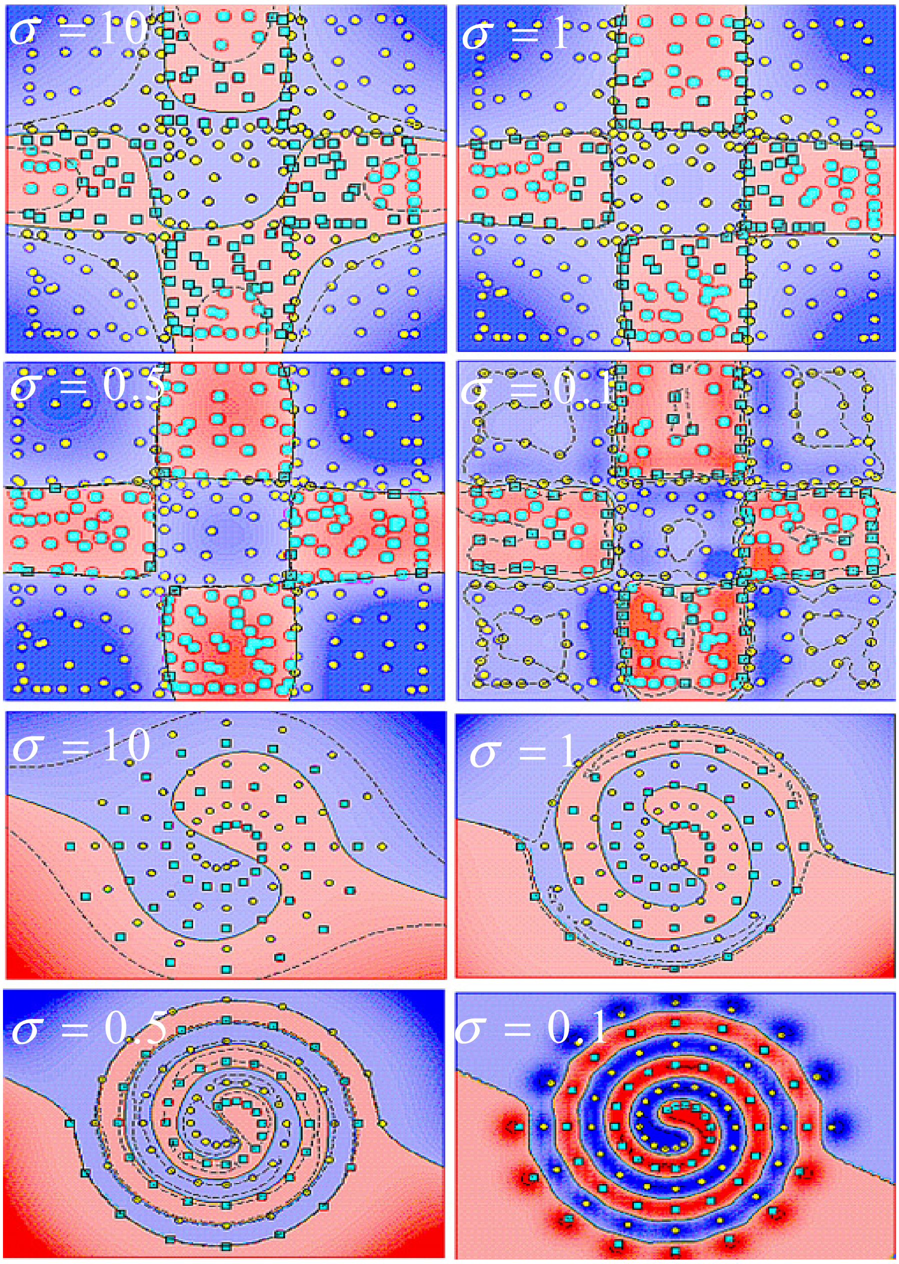


Figure 3.7: The chessboard (top four images) and spiral (bottom four images) classification problem, RBF kernel for a hard margin SVM problem with no errors allowed $C = \infty$ *and variable kernel scale parameter* $\sigma = \{10, 1, 0.5, 0.1\}$, $\gamma = \{0.05, 0.5, 2, 50\}$

3.3.4 Optimisation techniques for SVMs

Solving the SVM optimisation problem given in (3.27), under the given constraints, is a quadratic convex optimisation problem with local minimum being at the same time a global minimum. Even though this problem can be solved using classical quadratic optimisation approaches specially develop algorithms can be used to allow fast convergence.

Chunking (Vapnik, 1982) is an approach that exploits the sparsely of the solution for $\boldsymbol{\alpha}$ and the KKT conditions. At every step the problem containing non-zero α_i and some violating KKT conditions is solved. The size of the problem varies, and eventually it is reduced to the number of non-zero coefficients. This approach is suitable for large problems, however it is limited with maximal number of SVs that it can handle and requires a quadratic optimizer to solve the smaller problems.

Decomposition method (Osuna et al., 1997) solves a large quadratic problem by decomposing it into a sequence of small quadratic problems, with a fix size of sub-problems. As long as an example violates the KKT conditions it is added to the examples from previous sub-problems. In the end the sequence converges to an optimal solution. At each iteration one sample is added and removed; this allows arbitrary large training set however the convergence rate is low. And a quadratic optimizer is still necessary for solving the sub-problems.

Sequential Minimal Optimisation (SMO) proposed by Platt (1999a) can be viewed as a variation of the decomposition method. In each iteration a quadratic problem of size two is solved, since this can be done analytically there is no need for a quadratic optimizer. However the problem is how to choose good pair of variables to optimize in each iteration. This has been the method of choice in this work, and the algorithm implemented is an improvement of the initial Platt's idea (Keerthi et al., 2001). It is explained in more details in Appendix C.

3.3.5 Evaluation Results

In this part, the performances of several kernels with respect to changes in the scale parameter are compared. Sensitivity to changes in scale is very important when using RF with a generic image database, because the image classes may have different scales. Experiments on several ground truth databases show changes in effectiveness for different kernels.

Two classical kernels were considered, the RBFK and LK from Table 3.1. The RBF kernel is highly sensitive to scale parameter, and LK was proposed by Chapelle et al. (1999) for CBIR with RF. The third kernel is our adaptive convolution kernel (ACK) detailed in Chapter 4, and used here to investigate its sensitivity to scale. The experiment were performed on all six databases from Chapter 2 (for visual preview see

Appendix A) by varying the values for the scale parameter. In Figure 3.8-Figure 3.10 mean precisions over all classes for different values of the scale parameter γ (gamma), for each database from section 2.5.1 are shown. Sensitivity of the presented kernels to the changes in the scale parameter γ shows that depending on scales of the classes present in different databases each kernel has an optimal value. There is a huge instability in performances of both RBFK and LK specially when the scale parameter is larger than 10, whiles the proposed kernel ACK kernel shows higher stability even though not highest precision for all values of γ.

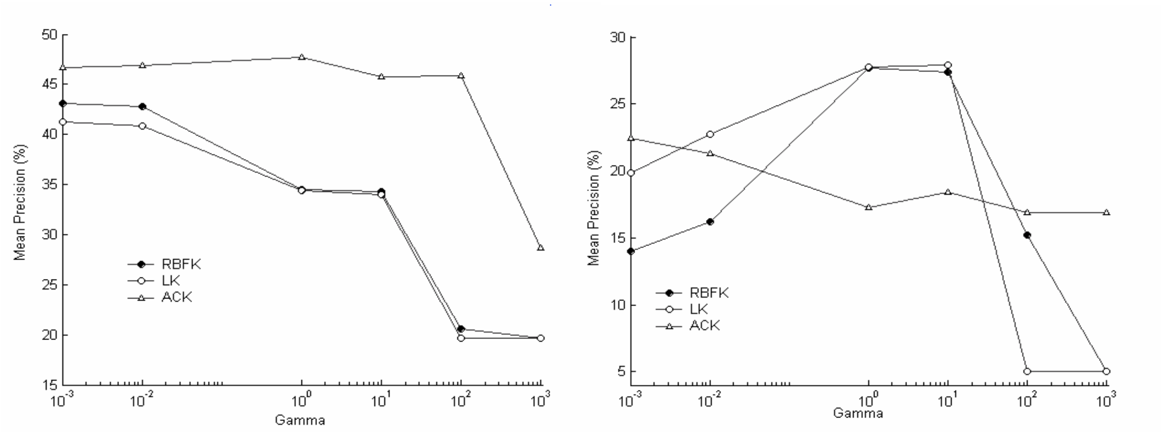


Figure 3.8: Mean precision depending on the scale parameter for SVMs (DColour and VisTex database).

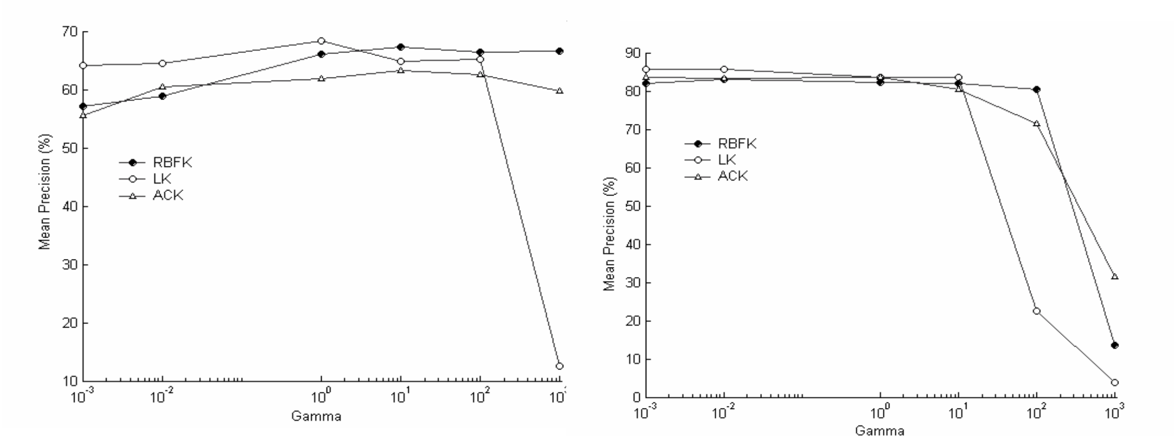


Figure 3.9: Mean precision depending on the scale parameter for SVMs (D8 database and D25-1800 database).

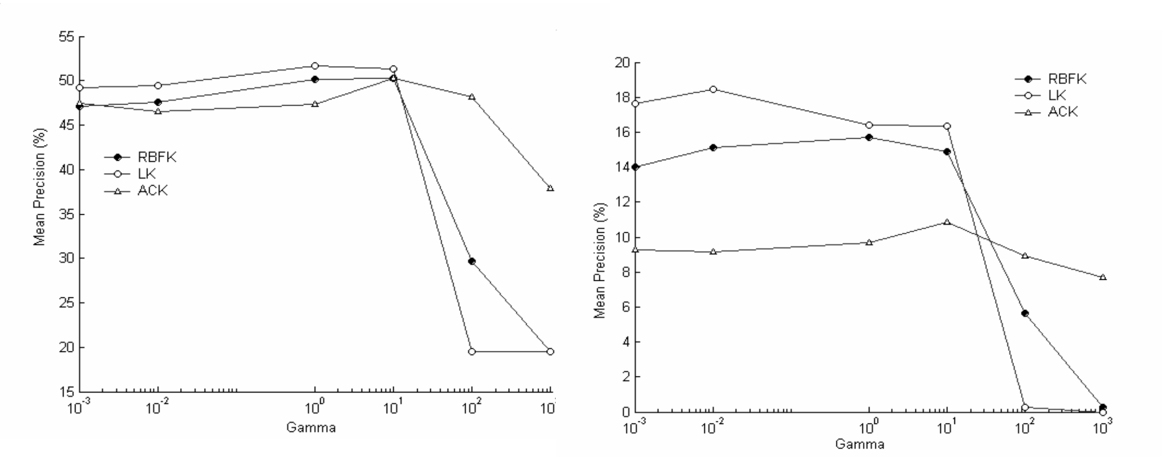


Figure 3.10: Mean precision depending on the scale parameter for SVMs (D7-700 and Caltech 101 database).

In real life applications, the scales of the user-defined classes cannot be known in advance therefore considerable variations performance of RF-based retrieval systems can be expected. For that reason selecting a kernel that is more scale-invariant is desirable for any RF system.

Comparisons between the two classical kernels RBFK and LK are shown in the following part of this sub-section (Figure 3.11-Figure 3.15). For the two mentioned kernels the most positive selection strategy was used (Ferecatu et al., 2004). Various descriptors or descriptor combinations, as in section 2.5.5 were also analyzed in conjunction with SVMs. The descriptors are either individual descriptors or normalized concatenations for CLD and EHD (denoted as CLEH), CSD and EHD (denoted as CSEH), as well as all the descriptors together, except DCD (denoted as CONC). Though there is no mathematical justification for simple feature concatenations (see Chapter 4), Chapelle et al. (1999) have experimentally proven that feature combinations usually perform better then individual descriptors. For both kernels the scale parameters were set to their optimal values for a particular database.

The VisTex database was excluded form these experiments since some of the classes are of very small size and are not suitable for RF approaches with at least 10 samples added in each iteration. Please note that comparisons to the newly devised kernel ACK are presented in the following chapter along with a detailed description of the kernel itself.

The search sessions were initialized with three randomly chosen sets with six labeled image, half for relevant samples and half for irrelevant samples. The target of each RF session is to find all images in a class based on the initial relevant set. The user is presented with images that have the highest certainty of being positive “most positive”. Each relevance feedback session has four to five iterations. Precision within the returned samples size equaling to the class size is measured. In the ideal case the system would return all relevant images and obtain recall value of one. Average precision over all classes for a particular database provides a measure of how well the relevance feedback system performs through iterations, in a category search scenario.

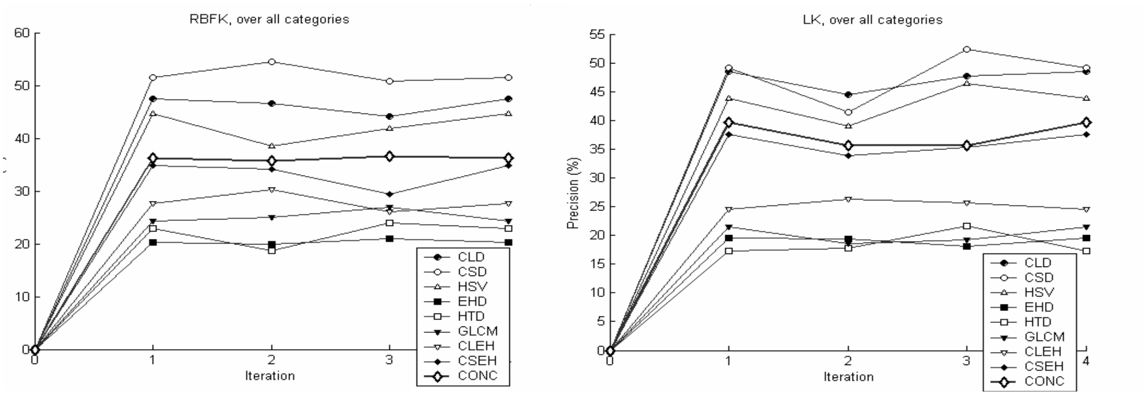


Figure 3.11: Average precision trough iteration for RBFK (left) and LK (right), for the DColour image database.

For the DColour database (see Figure 3.11) with classes based on low-level visual features (e.g. red, blue, yellow, green, and orange), as expected the descriptor combinations fail to give good precision results and winning performance is obtained for SVMs with only colour features. However, this type of databases are not real world databases, they were just used to confirm effectiveness of SVMs in image based RF methods.

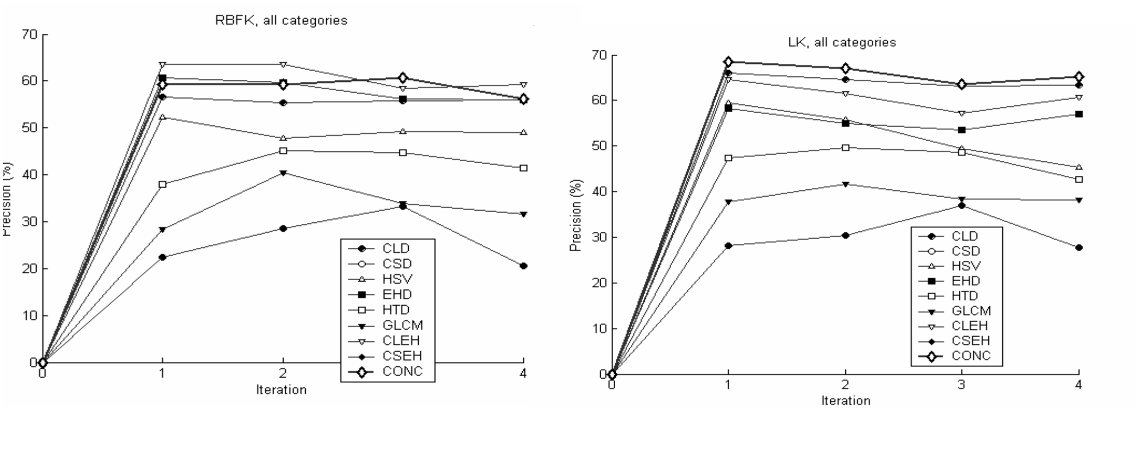


Figure 3.12: Average precision trough iteration for RBFK (left) and LK (right), for the D8 image database.

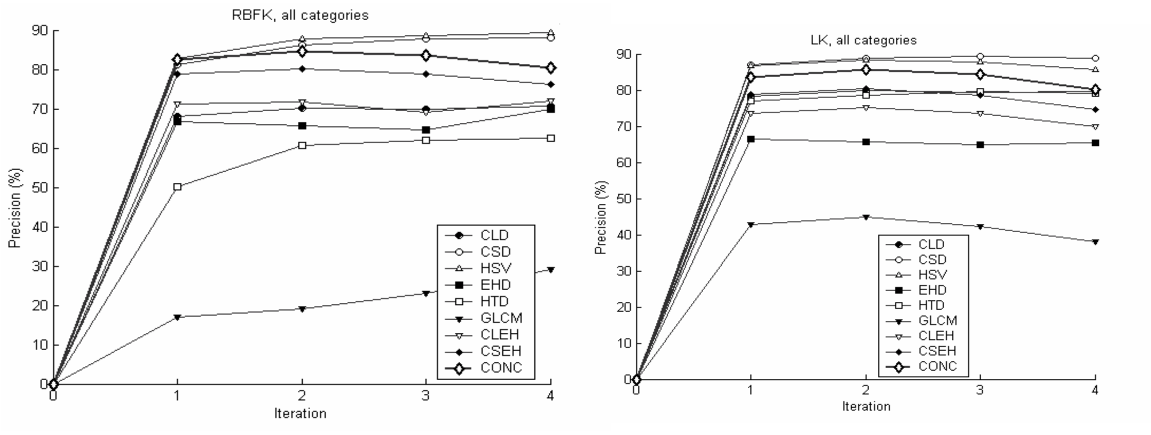


Figure 3.13: Average precision trough iteration for RBFK (left) and LK (right), for the D25-1800 image database.

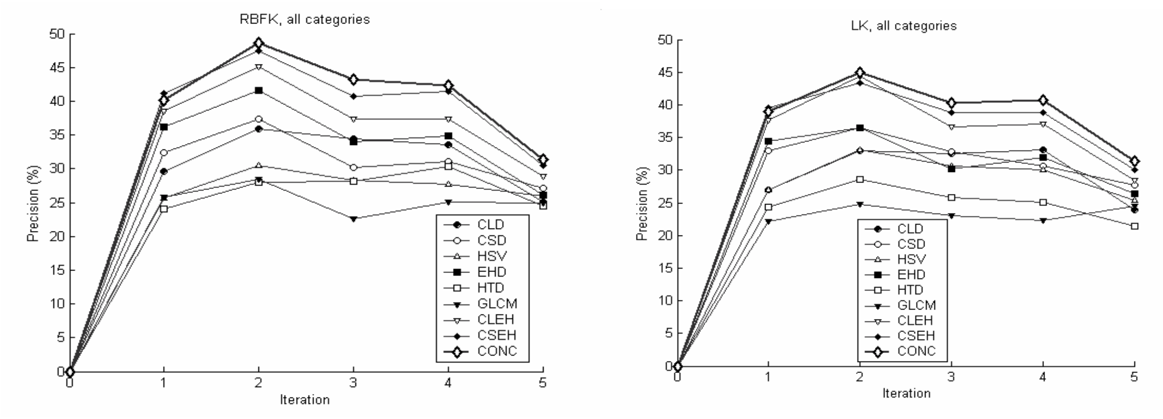


Figure 3.14: Average precision trough iteration for RBFK (left) and LK (right),for the D7-700 image database.

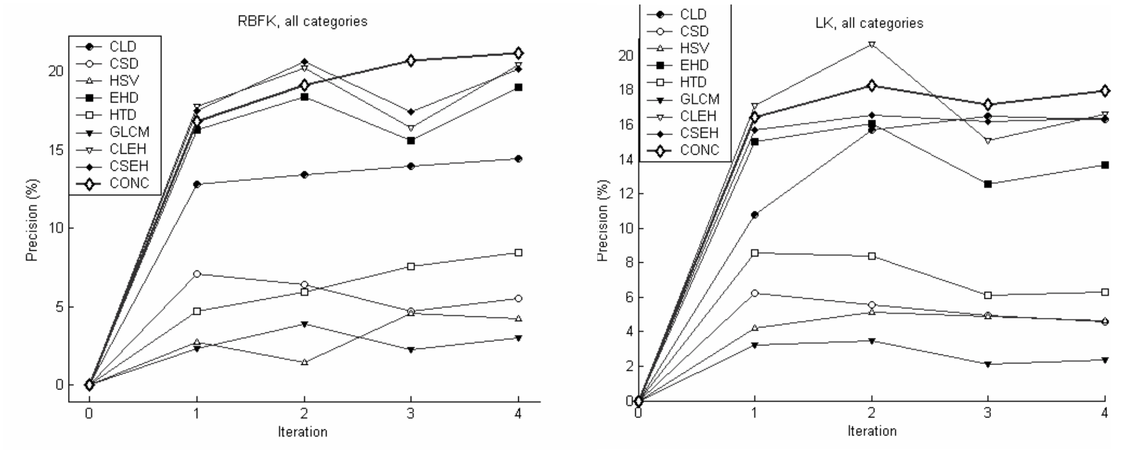


Figure 3.15: Average precision trough iteration for RBFK (left) and LK (right), for Caltech 101 image database.

For the rest of the databases with “more” semantically meaningful classes (see Figure 3.12- Figure 3.15) SVMs in combination with concatenated descriptors mainly give highest performances over all iterations as expected.

In the next paragraph precision-recall curves are presented. The precision values are averaged over same relative scopes, that is the ratio of the relevant retrieved images and the size of that class. Here the aim is to show how more training information influences the results of SVMs through iterations. Since it is difficult to show all results, only the result for the database that is closest to real world databases D7-700 database (see Appendix A, Figure A. 5) is presented.

Both for RBFK Figure 3.16 and LK Figure 3.17 average precision-recall curves for a combination of multiple descriptor perform better then for an individual feature. However, in both cases the precision-recall curve at bottom right of each figure representing precision-recall over different iterations does not show a conclusive improvement. That is more training samples does not necessarily improve the quality of results.

Since both, the feature space and the learning method have been investigated, the way of improving performances is to adapt one to another, that is to improve the learning approach to obtain as much as possible relevant information from the available features, and ignore the nosy data. This is the subject of Chapter 4.

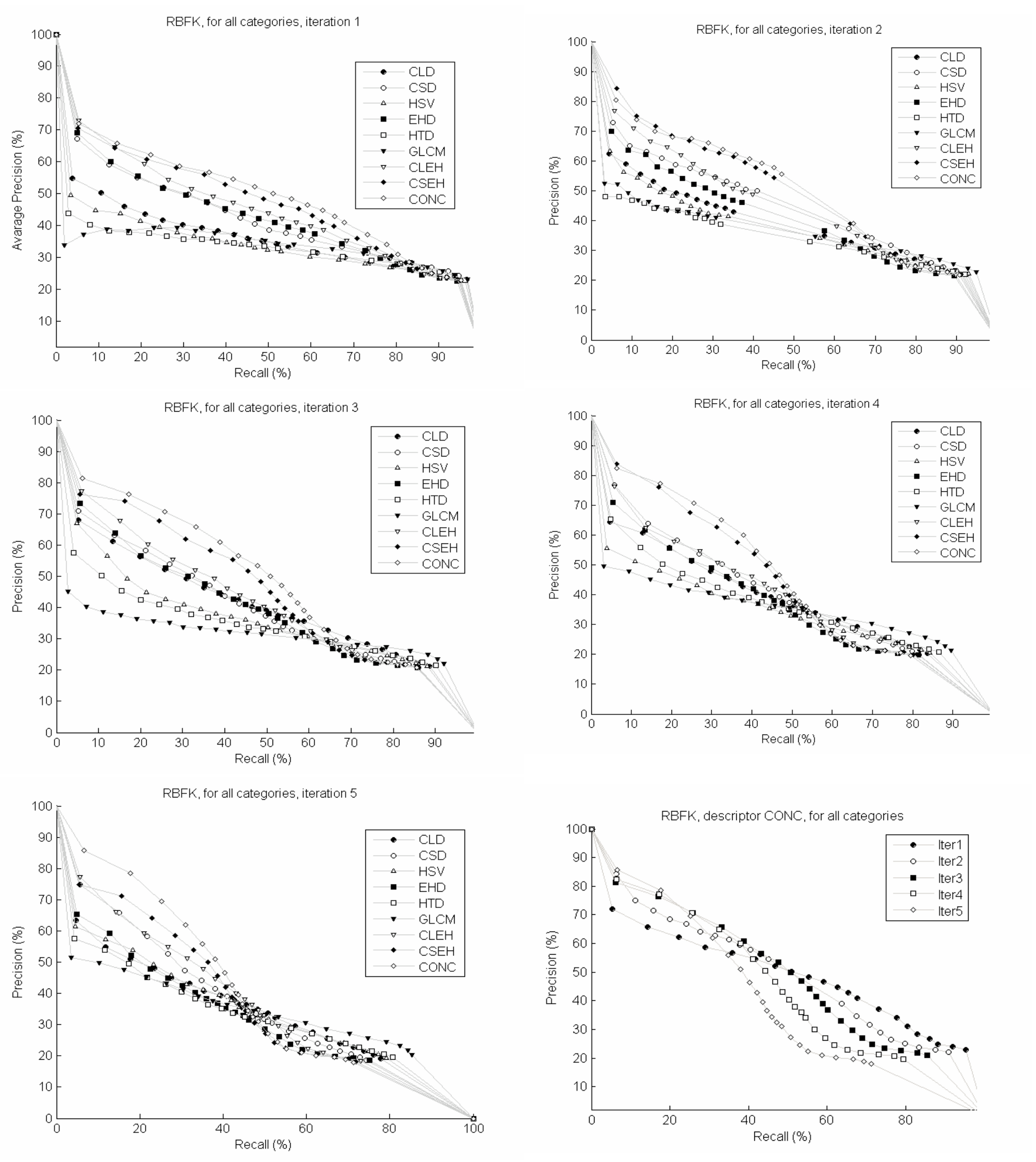


Figure 3.16: Average precision-recall curves for RBFK over all categories. Five iterations (top to bottom, left to right). Comparative results through iterations for the best performing feature CONC are given in the bottom right corner.

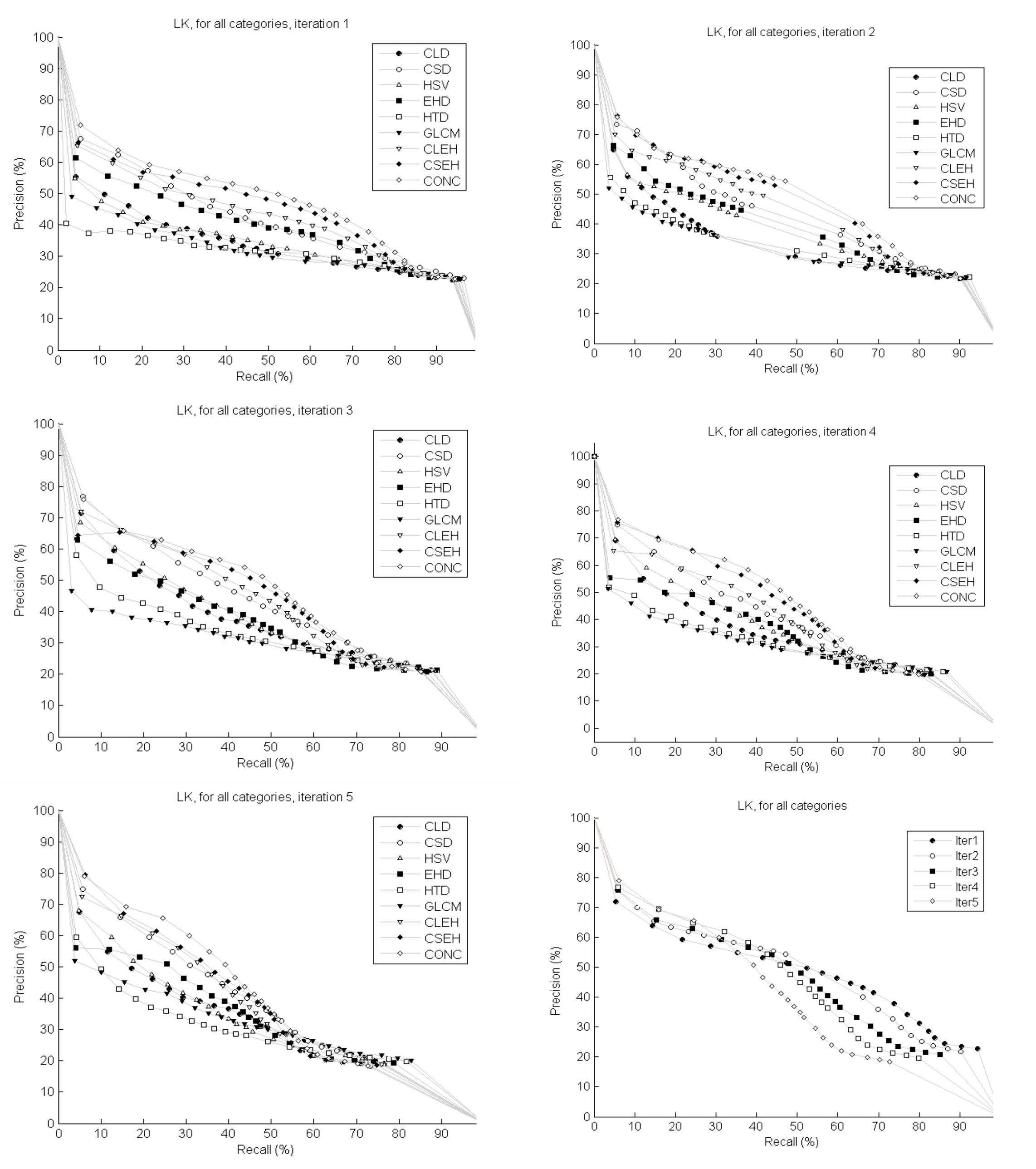


Figure 3.17: Average precision-recall curves for LK over all categories. Five iterations (top to bottom, left to right). Comparative results through iterations for the best performing feature CONC are given in bottom right corner.

CHAPTER 4 : A Multi-feature Scenario for Relevance Feedback

4.1. Adaptive Convolution Kernels in Multi-feature Spaces

Since no single descriptor is able to represent all the properties and patterns encapsulated in natural images, the combination of several descriptors appears to be a sensible strategy to increase their discrimination power and classification properties.

Several approaches for image retrieval are based on distance re-weighting, by exploiting relevance feedback from the user. The evolution of these learners for RF is presented in Chapter 3. In the approach proposed by Jing et al. (2004) regions are combined in image-to-image similarity by using Earth Mover's Distance and SVMs. However the mathematical aspect of positive–definite kernels for SVMs that guarantees convergence and uniqueness of the optimisation problem has not been analyzed. Chapelle et al. (1999) also obtained promising results when using global image features and SVMs. The authors show improvements when introducing different similarity measures but acknowledge the lack of proof for positive definiteness of several used kernels.

In order to effectively approach the problem, a new kernel based on specific distances for different feature subspaces is presented. The new kernel exploits the nature of the data and weights differently each feature subspace. The kernel can be dynamically adapted to user preferences when applied in a relevance feedback scenario. To assure convergence of the optimization problem in SVMs and uniqueness of the obtained solution, positive definiteness of the proposed kernel is analyzed.

4.1.1 Feature Subspaces, Similarity Measures

A number of different feature can be extracted from image content in order to obtain information at various levels of abstraction. Te goal is to as much as possible simulate human visual perception and extract information that could allow higher levels of conceptual abstraction.

The difficulty of the problem rises from generic and intrinsically different nature of visual descriptors. They are usually formed using a number of different algorithms and they have individually specific syntax. The underling consequence is that different descriptors 'live' in completely different feature space with their own similarity measures. Though their extraction, representation, statistical behaviour and similarity measures are designed, as much as possible, to simulate human perception, they do not naturally and straightforwardly mix into a meaningful combination.

In this approach, the descriptors introduced in Chapter 2 were considered. The multi-feature space for images is defined as a structured data space out of the following low-level descriptors: CLD, CSD, DCD, EHD, HTD, HSV histogram space and GLCM. Note that the first five descriptors are MPEG-7 standard descriptors. Some of the descriptor components are colour space values of histograms e.g., CLD, CSD, some of them include statistical moments of the coefficient e.g., HTD, GLCM. Hence the metric space induced for each descriptor tries to exploit the specific syntax and physical nature of the coefficients obtained. The distance between MPEG-7 descriptors is estimated using metrics recommended by the non-normative part of MPEG-7 standard, specifically specified for retrieval and browsing applications. For the rest of the descriptors the distances were designed to suit their syntax based on the feature extraction procedure.

In this section we give a brief summery of the distance measures used for each descriptor, as explained in Chapter 2, this is done through introducing the following notation.

Let $X^{(l)}, l = 1,..,L$ be a feature space endowed with a similarity measure $d^{(l)}$. Where L is the number of different feature spaces. Observe that $d^{(l)}$ is a distance function, in case $d^{(l)}$ is a metric the feature space $(X^{(l)}, d^{(l)})$ becomes a metric space. Let $\mathbf{x}_i^{(l)}$ be a

i-th vector element of $X^{(l)}$ with dimension $N(X^{(l)})$. The dimension of the feature space $N(X^{(l)})$ depends on the space itself . The superscript of the feature space, and its elements are not important in case only an individual feature space is considered, hence in these cases for simplicity of notation it can be neglected, then $\mathbf{x}_i \in X$, $\mathbf{x}_i = [x_{i,1}, x_{i,2}, ..., x_{i,N}]$.

The mentioned descriptors ad their distances are:

- CLD, the distance function is the recommended MPEG-7 distance (2.11):

$$d_{cld}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{3} \sqrt{\sum_{s \in S_i} a_{rs} \cdot (x_{i,s} - x_{j,s})^2} . \quad (4.1)$$

  Where a_{rs} are weights for different components in the YCrCb colour system. And S_i is a set of free parameters enabling various numbers of coefficients for each of the three components.

- CSD, the distance metric is normalized L_1 distance:

$$d_{csd}(\mathbf{x}_i, \mathbf{x}_j) = \sum_r \left| \frac{x_{i,r} - x_{j,r}}{a} \right| . \quad (4.2)$$

- DCD, the feature vector is made up of 4-touples of elements, $\mathbf{x}_i = \{(x_{i,r}, \mathbf{c}_{i,r})\}$, $r = 1,..., N(\mathbf{x}_i)$, where $\mathbf{c}_{i,r}$ is the r-th 3-dimensional colour component in RGB colour system, $x_{i,r}$ is the percentage of pixels that have corresponding colour values for the r-th dominant colour. This feature space is not a conventional vector or even metric space. Furthermore, the dimension of each feature $N(\mathbf{x}_i)$ is variable. The distance measure for the DCD is the quadratic form given as in (2.14):

$$d_{dcd}(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{r=1}^{N(\mathbf{x}_i)} \sum_{s=1}^{N(\mathbf{x}_j)} (x_{i,r}{}^2 a_{rr} + x_{j,s}{}^2 a_{ss} - 2x_{i,r} x_{j,s} a_{rs}) \right)^{1/2} . \quad (4.3)$$

  Here, $0 < a_{rs} \le 1$ is the similarity coefficient between two colours.

- EHD, the distance is a sum of L_1 distances over the original features, as well as global $x^g(\cdot)$ and semi-global $x^s(\cdot)$ histograms values (see Figure 2.3):

$$d_{ehd}(\mathbf{x}_i,\mathbf{x}_j)=\sum_{r=0}^{79}\left|x_{i,r}-x_{j,r}\right|+5\cdot\sum_{r=0}^{4}\left|x_{i,r}^{g}-x_{j,r}^{g}\right|+\sum_{r=0}^{64}\left|x_{i,r}^{s}-x_{j,r}^{s}\right|. \qquad (4.4)$$

- HTD, the distance metric is normalized L_1 distance:

$$d_{htd}(\mathbf{x}_i,\mathbf{x}_j)=\sum_{r}\left|x_{i,r}-x_{j,r}\right|/a(r) \qquad (4.5)$$

Here $a(r)$ stands for normalizing standard deviation of appropriate features components.

- HSV histogram, the distance metric is histogram intersection

$$d_{hsv}(\mathbf{x}_i,\mathbf{x}_j)=1-\sum_{r}\min(x_{i,r},x_{j,r}), \qquad (4.6)$$

- GLCM, the metric is a typical L_2 metric with the values of all coefficients normalised over the database, where $a(r)$ are column wise normalisation factors:

$$d_{glcm}(\mathbf{x}_i,\mathbf{x}_j)=\left(\sum_{r}\left(\frac{x_{i,r}-x_{j,r}}{a(r)}\right)^2\right)^{1/2}. \qquad (4.7)$$

It is apparent that for instance even though GLCM coefficients have higher syntactical meaning, the L_2 norm is still used risking that the underlining meaning of particular coefficients loses meaning. Hence it is important not only to develop descriptors that would as much as possible simulate human understanding of content but also to develop the measures used for human based retrieval and browsing scenario.

The problem at hand is how to combine presented distance functions since the result of each distance is a scalar value representing a level of similarity for that low-level primitive. To sensibly combining the descriptors a logical distance measures in line with the particular syntax of each descriptor needs to be used. Since it is difficult to assume levels of importance for each of the features a linear combination of individual metrics is a straightforward way to obtain similarity in the joint feature space.

Now, let $X=X^{(1)}\times X^{(2)}\times...\times X^{(L)}$ denote a Cartesian product of L individual feature spaces, and an element of that space $\mathbf{x}_i\in X$, where $\mathbf{x}_i=[\mathbf{x}_i^{(1)},..,\mathbf{x}_i^{(L)}]$. Here $i=1,...,m$

is the number of elements in the joint feature space X. The distance between feature vectors ca be denoted as follows:

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{L} \omega_l d^{(l)}(\mathbf{x}_i^{(l)}, \mathbf{x}_j^{(l)}), \tag{4.8}$$

Here $d^{(l)}$ represents one of the distance from (4.1)-(4.7), in the feature space $X^{(l)}$, and as mentioned L is the number of feature spaces. The weights ω_l are updated based on user relevance feedback as explained in the next paragraph.

The learning approach for RF is based on SVMs. A kernel based on linear combination of distances (4.8) is analyzed. It exploits the nature of the data and weights differently each feature subspace. Weights for each feature sub-spaces within the kernel are dynamically calculated from the intrarclass variance of user's feedback.

The weights ω_l are calculated from elements labelled as relevant and irrelevant by the user, since SVM is supervised statistical learning approach. Hence,

$$\omega_l = Var_{neg}^{(l)} \Big/ Var_{pos}^{(l)} . \tag{4.9}$$

Normalized to $\sum_{l=1}^{L} \omega_l = 1$. If positive examples have some commonality in low-level feature, the variance of distances $Var_{pos}^{(l)}$ is small. In that case the weight to that distance measure is large. However if the negative examples have also small variance $Var_{neg}^{(l)}$ then that distance does not have good discriminative properties and the weight is lowered. Instead of fixing the weighting factors based on the input, the approach keeps updating through iterations.

Therefore by embedding the distance function into a classical kernel, the new kernel is modelled not only to follow the rules and nature of the patterns used but also to learn over time a better representation of similarity for the structured data based on user's feedback. However to guarantee convergence of the optimization problem in SVMs as well as uniqueness of the obtained solution, the proposed kernel is analyzed in the following section.

4.1.2 Positive Definite Property of Kernel Functions

Whilst the basic algorithms for SVMs (Chapter 3) are theoretically well defined, finding optimal representation of real life data e.g. natural images and appropriate kernel based similarity matching is still an open issue. SVMs are optimal hyperplane classifiers acting over a well-defined inner product in feature space X.

In order to obtain a separable classification problem in a feature space, many dimensions have to be added to the input data. Kernels perform mapping of boundary between classes, from feature space X into a non-linear separable bound in a higher dimensional feature space. Given a symmetric positive kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ Mercer's theorem indicates that there exist a mapping $\mathbf{\Phi}$:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \right\rangle . \tag{4.10}$$

For a kernel satisfying the assumptions of Mercer's theorem the expression in (4.10) holds true. This theorem states that the kernel needs to be symmetric positive definite to guarantee convergence of the convex optimisation problem and thus uniqueness of the global solution (e.g., optimisation problem in SVMs (3.27)) .

Definition 4.1: A real valued kernel $K : X \times X \to \mathbb{R}$ is positive definite (PD) if and only if K is symmetric $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$, for all $\mathbf{x}_i, \mathbf{x}_j \in X$ and

$$\sum_{i,j=1}^{m} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

for $m \in N,\ c_i \in \mathbb{R}, \mathbf{x}_i, \in X,\ i = 1, ..., m$.

To imply a direct connection between Mercer's theorem (see Chapter 3) and PD kernels let $X = [a, b]$ be a compact interval and let $K : X \times X \to \mathbb{R}$ be continuous and symmetric real valued function then K is positive definite kernel if and only if

$$\int_a^b \int_a^b f(\mathbf{x}_i) f(\mathbf{x}_j) K(\mathbf{x}_i, \mathbf{x}_j) d\mathbf{x}_i d\mathbf{x}_j \geq 0$$

holds for all continuous functions $f : [a, b] \to \mathbb{R}$.

Though this shows how positive definite kernels are suitable for SVM, a relaxation of PD property for convexity of SVMs was presented through a family of conditionally positive definite kernels (Scholkopf, 2000).

Definition 4.2: A real valued kernel $K : X \times X \to \mathbb{R}$ is conditionally positive definite (CPD) if and only if K is symmetric $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$, for all $\mathbf{x}_i, \mathbf{x}_j \in X$ and

$$\sum_{i,j=1}^{m} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \quad \text{for } m \in N,\ c_i \in \mathbb{R}\ , \mathbf{x}_i, \in X, i = 1, ..., m$$

$$\text{with constraint } \sum\nolimits_{i=1}^{m} c_i = 0 .$$

Berg et al. (1984) considered a link between PD kernels and CPD kernels, and the possible application to SVM design, as given in the following proposition.

Proposition 4.3: Let K be a symmetric kernel on $X \times X$ then

$$\hat{K}(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) - K(\mathbf{x}_i, \mathbf{x}_0) - K(\mathbf{x}_j, \mathbf{x}_0) + K(\mathbf{x}_0, \mathbf{x}_0) \tag{4.11}$$

is PD if and only if K is CPD kernel.

It is necessary to show that CPD kernels are sufficient to enable convexity of the following SVM optimisation problem. Let $\hat{K}(\mathbf{x}_i \cdot \mathbf{x}_j)$ be a PD kernel, then SVM optimisation problem is :

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \hat{K}(\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 ,\ 0 \leq \alpha_i \leq C,\ i = 1, ..., m .$$

Based on previous proposition and condition $\sum_{i=1}^{m} \alpha_i y_i = 0$,

$$\begin{aligned}
&\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \hat{K}(\mathbf{x}_i \cdot \mathbf{x}_j) \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2} \underbrace{\sum_{j=1}^{m} \alpha_j y_j}_{0} \sum_{i=1}^{m} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_0). \\
&+ \frac{1}{2} \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{0} \sum_{j=1}^{m} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_0) - \frac{1}{2} \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{0} \underbrace{\sum_{j=1}^{m} \alpha_j y_j}_{0} K(\mathbf{x}_0, \mathbf{x}_0) \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j).
\end{aligned}$$

Hence, the optimisation problem defined with CPD kernels can be re-written so it is expressed through PD kernels, only. Therefore the use of CPD kernels in SVMs is equivalent to using corresponding PD kernels.

Furthermore the class of kernels on a set is closed under properties of addition, multiplication by a positive constant, product and pointwise limits (Haussler 1999).

4.1.3 Adaptive R-convolution Kernels

R-convolution kernels have been introduced by Haussler (1999) leading to a new class of kernels on structured data. Let $X^{(l)}$, $1 \le l \le L$ be separable non-empty metric spaces and $X = \bigcup_{l=1}^{L} X^{(l)}$ a composite structure. Decomposition of structured data object $\mathbf{x} \in X$ is specified with relation R into a finite set of tuple sub-components. $R((\mathbf{x}^{(1)},...,\mathbf{x}^{(L)}),\mathbf{x})$ indicates decomposition of $\mathbf{x}$ into components $\mathbf{x}_1,...,\mathbf{x}_L$, with $\mathbf{x}_l \in X^{(l)}$ and accompanying kernels $K_1,...,K_L$ for every component. Hence the set of all possible decomposition of $\mathbf{x}$ is given by $R^{-1}(\mathbf{x})$.

Let R be a decomposition structure for a particular structure data type on a set X, $\mathbf{x}_i, \mathbf{x}_j \in X$ are decomposition into $\mathbf{x}_i = (\mathbf{x}_i^{(1)},...,\mathbf{x}_i^{(L)})$, $\mathbf{x}_j = (\mathbf{x}_j^{(1)},...,\mathbf{x}_j^{(L)})$. For each of the subcomponents $\mathbf{x}_i^{(l)}, \mathbf{x}_j^{(l)}$ the measure of similarity is given with kernel $K_l(\mathbf{x}_i^{(l)}, \mathbf{x}_j^{(l)})$ defined on $X^{(l)}$. The associated convolution kernel is defined as:

$$K_R(\mathbf{x}_i, \mathbf{x}_j) = \sum_{\mathbf{x}_i \in R^{-1}(\mathbf{x}_i)} \sum_{\mathbf{x}_j \in R^{-1}(\mathbf{x}_j)} \prod_{l=1}^{L} K_l(\mathbf{x}_i^{(l)}, \mathbf{x}_j^{(l)}) \tag{4.12}$$

Now all the necessary elements are obtained to define the adaptive convolution kernel (ACK) in respect to specific descriptor syntax. The multi-feature space can be defined as a structured data space out of a number of descriptor spaces. Every feature is a structured data vector, formed by combining simpler components into a complex representation. The decomposition structure is given as $\mathbf{x}_i = \{\mathbf{x}_i^{(l)}\}$, $i = 1,...,L$, with every element being a vector on its own of dimension m_i, $\mathbf{x}_i^{(l)} = (\mathbf{x}_{i,1}^{(l)},...,\mathbf{x}_{i,m_i}^{(l)})$. As mentioned in previous section the dimension depends on the nature of the feature space.

The following theorem is needed to prove that the convolution kernel satisfies required conditions to be a PD kernel. For the proof the reader is referred to Haussler (1999)

Theorem 4.4 [R-convolution kernels]: If $K_1, K_2, ..., K_L$ are PD kernels on $X_1 \times X_1, X_2 \times X_2, .., X_L \times X_L$ and R is finite decomposition on X then convolution $K_1 \cdot K_2 \cdot \ldots \cdot K_L$ is a PD kernel on $X \times X$.

It is expected that the complex nature of the patterns in images can be learned over time when using the combined distance function (4.8). Clearly, this distance function can be embedded in any classical kernel. In the sequel the exponential kernel will be used for the SVM based learning approach. This kernel is referred to as adaptive convolution kernel (ACK)(Djordjevic and Izquierdo, 2006c):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\rho D(\mathbf{x}_i, \mathbf{x}_j)) \tag{4.13}$$

where $D(\mathbf{x}_i, \mathbf{x}_j)$ can be any distance and ρ is a positive constant multiplier. However without further analysis the kernel presented in (4.13) does not necessarily satisfies positive definiteness conditions of Mercer's theorem. By using the decomposition structure defined for convolution kernel (4.12) the following expression is obtained:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\rho \sum_{l=1}^{L} \omega_l \cdot d^{(l)}(\mathbf{x}_i^{(l)}, \mathbf{x}_j^{(l)})) = K_1 \cdot K_2 \cdot \ldots \cdot K_L \tag{4.14}$$

$K(\mathbf{x}_i, \mathbf{x}_j)$ from (4.14) is a convolution PD kernel if the individual components satisfy conditions to be PD kernels, based on Theorem 4.4 . Using the following few theorems and proposition positive definiteness for the particular descriptor combination from S4.1.1 is proven.

Theorem 4.5: K is CPD kernel over $X \times X$ if and only if $\exp(uK)$ is PD for all $u > 0$

Proof: If $\exp(uK)$ is PD, then knowing that a class of kernels is closed under pointwise limit, $K = \lim_{t \to 0+} \frac{1}{t} \exp(uK)$ is PD and also CPD. If $K(\mathbf{x}_i, \mathbf{x}_j)$ is CPD and $\hat{K}(\mathbf{x}_i, \mathbf{x}_j)$ is PD kernel, based on (4.11) and $\sum_{i=1}^{m} c_i = 0$ for CPD kernels, then:

$$\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \exp(uK(\mathbf{x}_i,\mathbf{x}_j)) = \sum_{i=1}^{m}\sum_{j=1}^{m} \hat{c}_i \hat{c}_j \exp(u\hat{K}(\mathbf{x}_i,\mathbf{x}_j)) + \underbrace{\sum_{j=1}^{m} c_j}_{0} \sum_{i=1}^{m} \exp(-uK(\mathbf{x}_i,\mathbf{x}_0))$$

$$+\underbrace{\sum_{i=1}^{m} c_i}_{0} \sum_{j=1}^{m} c_j \exp(-uK(\mathbf{x}_j,\mathbf{x}_0)) + \underbrace{\sum_{i=1}^{m} c_i}_{0} \underbrace{\sum_{j=1}^{m} c_j}_{0} \exp(uK(\mathbf{x}_0,\mathbf{x}_0))$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{m} \hat{c}_i \hat{c}_j \exp(\hat{K}(\mathbf{x}_i,\mathbf{x}_j)) \geq 0$$

Hence $\exp(uK)$ is PD kernel. ∎

Proposition 4.6: If K is a negative function $K \leq 0$, and CPD kernel then $-(-K)^{\alpha}, 0 < \alpha < 1$ is also CPD.

From Berg et al. (1984) it follows that $-(-K)^{\alpha} = \frac{\alpha}{\Gamma(\alpha-1)} \int_0^{\infty} (e^{\lambda K} - 1) \frac{d\lambda}{\lambda^{\alpha+1}}$. Assuming that K is CPD and based on Theorem 4.5, $e^{\lambda K}$ is CPD; the right hand side is now a sum of CPD kernels and hence it is a CPD kernel itself.

In order to prove that the proposed kernel in (4.14), is PD the following corollary is used to build valid kernels out of presented similarity measures.

Corollary 4.7: Let K be a symmetric, negative kernel $K : X \times X \to (-\infty, 0)$ which is also conditionally positive definite (CPD), then

$$\exp(-\rho \cdot (-K)^{\alpha}) \quad 0 < \alpha < 1, \rho > 0 \tag{4.15}$$

is a PD kernel. The proof is implicitly derived from Theorem 4.5 and Proposition 4.6.

Positive Definite Kernel based on L2 distances

In this section positive definiteness of K_{cld} and K_{glcm} kernels is proven. These kernels were formed by replacing the distance metric for CLD and GLCM from (4.1) and (4.7), into individual kernels from (4.14), designed based on Corollary 4.7 and L_2 distance.

Positive definite kernels can be considered to be a nonlinear generalisation of the simplest similarity measure, the inner product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle, \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^n$. A norm on a space X can be defined based on a strict inner product as :

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \tag{4.16}$$

In this case the associated distance between two vectors is given as $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$. In case of GLCM descriptor the distance given in (4.7) is inferred from L_2 norm. Based on (4.15) the aim is to prove that an exponent of negative L_2 distance is CPD. Though this is a trivial case it is important to describe it for the completeness of the approach.

Initially the kernel defined with a negative value of squared distance of the norm from (4.16), $K(\mathbf{x}_i, \mathbf{x}_j) = -\|\mathbf{x}_i - \mathbf{x}_j\|^2$, needs to be symmetric, negative and CPD. Indeed it is symmetric $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i) = -\|\mathbf{x}_i\|^2 + 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle - \|\mathbf{x}_j\|^2$, and CPD with the condition $\sum_{i=1}^{m} c_i = 0$:

$$\sum_{i,j=1}^{m} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) = -\sum_{i=1}^{m} c_i \|\mathbf{x}_i\|^2 + 2\sum_{i,j=1}^{m} c_i c_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{j=1}^{m} c_j \|\mathbf{x}_j\|^2 = 2\sum_{i,j=1}^{m} c_i c_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq 0$$

The third necessary property is that it is negative: $K(\mathbf{x}_i, \mathbf{x}_j) = -\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq 0$. Therefore based on these three properties and Corollary 4.7 , the following kernel is PD:

$$\begin{gathered}
\exp(-\rho \cdot (-K)^{\alpha}) = \exp(-\rho \cdot (-(-\|\mathbf{x}_i - \mathbf{x}_j\|^2))^{\alpha}) = \\
= \exp(-\rho \cdot \|\mathbf{x}_i - \mathbf{x}_j\|^{2\alpha}) = \exp(-\rho \cdot \|\mathbf{x}_i - \mathbf{x}_j\|^{\beta}) \\
0 < \alpha < 1 \Rightarrow 0 < \beta \leq 2, \rho > 0
\end{gathered}$$

K_{glcm} is a version of this kernel and hence it is a symmetric positive definite kernel for $\beta = 1$:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\rho \cdot \|\mathbf{x}_i - \mathbf{x}_j\|) \tag{4.17}$$

In the case of CLD, K_{cld} represents a convolution of 3 kernels, since this feature space is further decomposed into 3 subcomponents based on d_{cld} from (4.1). The kernel for CLD is given as:

$$\begin{aligned} K_{cld}(\mathbf{x}_i,\mathbf{x}_j) &= \exp(-\rho\cdot d_{cld}(\mathbf{x}_i,\mathbf{x}_j)) \\ &= \exp\left(-\rho\cdot\left(a_1\left\|\mathbf{x}_{i,1}-\mathbf{x}_{j,1}\right\|+a_2\left\|\mathbf{x}_{i,2}-\mathbf{x}_{j,2}\right\|+a_3\left\|\mathbf{x}_{i,3}-\mathbf{x}_{j,3}\right\|\right)\right) \\ &= K_{cld,1}\cdot K_{cld,2}\cdot K_{cld,3} \end{aligned} \tag{4.18}$$

Here $a_i, i=1:3$ are constants; terms above are of form (4.17). Furthermore (4.18) is a positive definite kernel as convolution of three PD kernels.

Positive Definite Kernel based on L1 distances

Kernels for CSD, EHD, and HTD are based on the distance metrics from (4.2) , (4.4), (4.5) which are versions of L_1 norm. These kernels K_{csd}, K_{ehd} and K_{htd} are induced with Corollary 4.7 and L_1 distance.

Lemma 4.8 : Let $K: X\times X\to\mathbb{R}$ be a symmetric function then

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\left|\mathbf{x}_i-\mathbf{x}_j\right|) \tag{4.19}$$

is a PD kernel. Here where $|\cdot|$ denotes the L_1 norm.

Proof: Let $K: X\times X\to\mathbb{R}$ there exist m vectors of dimension N, $\mathbf{x}_i\in\mathbb{R}^N, i=1,....,m$. An element of matrix K at position (i,j) is $K_{ij}=K(\mathbf{x}_i,\mathbf{x}_j)=-\left|\mathbf{x}_i-\mathbf{x}_j\right|$, negative value of the L_1 distance. Furthermore $K_{ij,r}=-\left\|x_{i,r}-x_{j,r}\right\|_2$ is Euclidian distance L_2 on scalar components $x_{i,r}, x_{j,r}$ for $r=1,...,N$. Where $x_{i,r}$ denotes the r-th coordinate of vector $\mathbf{x}_i$. It can be noticed that $K_{ij}=\sum_{r=1}^{N}K_{ij,r}$. Based on discussion in previous section $K_{L2}=-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|_2$ is a conditionally positive definite kernel. Therefore a sum of CPD kernels is a CPD kernel. To assure that the kernel matrix K is not singular, a strictly positive definite function as an exponent is used:

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\rho\left|\mathbf{x}_i-\mathbf{x}_j\right|)$$

Based on Theorem 4.5 an exponent of a CPD kernel is a PD kernel, hence so is the above kernel. For more details the reader is referred to Baxter (1991). ∎

Therefore K_{csd} and K_{htd} are PD kernels as forms of the kernel given in (4.19). For the case of EHD, the expression for d_{ehd} in (4.6) has an additional transformation on the input feature space into a global and semi-global histogram spaces. However this does not influence the kernel for Mercer's condition, only the input space. Therefore based on previous discussion K_{ehd} is also a symmetric PD kernel.

Positive Definite Kernel based on Quadratic distance

Finally, the DCD distance, leads to a non-trivial kernel based on quadratic similarity measure (Djordjevic and Izquierdo, 2006b) .

Proposition 4.9: The kernel induced by the quadratic distance d_{dcd} from (4.3) on the DCD feature space is PD.

Let's define a mapping from the DCD feature space into the N-dimensional feature space $\hat{X}$, for $N = \max_{i,j}(N(\mathbf{x}_i)\cdot N(\mathbf{x}_j))$. That is, each pair of features $\mathbf{x}_i, \mathbf{x}_j$ is mapped into a pair of vectors $\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j$ padded with zeros if $N(\mathbf{x}_i)\cdot N(\mathbf{x}_j) < N$:

$$\hat{\mathbf{x}}_i = [\underbrace{x_{i,1}\dots x_{i,1}}_{\times N(\mathbf{x}_j)}\dots\underbrace{x_{i,N(\mathbf{x}_i)}\dots x_{i,N(\mathbf{x}_i)}}_{\times N(\mathbf{x}_j)}],\ \hat{\mathbf{x}}_j = [\underbrace{\underbrace{x_{j,1}\dots x_{j,N(\mathbf{x}_j)}}_{\mathbf{x}_j}\dots x_{j,1}\dots x_{j,N(\mathbf{x}_j)}}_{\times N(\mathbf{x}_i)}].$$

Now, K_{dcd} can be written as:

$$K_{dcd}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = \exp(-(-\dot{K}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j))^{1/2}), \tag{10}$$

where $\dot{K}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) = -\left(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\right)^T A\left(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\right)$, A is the cross-similarity matrix between feature components of size $N \times N$ and its elements are cross-similarity coefficients $0 < a_{rs} \le 1$ from (4.3) .Based on Corollary 4.7 to prove that K_{dcd} is PD kernel, it is enough to show that $\dot{K}$ is negative and CPD. Clearly, $\dot{K}$ is negative since $\dot{K}(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j) < -\left\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\right\|^2 \le 0$. Moreover, it is straightforward to verify that:

$$\sum_{i,j=1}^{m} c_i c_j \dot{K}(\hat{\mathbf{x}}_i,\hat{\mathbf{x}}_j) = -\overbrace{\sum_{j=1}^{m} c_j}^{0} \sum_{i=1}^{m} c_i \hat{\mathbf{x}}_i^2 - \overbrace{\sum_{i=1}^{m} c_i}^{0} \sum_{j=1}^{m} c_j \hat{\mathbf{x}}_j^2 + 2\cdot a \left\langle \sum_{i=1}^{m} c_i \hat{\mathbf{x}}_i, \sum_{j=1}^{m} c_j \hat{\mathbf{x}}_j \right\rangle .$$

$$= 2\cdot a \cdot \left\| \sum_{i=1}^{m} c_i \hat{\mathbf{x}}_i \right\|_2^2 \geq 0.$$

Thus, $\dot{K}$ is CPD and K_{dcd} PD kernel. ∎

Positive Definite Kernel based on Histogram Intersection distance

In case of HSV descriptor, it is necessary to prove that the following kernel K_{hsv} is PD:

$$K_{hsv}(\mathbf{x}_i,\mathbf{x}_j) = \exp(-\rho \cdot d_{hsv}(\mathbf{x}_i,\mathbf{x}_j)) = \exp(-\rho \cdot (1 - \textstyle\sum_i \min(\mathbf{x}_i,\mathbf{x}_j))) \qquad (4.20)$$

Shawe-Taylor and Cristianini (2004) defined an intersection kernel as a kernel on a set of subsets of a measurable set $\mathcal{X}$ (Appendix B, Definition B.13-B.16). In a finite case like in the one of HSV descriptor, the measure is a mapping to positive real numbers whilst in the case $\mathcal{X}$ is infinite this assumes measurable sets and σ-algebra. Integration over the set $\mathcal{X}$ is defined for a measurable function f as (Royden, 1988):

$$\mu(f) = \int_{\mathcal{X}} f(x)\,d\mu(x)$$

For indicator function I_X on a measurable set X, which is $I_X(x) = 1$ for $x \in X$ and 0 otherwise, the measure is define as:

$$\mu(X) = \mu(I_X) \in \mathbb{R}^+$$

Definition 4.10: Intersection kernel can be defined on a subset of measurable set $\mathcal{X}$, as a measure of the intersection of two sets $X_1, X_2 \in \mathcal{X}$: $K_I(X_1,X_2) = \mu(X_1 \cap X_2)$.

The feature space of all measurable functions has the inner product defined as:

$$\langle f_1, f_2 \rangle = \int_{\mathcal{X}} f_1(x) f_2(x)\,d\mu(x).$$

In this non-vector space the feature mapping is $\Phi: X \to I_X$. Since $I_{X_1 \cap X_2} = I_{X_1} \cdot I_{X_2}$:

$$K_I(X_1,X_2) = \mu(X_1 \cap X_2) = \int_{\mathcal{X}} I_{X_1 \cap X_2}(x)\,d\mu(x) = \int_{\mathcal{X}} I_{X_1}(x) I_{X_2}(x)\,d\mu(x) = \langle I_{X_1}, I_{X_2} \rangle$$
$$= \langle \Phi(X_1), \Phi(X_2) \rangle$$

Hence the intersection kernel is a valid kernel.

In case of real numbers, $\mathbf{x}_i$ and $\mathbf{x}_j$ represent sets in ranges $[0, x_i]$ and $[0, x_j]$, (e.g., histogram bins). The intersection kernel with a standard integration measure is a PD kernel with normalization:

$$K_{hi}(\mathbf{x}_i, \mathbf{x}_j) = \min(\mathbf{x}_i, \mathbf{x}_j)$$

Since in (4.20) the normalization is done beforehand on the distance function level, the following kernel for the histogram intersection can be obtained:

$$K_{hsv}(\mathbf{x}_i, \mathbf{x}_j) = const. \cdot \exp(\rho \cdot \sum_i \min(\mathbf{x}_i, \mathbf{x}_j)) = const. \cdot \exp(\rho \cdot K_{hi}(\mathbf{x}_i, \mathbf{x}_j)) \qquad (4.21)$$

Since as mentioned exponential functions can be approximated by polynomials with positive coefficients and polynomials are well-defined class of kernels. Considering the class of kernels is closed under point wise limits, (4.21) being a limit on positive definite polynomial kernels is a positive definite kernel itself. In a alternative approach Boughorbel et al. (2005), defined a generalized histogram intersection considering directly real numbers.

This proposition along with the previous analysis and discussions enables the use of the proposed kernel to drive a SVM based learning approach. In this case, convergence and uniqueness of the underlying optimisation problem is guaranteed.

4.2. Evaluation Results

In our specific problem the subcomponent dimensionalities are dependant on nature of the feature spaces from Chapter 2. For a specific combination of descriptors the dimensionalities are given in Table 4.1.

Table 4.1: Dimensions of individual feature spaces

Descriptor	CLD	CSD	DCD	HSV	EHD	HTD	GLCM
Dimension	58	32	variable	62	80	32	4

The experimental setup is similar to the one in previous chapter, however the ACK kernel and individual kernels from (4.14) are included into the comparison, together with classical Gaussian kernel (RFBK) and Laplace kernel (LK).

For RBFK and LK the descriptors are either individual descriptors or normalized concatenations CLEH, CSEH, as well as CONC. For the individual kernels based on distances (denoted as KCLD, KCSD, KDCD, KHSV, KEHD, KHTD, KGLCM) appropriate single descriptors were used. The proposed ACK kernel is defined on structured spaces of all seven descriptors with weights based on relevant and irrelevant training samples of the supervised user feedback.

The initial training set has 6 training examples 3 labelled as relevant and 3 as irrelevant, in each iterations appropriately labelled samples, from the retrieved user interface are added (10 per iteration). For each kernel three independent random runs were performed and the averaged results used in the analyses. Each relevance feedback session is followed for 4 iterations and the precision measured within a window of size equal to the class size.

Comparative results between the kernels described in this chapter are also considered. That is comparison of precision in individual kernels based on distances and the ACK kernel. The retrieved set of images always equals size of a certain class (category). This results were presented on different databases see Figure 4.1-Figure 4.5 (the image on the left). In the same set of figures (the image on the right) comparative results for the two best performing kernels for kernels based on individual distances and ACK are given together with the best performing combinations kernel-descriptor for RBFK, LK

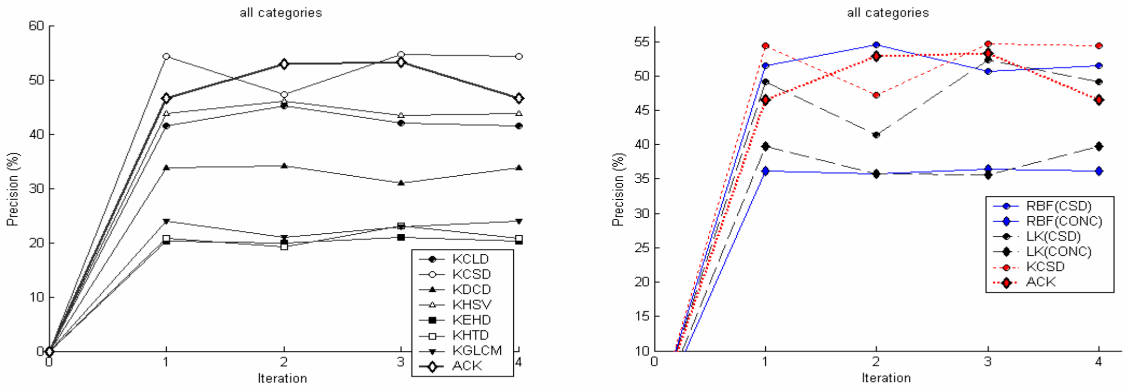


Figure 4.1: Average precision over iterations for the individual distance kernels and ACK (on the left). Comparison of two best performing cases for each type of kernel (on the right). DColour database.

For the DColour database with classes based on low-level visual features, in Figure 4.1 it can be seen that the ACK manages to outperform most of the kernels based on individual distances, even kernels based on colour. This is due to the adaptive nature of the kernel and the ability to select relevant features on-line. A similar high performance is also obtained when comparing with best results for RBFK and LK (image on the right).

.

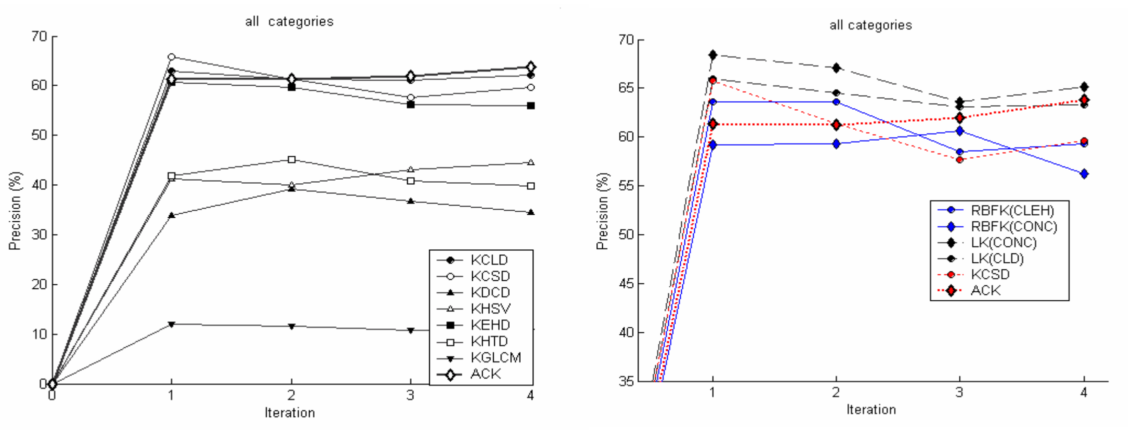


Figure 4.2: Average precision over iterations for the individual distance kernels and ACK (on the left). Comparison of two best performing cases for each type of kernel (on the right).D8 image database.

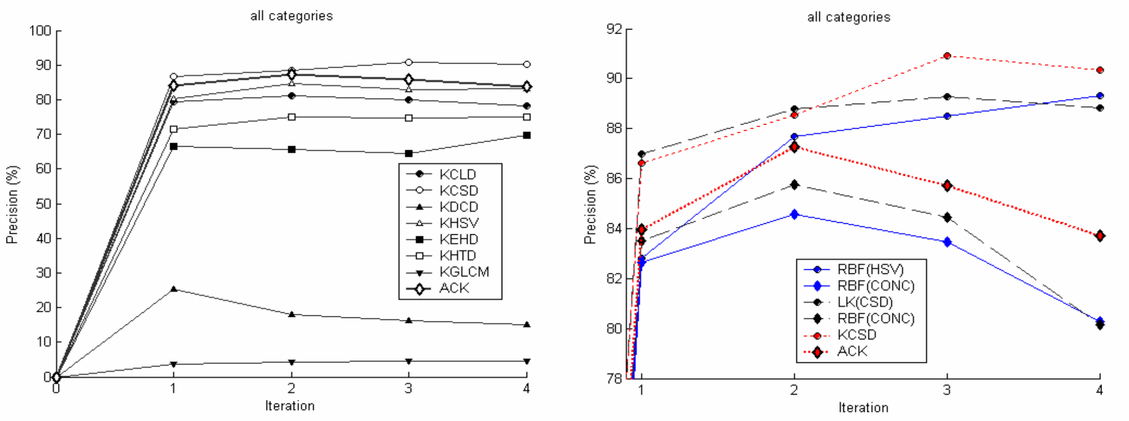


Figure 4.3: Average precision over iterations for the individual distance kernels and ACK (on the left). Comparison of two best performing cases for each type of kernels (on the right). D25-1800 image database.

For the two object based databases with homogeneous background Figure 4.2-Figure 4.3. The ACK kernel gives high performances when comparing to individual distance kernels (images on the left). However when comparing to other classical kernels it is performing averagely. This led to a conclusion that the weighting scheme in ACK kernels actually gives more weights to the background colour and hence the results are worse.

So even though the ACK adapts to the data it is modelled for natural images not object based categories with artificial uniform background.

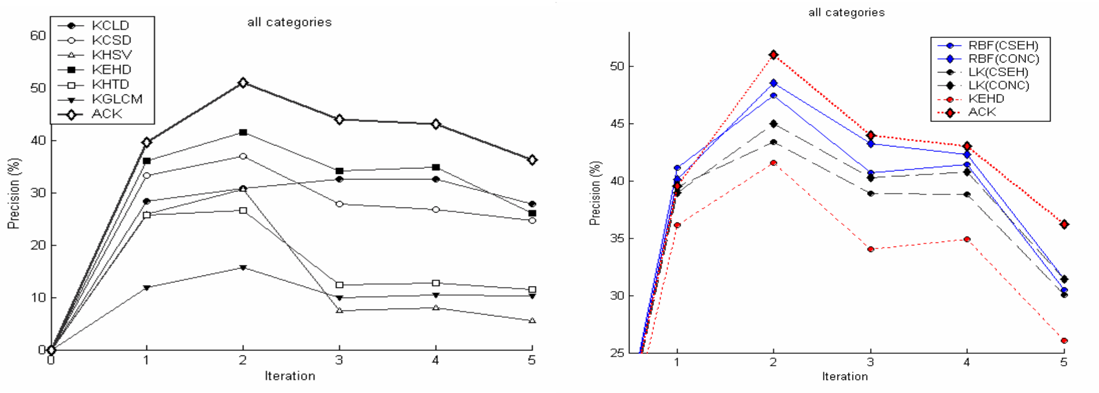


Figure 4.4: Average precision over iterations for the individual distance kernels and ACK (on the left). Comparison of two best performing cases for each type of kernel (on the right). D7-700 image database.

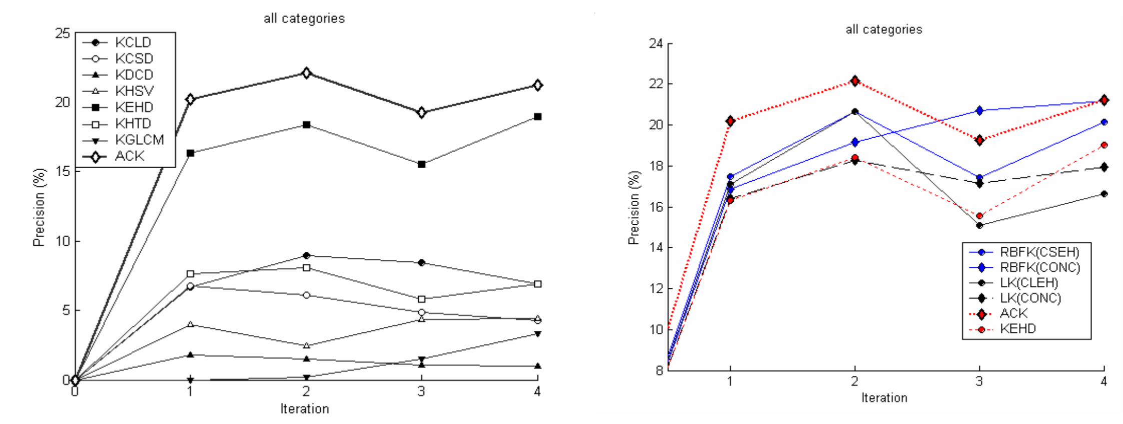


Figure 4.5: Average precision over iterations for the individual distance kernels and ACK (on the left). Comparison of two best performing cases for each type of kernel (on the right). Caltech 101 image database.

In natural images with diverse background as in D7-700 and Caltech 101 image database, the ACK kernel performs better than individual distance kernels as expected (image on the left) Figure 4.4- Figure 4.5. And at the same time gives highest precision comparing to other kernels (image on the right) adapting itself as more training data becomes available.

In the next paragraph, as in the previous chapter precision-recall curves are presented for the D7-700 image database. The precision values are again averaged over same relative scopes. The aim is to show how more training information influences the results of SVMs through iterations.

Average precision-recall curves between the kernels described in this chapter are considered. That is comparison of precision-recall values is given, between individual kernels based on distances and the ACK kernel. In Figure 4.6, in all of the curves there is a very clear and distinctive improvement when using the ACK kernel The precision-recall curves at bottom right in Figure 4.6 present precision-recall over different iterations for the best performing ACK kernel. This curves show a conclusive improvement, that is more training samples lead to improved quality of results, since the precision-recall values for the fifth iteration with the higher number of training samples show best performances. This is a direct consequence of adaptive nature of the ACK with adjusts the kernel , and weights differently each feature based on intra class variance for relevant and irrelevant samples.

Adaptive convolution kernel that deals with multi-feature spaces and guarantees convergence of the SVM optimisation problem has been introduced. A feature space as a structure of individual visual feature spaces with different distances was considered. The ACK kernel follows the nature of the patterns used and learns over time a better representation of similarity for the structured data feature space. For a particular combination of distances it has been proved that the ACK kernel is positive definite with high performance values. Furthermore this approach gives possibility of defining new convolution kernels and combining features very different in nature as visual, audio and motion descriptors and therefore effective classification of multimedia content.

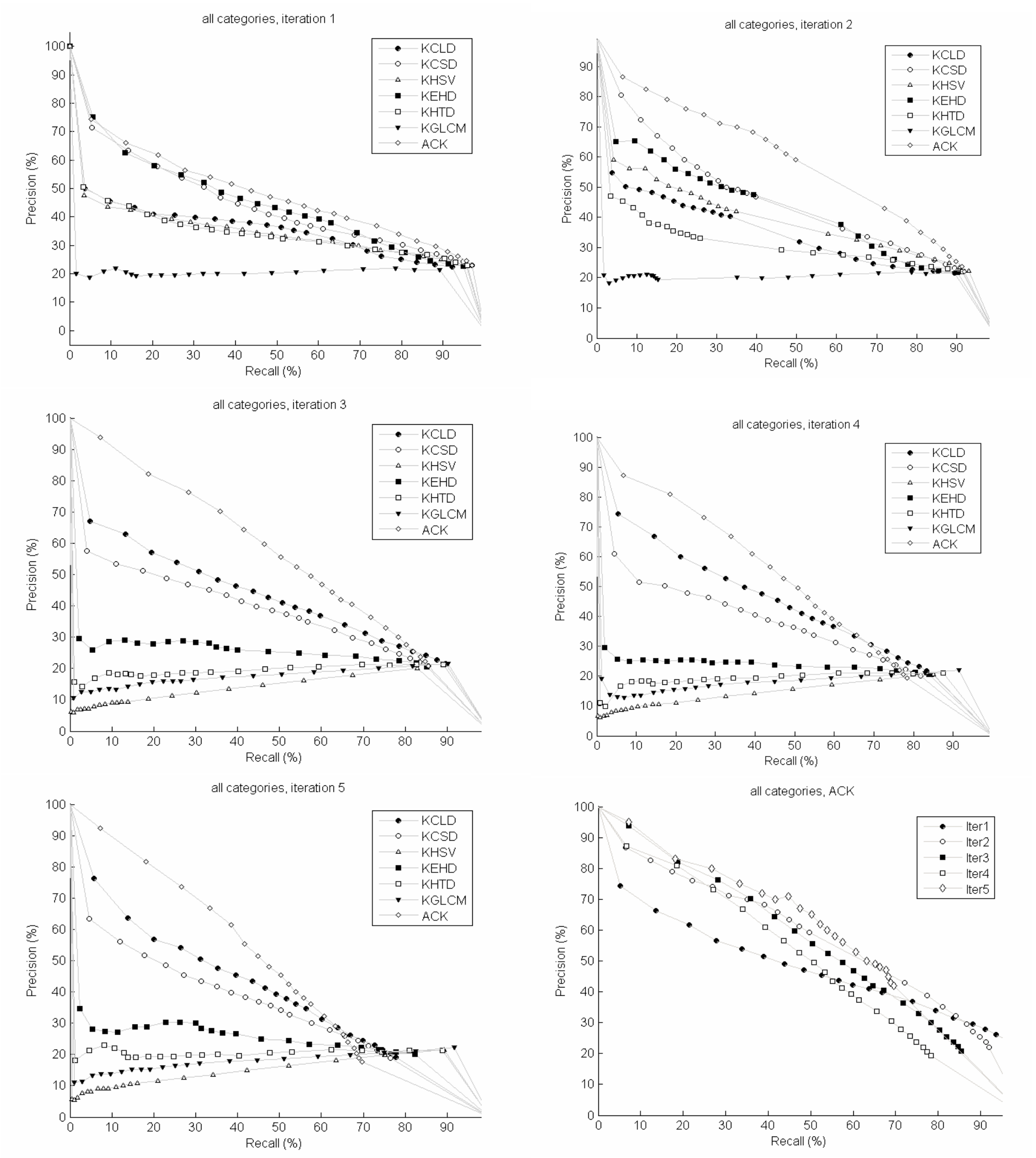


Figure 4.6: Average precision-recall curves for kernels based on individual distances and for the proposed kernel ACK in the given structured multi-feature spaces. (top to bottom, left to right). Comparative results through iterations for the best performing ACK kernel are given in bottom right corner.

CHAPTER 5 : An Object- Driven System with Relevance Feedback and Kernels on Sets

5.1. Introduction

In this chapter user's interest in single semantic objects is considered. Bearing in mind that object segmentation is arguably as hard as the semantic gap problem, a block-based structure is introduced to label single objects in images. The reason for using this strategy came out of conclusions of our previous thorough empirical analysis of retrieval systems in a relevance feedback scenario (Dorado et al. 2006). It was noticed that labelling complete images as relevant to a given key-word introduces a lot of noise due to the variety of non relevant objects in complex scenes. The same aspect is studied in few other approaches from the literature. Mostly interest points and models composed of local characteristics of image parts and spatial relations among them are exploited. For instance, in Fergus et al. (2003) global object models are learned based on scale-invariant image regions. In Agarwal and Roth (2002) a vocabulary of image parts is used together with spatial relation among the parts. Another interesting approach, proposed by Csurka et al. (2004), detects interest points and patches around them, then these points are clustered to create a fixed length histogram feature vector. However most of these methods are parametric and assume the input data can be faithfully modelled by some probability distribution.

Though the advantage of using interest points in an object recognition scenario is apparent, there is no guarantee that in a retrieval scenario for natural images the points and features describing the local regions around the points will be representative enough. Natural image databases usually do not contain different views of the same object but a variety of pictures of the same conceptual object, e.g. tigers; with high variance in visual appearance, size, occlusion and posture. Conventional computer vision approaches for object recognition are likely to fail in such scenarios.

Contrasting this, the structured feature space proposed in this chapter consists of several low-level feature representations of texture, colour and edges for small image blocks. This space incorporates both the low-level similarity in the multi-feature space described before and spatial correlation among neighbouring blocks. The idea is to combine the low-level similarity of features and spatial relations of image parts into a semi-local, semi-global feature representation level and to devise a kernel that could handle this data (Djordjevic and Izquierdo, 2006d).

5.2. Overview of Kernels on Sets

In this section an overview of approaches for kernels with vector sets is given. The use of localized features and discriminative learning approached has recently given rise to a new class of methods that model the kernel to deal with not only input vectors of equal dimension, with each vector corresponding to a particular global feature, but also sets of interest points descriptors of different cardinality. Furthermore since several types of features are collected together they all need to be fused in a kernel. Several approaches have been recently proposed when dealing with kernels on sets. Observe that using kernels on sets enables the handling of sets of descriptors with different dimensions. Most relevant approaches form the literature use kernels on sets over local low-level features. A kernel that derives an average similarity of the best matching local primitives in two features sets was introduced by Wallraven et al. (2003). However, the underlying maximum operator proposed in leads to a non-Mercer kernel. Thus, convergence and uniqueness of the solution of the optimization problem underpinning the used SVM classifier is not guaranteed.

Similarly a kernel in Lyu (2005) uses a power of all possible local feature matches in sets; Though with good performances this kernel is computationally very expensive.

The authors in Boughhorbel et al. (2004) proved that statistical positives of a non-Mercer kernel can be guaranteed with high probability by controlling the hyperparameters in SVMs, however this does not stand for the general case.

Another interesting approach for kernels on sets deals with principal angle based similarity kernel between two linear subspaces spanned by mapping from real to Hilbert space (Wolf and Shashua, 2003). However in this case the kernel is only positive-definite for sets of equal cardinality.

In Shashua and Hazan (2005) a family of algebraic kernels was used to combine similarities given by vector-based kernels. In this case various weighs indicate the level of alignment between feature parts.

Several methods design kernels based on probabilistic models of inputs. Instead of defining a kernel directly between input sets of different cardinality, the inputs are regarded as independent and identically distributed samples from unknown distributions from a same parametric family. The vector set kernel is defined as a kernel between these distributions. For instance, the Bhattacharyya kernel uses multivariate Gaussian distributions (Kondor and Jebara, 2003), while KL-divergence is used in Moreno et al. (2003). In Grauman and Darrell (2005) an explicit histogram pyramid was formed on sets, and used for partial matching with hierarchical weighted histogram intersection as similarity measure.

However all of these approaches have limitations in complexity, parametric representation, positive definiteness and they are mainly defined for interest points like SIFT or jets (Lowe, 1999; Schmid and Mohr, 1997; Mikolajczyk and Schmid, 2005) or specific histogram representations (Grauman and Darrell, 2005).

Hence in this chapter low-level similarity of features and spatial relations of image parts are combined into a structure and localized feature space and then a kernel on sets that could accommodate this data is considered.

5.3. System Overview and Feature Spaces

To simulate human visual perception several primitives or low-level features extracted from image content need to be considered. The aim is to obtain information from different low-level visual cues at various levels of complexity and to jointly exploit that information to infer higher levels of conceptual abstraction. Low-level descriptors are very useful to search for patterns of interest and similarities in image database. However, if the aim is to retrieve audiovisual content using semantic structures, e.g., key-words, which are natural to humans, three profound challenges become evident: how to merge different low-level content features into meaningful descriptors with high semantic discrimination power (as discussed in Chapter 4); how to deal with the subjective interpretation of images by different users under different conditions (the relevance feedback); and how to recognize single semantic objects in complex images.

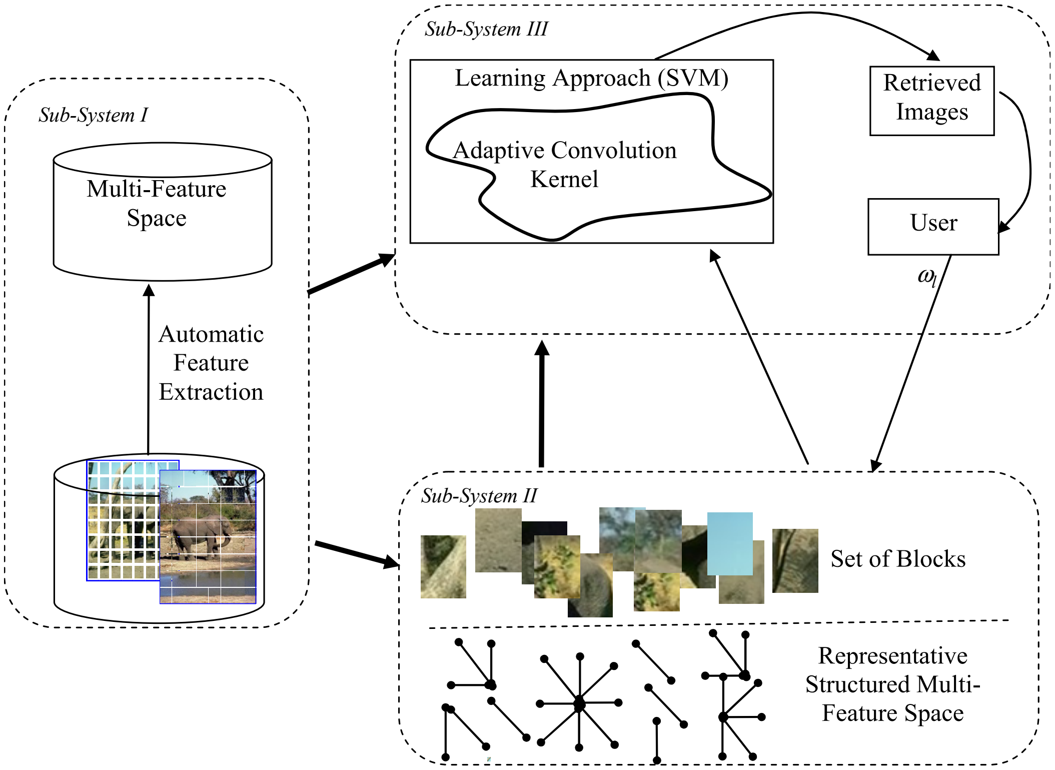


Figure 5.1: Framework architecture.

The proposed system, as outlined in Figure 5.1, consists of three main sub-systems. Each one of these sub-systems relates to one of the three challenges described previously: merging different low-level content descriptions; object based model; and user relevance feedback. The first sub-system runs offline and embraces three processing steps. Initially all images in the database are split into blocks of small size. The aim of this processing step is to handle single semantic objects as mosaics of elementary blocks. In the second step several low-level features are extracted automatically for all image blocks. Now the problem is how to merge these features into a single multi-feature descriptor and this has been addressed in Chapter 4. This step leads to the "multi-feature space" highlighted in Figure 5.1.

The second sub-system initially runs on-offline and addresses the challenge of effectively capturing single objects in natural scenes. This challenge is tackled by generating a structured block-based representation of each image. Image blocks are clustered together based on their joint low-level similarity and spatial proximity to each other. This step generates a number of representative structures per image as illustrated in Figure 5.1. However it is also used online, together with the third sub-system, in order to capture common structures in user selected relevant images and to exclude any structures that are common to the irrelevant images.

The third sub-system involves online interaction with the user and comprises a number of processing steps. The interaction is initialized by retrieving few previously annotated pictures related to a given semantic concept. The user marks the retrieved pictures as relevant or irrelevant. This information is used as the supervised input to the SVM based learning approach. The method uses the structured descriptor space generated in the offline stage of the second sub-system. It classifies images by matching structures using the proposed nonlinear multi-feature kernel. This kernel exploits the user input to weight dynamically each feature space accordingly. In each iteration the structuring sub-system is used to capture relevant structures across images labelled as relevant by the user, and to exclude those clustered among irrelevant images (e.g. homogeneous background). The weights assigned to the different feature-spaces are denoted by ω_l in Figure 5.1. Finally, the system outputs relevant images to a given semantic concept or keyword. This information can be used in a second iteration of user relevance feedback to improve the retrieval performance with respect to the semantic concept of concern.

After several online iterations with the user in the loop, the system outputs all relevant images and the learned weights defining the underlying multi-feature space for a specific semantic concept. At the same time, the semantic concept or key-word of concern is propagated through all retrieved relevant images in the database.

5.3.1 Low-level Feature Selection

The multi-feature space for images is defined as a structured data space as mentioned in Chapter 4 out of the following low-level descriptors: CLD, CSD, DCD, EHD, HSV histogram , Gabor Filters Feature (GFF) and GLCM. Note that instated of HTD from Chapter 4 a new descriptor is used, this is Gabor feature filter (GFF) . Both HTD and GFF are based on Gabor filters while HTD is standardized in number of scales and direction, these parameters can be freely adapted in GFF. A disadvantage of HTD is that it deals with images of size equal or larger to 128x128 pixels. And since the idea of this chapter is to incorporate smaller blocks of images into a mosaic structure, the GFF descriptor is more suitable. The GFF is used to extract localized texture information on a number of directions and scales. The mean value and standard variation of Gabor filtered image coefficients are used to construct the feature vector (2.1).The distance values between two feature is L_1 distance of normalized first and second moments of Gabor coefficients (2.5). Hence the positive definiteness discussed for HTD (Chapter 4) also stands for the appropriate exponential distance kernel built for GFF.

5.4. Object-based Structured Descriptor Space

Most annotation and retrieval approaches from the literature have dealt with either whole images or (not semantically meaningful) regions segmented according to colour or texture similarity. If segmentation is used the presence of noisy regions and oversegmentations is unavoidable. This leads to confusing and inadequate retrieval results. Consequently, on the one hand we need to consider the fact that even the best image segmentation techniques cannot extract meaningful semantic objects, hence it is not reasonable to assume object segmentation in a retrieval system. On the other hand, without segmentation learning techniques may fail to capture a specific object the user is interested in (Chapter 4, Figure 4.2 and Figure 4.3). However, semantic objects can be regarded as mosaics of small building blocks. In most cases these building blocks do not encapsulate the whole semantic concept and they can be regarded as being closer to low-level than to high level descriptions (Izquierdo and Djordjevic, 2006).

In this section a method to organize individual image blocks into meaningful structured data conveying spatial information is presented. The aim is to keep the descriptor resolution at block level, local to objects rather than whole images, while at the same time capturing higher-level semantic information. To do so, first elementary non-overlapping structures covering the whole picture are generated. Once this has been achieved key-representative blocks within each image are extracted using k-medoids clustering method (Kaufman and Rousseeuw, 1987). Meaningful clustering is achieved using the distance (4.8) as well as the spatial proximity of the generated structures.

Let $\{I_1, I_2, ..., I_n\}$ be the images in the database of concern. Each image is partitioned into a grid of $r \times s$ blocks. Let B_{ij}, $i = 1,...,r$, $j = 1,..,s$ be the set of image blocks for a given image. Here B_{ij} stands for the block at position (i, j). The similarity between two blocks is estimated according to (4.8). Next, each block in an image is assigned to a group made up of 3x3 neighbourhoods. These 3x3 blocks neighbourhoods are called (regular) structures. For border areas the structures are adapted not to cross over the bounds and hence can be smaller then 3x3. These mosaics are the initial structures denoted as S in Figure 5.2 a).

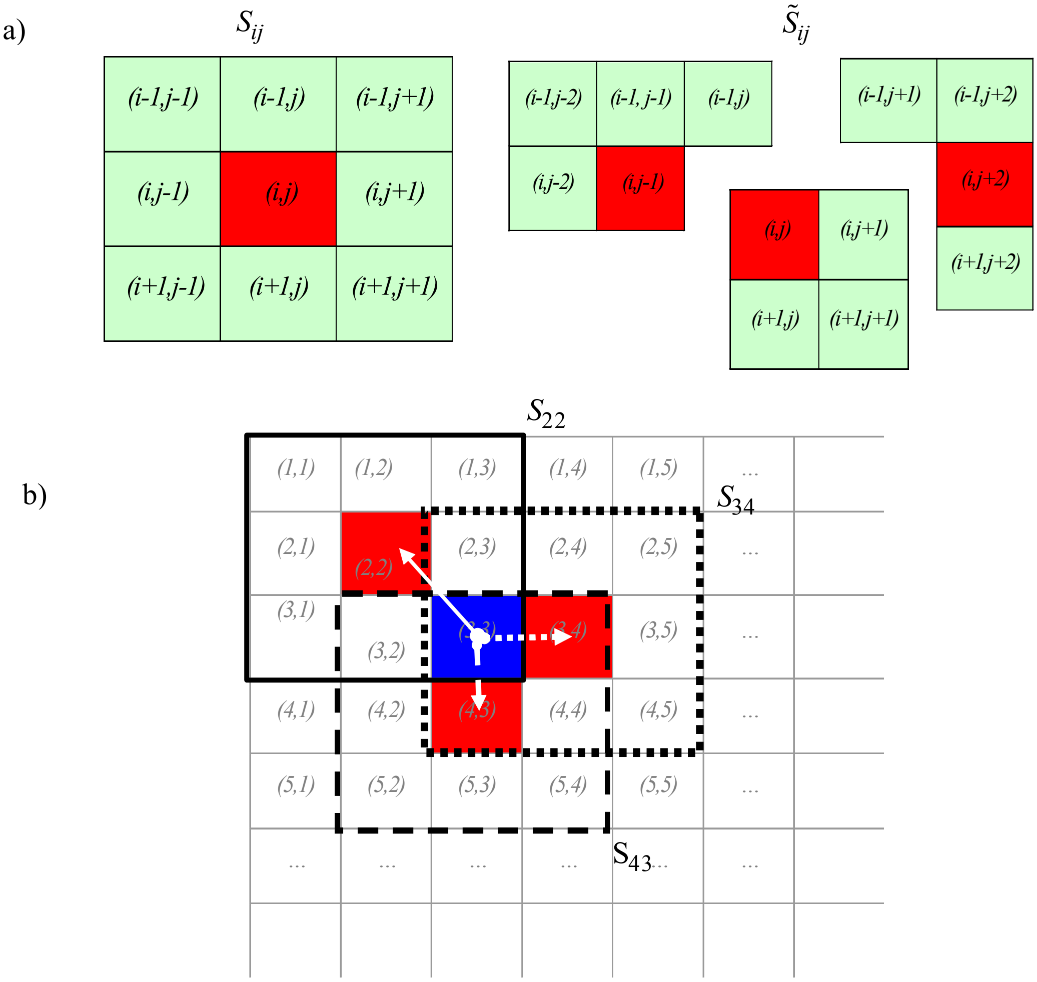


Figure 5.2: a) The 3x3 neighbour used to build initial regular structures and its breakdown into non-regular, non-overlapping structures b) A single block (blue colour) as a member of several neighbouring overlapping structures. White arrows correspond to low-level similarity between central blocks of each structure (in red). The blue block remains a member of a structure to which it is most similar.

The distance between block B_{ij} at the centre of each structure S_{ij} and any other blocks B is defined as "the distance between S_{ij} and B". The distance between S_{ij} and B is also estimated according to (4.8). The distance between two structures is defined as the distance between their central blocks. Observe, that the initial regular structures overlap each other. In the subsequent processing step non-overlapping structures are built according to the following condition, (see Figure 5.2 b):

If a block B is a member of two or more neighbouring structures then it is removed from all structures but the one closest to itself.

Using this condition the number of member blocks of each structure is reduced. And refined subset of structures $\tilde{S}_{ij}$ is formed. Observe that $\tilde{S}_{ij} \subseteq S_{ij}$ for some but not all i, j. The resulting refined set of structures is non-overlapping and of irregular shape as shown in Figure 5.3 (left).

In order to generate a set of key representative descriptors, k-medoid clustering is performed on the obtained structures by incorporating spatial information. Though k-means is computationally more efficient than the k-medoid, it requires a well-defined vector space. Unfortunately, the multi-feature space considered in this work is not even a metric space (as discussed Chapter 4), hence the k-means clustering technique cannot be applied. Observe, that k-medoid clustering can be used over any feature space endowed with a similarity function. The aim is to cluster together structures $\tilde{S}_{ij}$ using the similarity function (4.8) and spatial information about $\tilde{S}_{ij}$. The similarity function for the clustering algorithm is defined as:

$$\tilde{D}(\tilde{S}_{pq}, \tilde{S}_{ij}) = D(\tilde{S}_{pq}, \tilde{S}_{ij}) \cdot \Gamma(\tilde{S}_{pq}, \tilde{S}_{ij}), \tag{5.1}$$

for $\Gamma(\tilde{S}_{pq}, \tilde{S}_{ij}) = \min_{B_{vx} \in \tilde{S}_{pq}, B_{yz} \in \tilde{S}_{ij}} \gamma(B_{vx}, B_{yz}) + 1$, where γ is the Chebyshev distance over the positions of the block contained in the two structures. That is, $\gamma(B_{vx}, B_{yz}) = \max\left(|v-y|, |x-z|\right)$. Observe that the similarity measure $\Gamma(\tilde{S}_{pq}, \tilde{S}_{ij})$ weights (4.8) using the actual spatial distance between the two structures Figure 5.4. Armed with (5.1), and for the sake of completeness, the k-medoid clustering algorithm is explained in the following paragraph and depicted in Figure 5.3 (right).

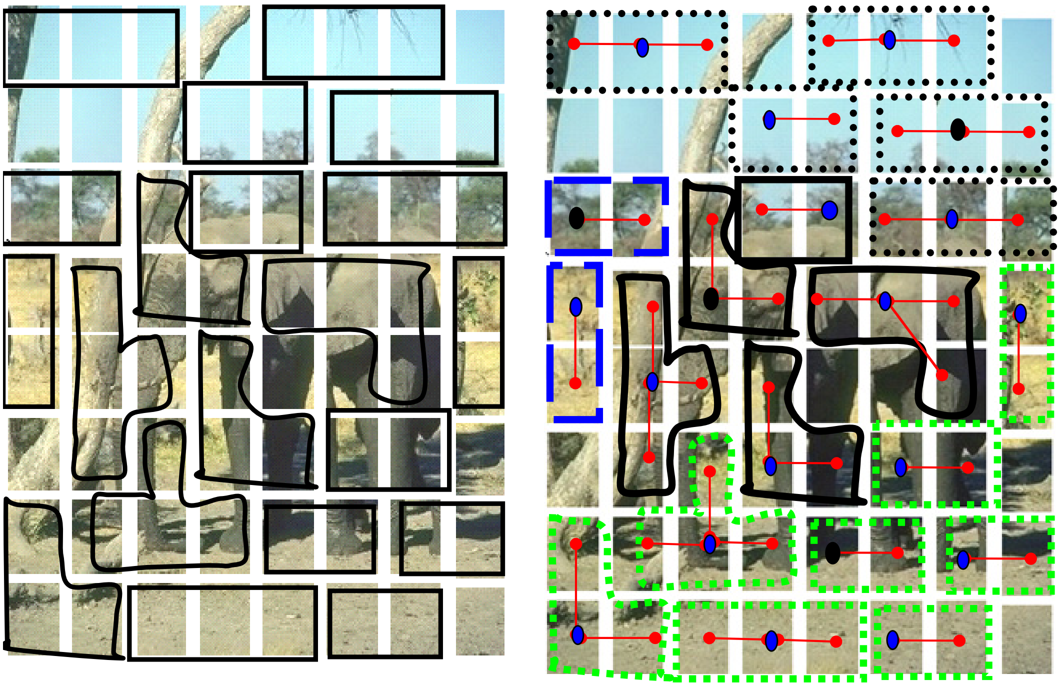


Figure 5.3: Mosaics of structured elements before and after clustering using low-level similarity and spatial proximity of structured elements. The elements with blue dots represent structure centres and the four elements with black dots representative centres of the four representative clusters.

Initially cluster medoids (prototypes) C_k^0, $k=1,\dots,K$ are randomly selected among all structures $\tilde{S}_{ij}$. Using these K medoids an iterative process is started. At a given iteration t the set of medoids is defined as $C_k^t=\{\tilde{S}_{pq}\}, p\in\{1,\dots,r\}, q\in\{1,\dots,s\}$. The membership of each element to a cluster is obtained by finding the minimum distance (5.1) from an element to cluster medoid. The new updated set of medoids is generated so that the sum of the distances between a new medoid and member elements of that cluster is minimal:

$$\min_{p,q}\sum_i\sum_j \tilde{D}(\tilde{S}_{pq},\tilde{S}_{ij}). \tag{5.2}$$

Then membership of each cluster element is updated and if any of the memberships changed a re-allocation of cluster medoids is performed and the iterations continue.

The previously described clustering process leads to a new descriptor with predefined number of representative elements. These descriptors have local block level discrimination power and convey spatial contextual information, Figure 5.3.

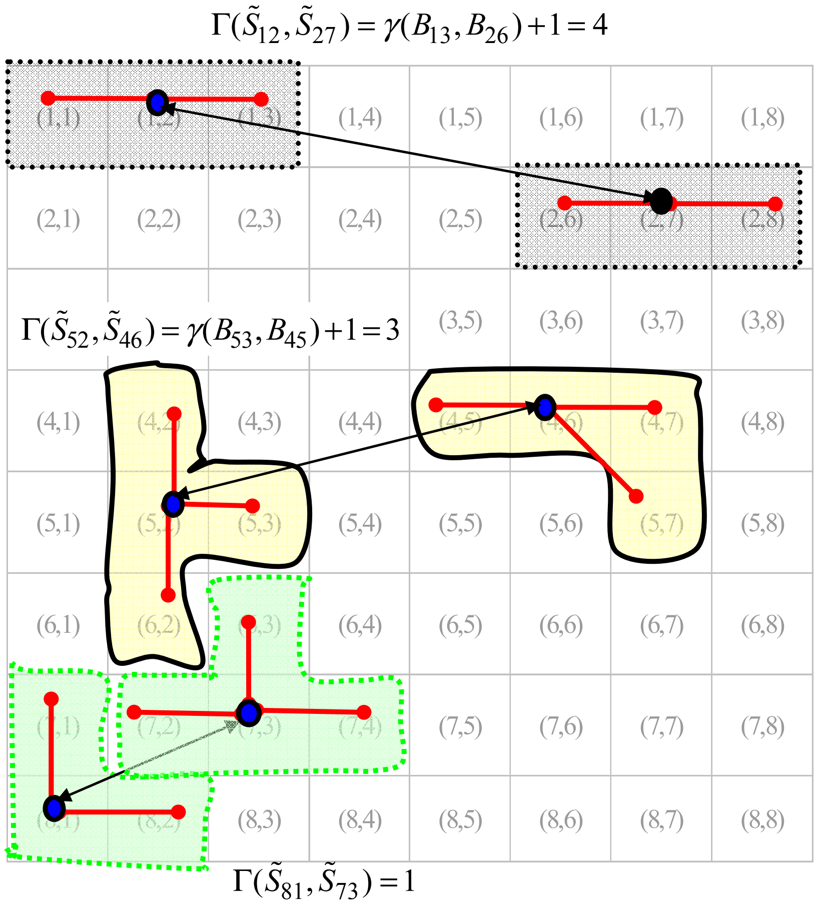


Figure 5.4: Spatial proximity between non-overlapping structures.

5.5. Kernel on Sets for Structured Feature Space

Following the previous process to generate arbitrarily shaped structures (potentially representing objects in the image), the similarity between such structures can be measured using the low-level descriptors and ACK (Chapter 4). However, the results obtained when individual structures were used with ACK kernel, were not as expected. The main reason for the observed moderate performance is that fact that the estimated structures are not always good representations of semantic objects. Indeed, in most cases the obtained structures are parts of object and more than one structure is needed to roughly represent an object, Figure 5.5. Considering this fact a different approach was taken to retrieve images containing similar objects.

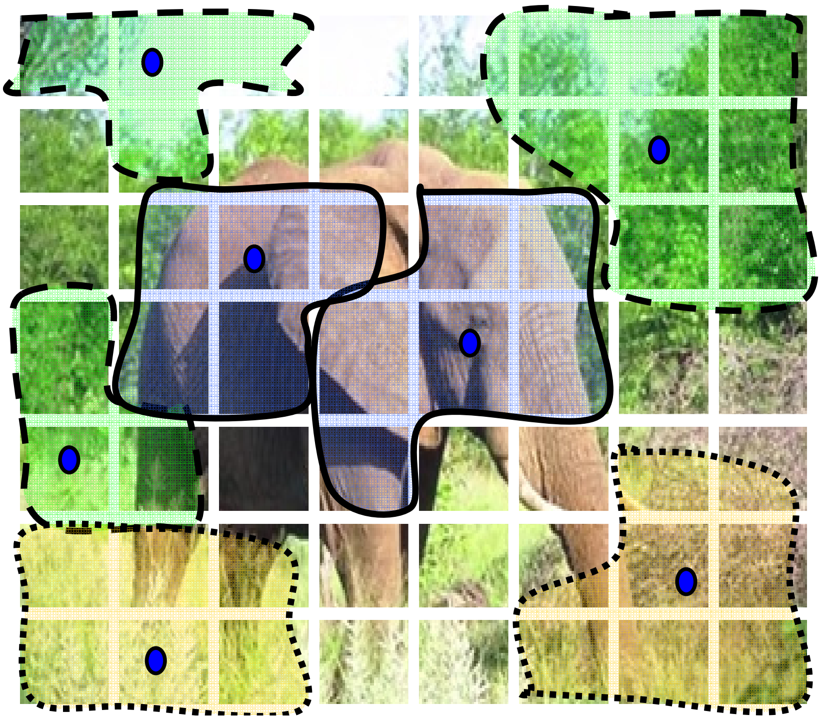


Figure 5.5: Example of clustered structures where the obtained structures are parts of object and more than one structure is needed to roughly represent an object.

In each iteration the obtained structures from the previous step of forming a structured space are further used to capture relevant structures across images labelled as relevant by the user, and to exclude those clustered among irrelevant images Figure 5.6.

In an iteration two sets are obtained, a set of relevant and irrelevant user labelled images. Since we have representative structures for each image, the process of obtaining even better key representations can be further refined by obtaining only structures that are common to all relevant images. The most similar structure across images can be obtained and further on ordered by decreasing levels of coherency.

In Figure 5.6 it is depicted how a structure is chosen to be a part or not of that particular "across image cluster". For each cluster in a set of relevant images we find the most similar structure from a new image, where the similarity is obtained as similarity to all elements that are already members of that function, and based on similarity measure (4.8). In Figure 5.6 a) the cluster is formed across images labelled as 1, 2 and 3. Now the choice of a structure from image 4 needs to be made. The structure from image 4 that has the smallest distance from all members of that cluster will be chosen as the new member. This allows for most similar structures to be clustered together. Hence in this way a set of most coherent clusters is obtained for each relevant and irrelevant set, and this clusters can now be ordered in decreasing levels of similarity Figure 5.6 b) and Figure 5.6 c).

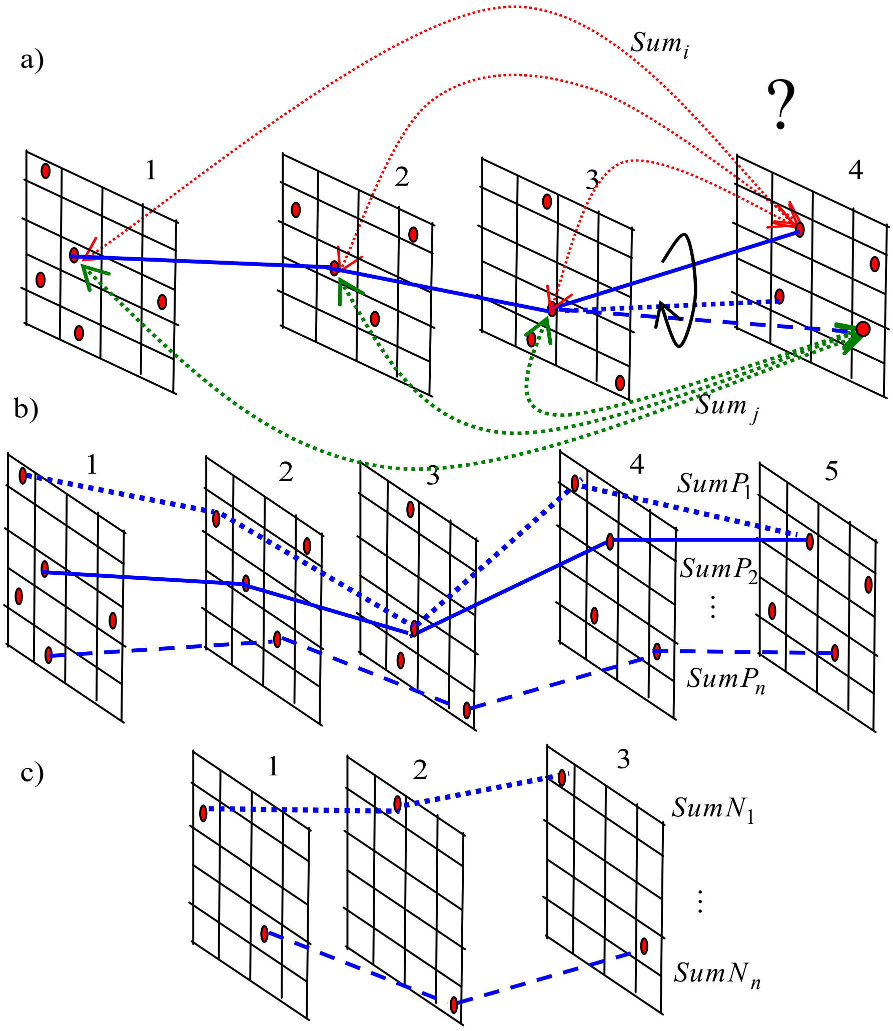


Figure 5.6:a) Clustering image structures across images b) clusters of representative structures across images labelled as relevant, ordered by decreasing cluster coherency, c) clusters among images labelled as irrelevant, ordered by decreasing cluster coherency.

The reasoning being, that if irrelevant images have very coherent clusters which are similar to clusters across relevant images, then all the structures that are members of this cluster need to be removed from the set of representative structures for each relevant image (e.g. coherent background across the database). As well as from the set of representative structures for negative images since they have low-level commonality not connected to the concept in question. When considering similarity between two "across mage clusters" the similarity is obtained between their medoids by using low-level similarity (4.8). The medoids are closest elements to all the other elements in that particular cluster. Hence, the obtained set of representative structures per relevant and irrelevant images is further refined, removing background noise and capturing similarity across images. Therefore this sets have variable cardinality of individual descriptors for each key representative.

Given two images and their decomposition into structures, the kernel to measure dissimilarity is defined as the sum of dissimilarities estimated using all possible combinations of structure pairs between the two images, "summation" kernel. The reason is that now the contributions of structures building a single object are added together and the actual object representation is better considered in the underlying measure. This kernel need: to satisfy the Mercer condition; to be computationally effective; and to be able to handle inputs of variable length.

Let a set of local features for the i-th image be $\boldsymbol{\mathcal{X}}_i = \{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_{N(\boldsymbol{\mathcal{X}})}^{(i)}\}$, $i = 1,...,m$, where $N(\boldsymbol{\mathcal{X}}_i)$ is the cardinality of the i-th image set and $\mathbf{x}_k^{(i)}$, $k = 1,...,N(\boldsymbol{\mathcal{X}}_i)$ is a multi-feature vector for key-representatives in an image. The dissimilarity of a pair of local vectors can be represented with a local Mercer's kernel $K_L(\mathbf{x}_k^{(i)}, \mathbf{x}_h^{(j)})$, in this case the ACK kernel from (4.14). Hence the summation kernel is PD kernel as a sum of PD kernels:

$$K(\boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j) = \sum_{k=1}^{N(\boldsymbol{\mathcal{X}}_i)} \sum_{h=1}^{N(\boldsymbol{\mathcal{X}}_j)} K_L(\mathbf{x}_k^{(i)}, \mathbf{x}_h^{(j)}) . \tag{5.3}$$

This kernels leads to much better results than in the case whole images (Chapter 4), results as reported in the next section.

5.6. Experimental results

For most categories there is a huge variance in low-level features over different images as presented in Chapter 4. Even though the results were averaged over a number of independent runs, the mentioned variance leads to inconstancy that inspired development of set kernels on structured feature spaces (5.3). According to the framework presented in Figure 5.1 the set kernels in a structured descriptor space, are experimentally evaluated. Performances for the ACK kernel from (4.14). with complete images and for set kernels with local ACK kernels (denoted as SET (ACK)) from (5.3).

were evaluated on all four databases with clear semantic concepts : D8 image database, D25-1800 image database, D7-700 image database and Caltech 101 image database. The same experimental setup as given in Chapter 4, was used for these experiments.

In Figure 5.7, Figure 5.8 precision for the ACK kernel is presented, and since it was not the best performing kernel for databases with homogeneous background (see Figure 4.2 and Figure 4.3), the best performing kernel from the mentioned figures is also shown.

The results from the SET(ACK) kernel described in this chapter outperform the ACK kernel on whole images as well as the best performing combination of kernels and descriptors LK(CONC) in Figure 5.7or individual kernel KCSD in Figure 5.8.

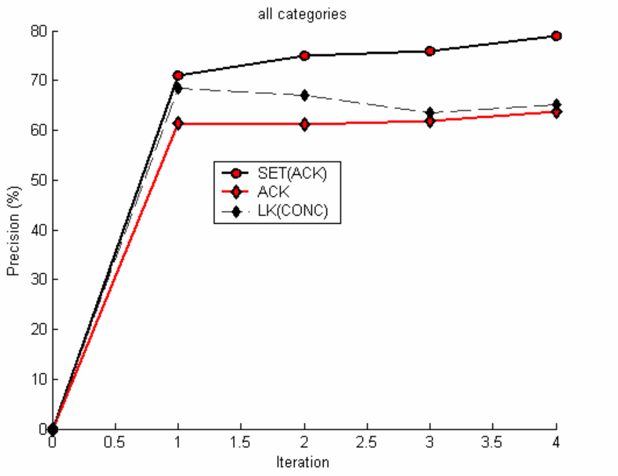


Figure 5.7: Average precision depending on the iteration, over all concepts for the D8 image database.

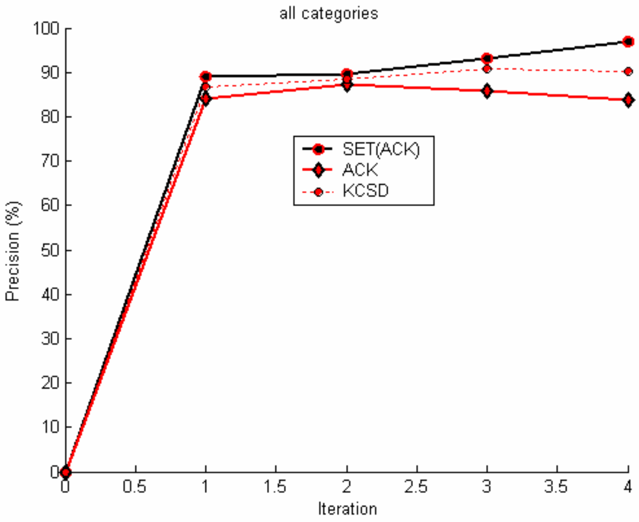


Figure 5.8: Average precision depending on the iteration, over all concepts for the D25-1800 image database.

Similar results can be seen from Figure 5.9 and Figure 5.10 with precision curves over different categories for the D7-700 image database and Caltech 101 database. In both cases the newly designed kernel incorporating spatial information outperforms the ACK kernel on whole images, which on its own had the highest precision when compared to other kernels on whole images (Figure 4.4 and Figure 4.5), in both databases.

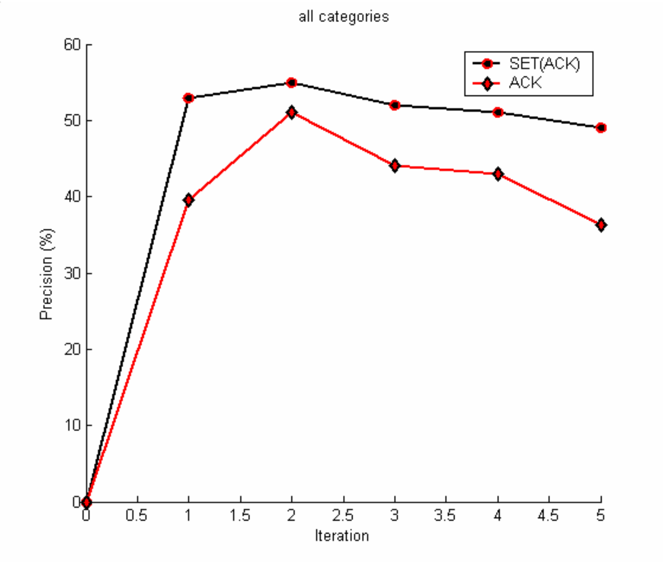


Figure 5.9. Average precision depending on the iteration, over all concepts for the D7-700 image database.

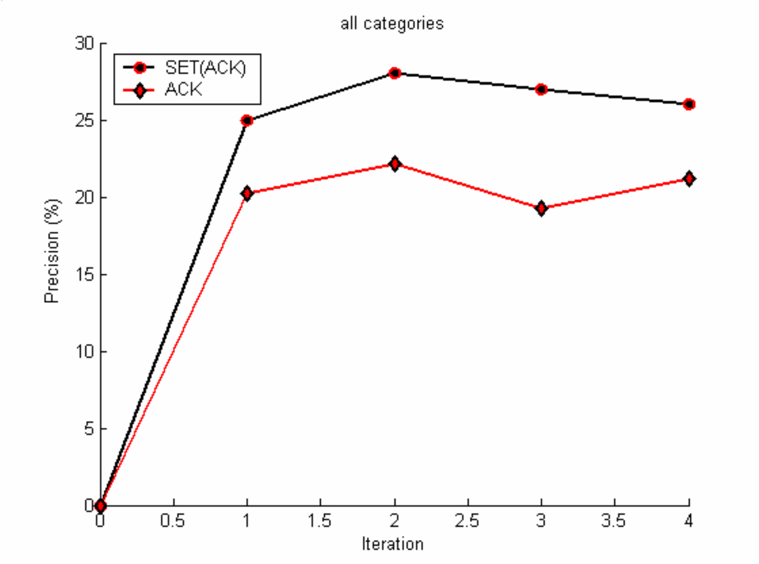


Figure 5.10: Average precision depending on the iteration, over all concepts for the Caltech 101 image database.

These results justify the idea for using structured descriptors with low-level similarity and also incorporating spatial information about building block s. The feature space was generated in a structured way exploiting both low-level content of object-based image blocks and their spatial location within an image. A set kernel with individual convolution kernels performing over multi-feature space was designed, to work with SVMs in a RF scenario. This leads to improved performance and reduction of background noise.

CHAPTER 6 : Application

6.1. Introduction

Two main applications of developed improvements for relevance feedback module have been presented in this chapter. The first one is focused on the task of adding knowledge to the image content in order to enable more “intelligent” classification. The second application describes the aceMedia system integrating a number of functionalities including the relevance feedback module described in this thesis.

6.2. A Framework for Image Selection in Concept Learning

Traditionally, proposed methods in machine learning and pattern recognition are used to select a path from visual features to semantic meaning. In this type of approaches the learning process is based on basic visual interpretation of the image content indicating observed elements in the scene, e.g. landscape, cityscape (Vailaya et al. 1998, 2001).

It is well known that two objects can be similar in their visual primitives but semantically different to a human observer. Therefore substantial noise could be introduced in propagating interpretations using only low-level similarity. On the other hand propagation based only on high-level similarity puts a heavy burden on the designer. Combined approaches that go both ways underpin the paradigm of “bridging the semantic gap.”

In this section a framework to assist concept learning from examples, in semantic-based image classification is presented (Dorado et al., 2006). The framework exploits the capability of support vector classifiers to learn from a relatively small number of samples (Jain et al. 2000). A straightforward way to choosing training samples is by random selection of images, however this does not guarantee quality or good representation of the concept. On the other hand, manual searching for good training samples has also drawbacks. One of them is how to define “a good” sample, this involves subjectivity and varies from one designer to another. Manual search could also imply the need to traverse the entire database in an effort to obtain higher efficiency. Consequently, selection of suitable examples becomes a critical design step. This framework uses unsupervised learning as the first step in designing the classifier. By applying clustering it organizes images based on low-level similarity in order to assist a designer in selection of positive and negative samples for a given concept. Basically, clustering outcomes are used to identify sensitive points that can define the hyperplane between groups of images associated with certain concepts.

Low-level feature similarity, relies only on machine’s interpretation of the content, and hence has a shortcoming in terms of efficiency due to introduction of misleading information. Here is where relevance feedback plays an important role through active learning by allowing additional training and system adaptation.

Therefore in order to refine the classifier model the initial design step is followed by an active learning step. After clustering the space of image descriptors, positive samples are selected from feature vectors situated in the well-populated regions in the neighbourhood of cluster prototypes relevant to the chosen concept. Negative samples are selected from feature vectors placed in regions with clashing cluster prototypes or in regions where two or more clusters overlap. The framework is designed to captures hints from the professional annotator observed from the clustering result.

6.2.1 The Problem of Learning Concepts

Semantic-based classifiers perform the task of using content-based descriptions to assign certain objects to a given semantic concepts. The training process in learning-from-examples is carried out by presenting declarative knowledge through a number of labelled objects. In this way human subjectivity is introduced to image classification. In order to refine the classifier model the initial design step is followed by active learning.

Bhanu and Dong (2002) proposed a framework for learning concepts based on retrieval experience, which combines partially supervised clustering and probabilistic relevance feedback. The challenge of finding suitable samples is also observed in training strategies as the one presented in Boutell et al. (2004). Several interactive approaches have been proposed to enable system adaptation based on long-term learning (Bhanu and Dong 2002; Yoshizawa and Schweitzer 2004). Tong and Chang (2001) proposed the use of support vector machine active learning algorithm for conducting effective relevance feedback for image retrieval. In a similar manner Zhang and Chen (2002) propagated annotations using training samples that enable maximum knowledge gain (reduction of uncertainty). Nguyen and Smeulders (2004) proposed a similar strategy, for active learning in classification by taking into account prior data distribution and cluster medoids. An image is considered to be either a positive or negative sample of a given concept, if it satisfies a criteria defined by a professional annotator. However, subjectivity of the selection criteria, amount of available examples, image occlusion, shadows, rotation etc., are just some of the identified drawbacks in collecting training patterns. Choosing samples based just on human perception misses out on the fact that in the end the classifier will be using descriptions with limited domain knowledge, and not the overall cognitive human perception. Then, the problem is how to assist designers in selecting samples to train semantic-based image classifiers.

6.2.2 A Framework for Concept Learning

Low-level features are organized by combining unsupervised and partially supervised training modes. The objective is to find a classifier model that roughly resembles the semantic categorization of images.

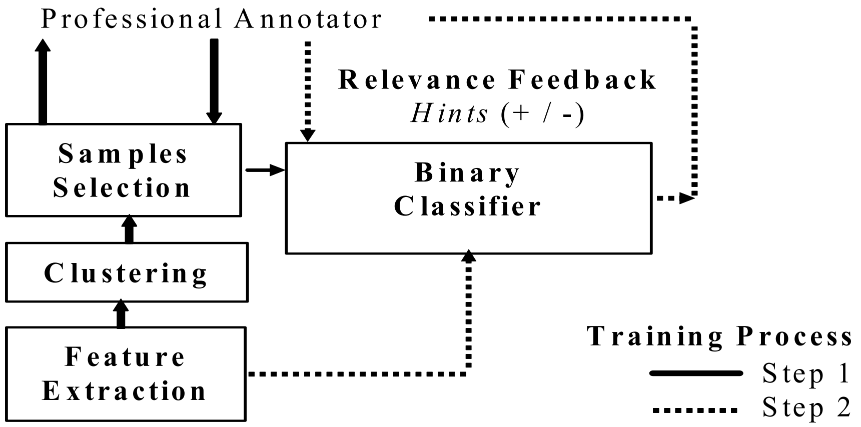


Figure 6.1: Framework for training a SVM classifier. The first step uses clustering to assist the professional annotator in selecting image samples. The latter step applies active learning through relevance feedback to refine the classifier model

An initial data set is built with the best-ranked images in the clusters, Figure 6.1. These images are associated with sensitive points of two types: *positive samples* with high membership to a relevant cluster; and *negative samples* with high membership to a non-relevant cluster, associated with a different concept than the relevant one

These positive and negative samples constitute the candidates of training patterns. Then in an active leaning procedure the annotator follows a sample selection procedure to decide if the candidates are suitable samples to obtain a classifier model (see Figure 6.2)

The second step in the training process uses relevance feedback to refine the classifier model. The classifier predicts positive examples for the category from the unlabelled images.

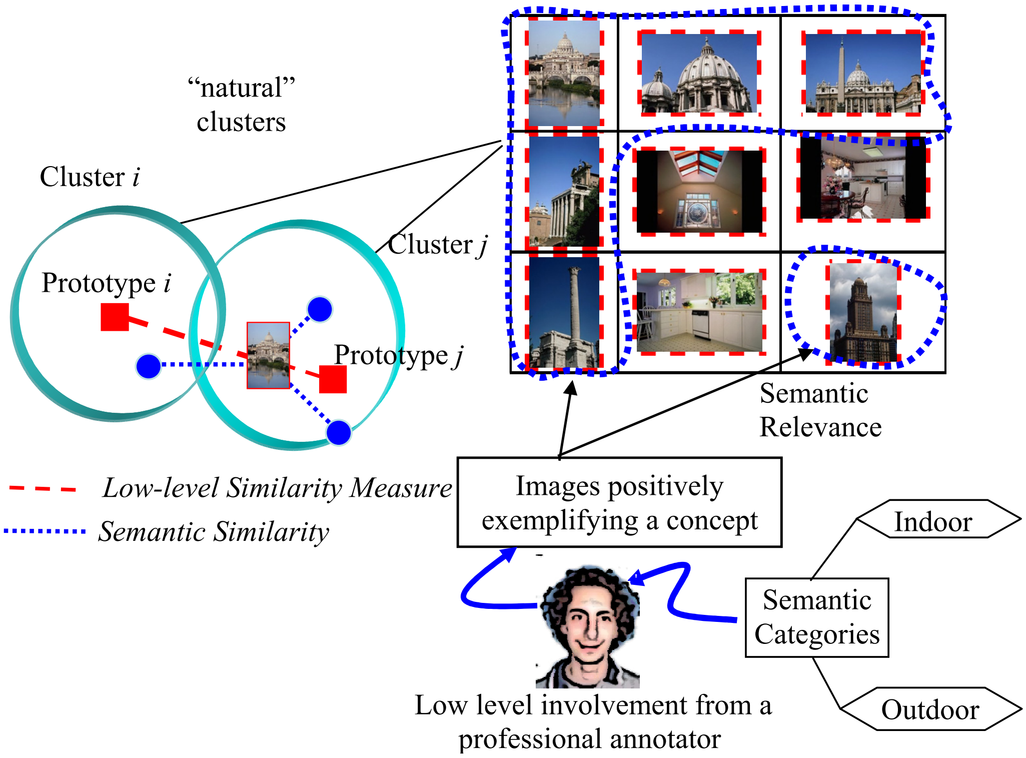


Figure 6.2: Finding design samples for a first training round. Low-level similarity is captured by the clustering algorithm. Professional annotator indicates relevant images to the concept.

The professional annotator provides hints indicating positive and negative images found among the classification results. These annotator's hints are collected to update the training data set. Furthermore, both positive and negative images are used to refine the classifier. The introduced knowledge accumulated during the training interactions is used to increase the problem domain knowledge and enable long-term learning. The framework's components are detailed below.

Unsupervised Clustering

The goal of the clustering task is to help organize low-level features into groups with interpretations that may relate to relevant concepts of the image content. Thus, features are clustered according to similarities among them (Jain et al. 1999). The cluster assignment is essentially unsupervised and no prior knowledge of the underlying content is used in the algorithm. Although any clustering mechanism helps in revealing the structure of the data set, the nature of the problem requires an extension to deal with the subjectivity and fuzziness of the human interpretation. In the proposed framework, the clustering task is carried out using Fuzzy C-Means (FCM) (Pedrycz 1990).

FCM is an optimisation technique based on minimization of objective function that measures the level of data space partitioning. The objective function indicates the quality of the partition and has the following form:

$$J(X,\mathbf{V},\mathbf{U}) = \sum_{i=1}^{m}\sum_{j=1}^{c} u_{ij}^{p} d^{2}(\mathbf{x}_i,\mathbf{v}_j)$$

where X is a data space, with m elements, feature vectors $\mathbf{x}_i$ of dimension N . $\mathbf{V}$ is a set of $c\ (2 \le c \le m)$ cluster prototypes with N -dimension elements $\mathbf{v}_j$. $p\ (1 < p < \infty)$ is a fuzzy exponent determining a degree of overlap of fuzzy clusters and $\mathbf{U}$ is a matrix defining fuzzy partitions:

$$\mathbf{U} = [u_{ij}],\ u_{ij} \in [0,1],\ \sum_{j=1}^{c} u_{ij} = 1,\ \ 0 < \sum_{i=1}^{m} u_{ij} < m \tag{6.1}$$

Where u_{ij} is the degree of membership of vector $\mathbf{x}_i$ in the cluster j. $d^{2}(\cdot)$ is any distance norm expressing the similarity between any feature vector and the prototype.

A drawback of the approach is that after a number of iterations the solution can converges to local minima, which is not necessarily the optimal one. The convergence is independent of the change in the distance function if the distances are all positive and the prototypes are calculated accordingly to the minimization of the objective function.

An illustrative example of using clustering as pre-processing mechanism to find suitable samples is presented in Figure 6.3. FCM provides the cluster prototypes as well as the feature space partition. Membership degrees of patterns to each cluster are used to collect candidate images of design samples.

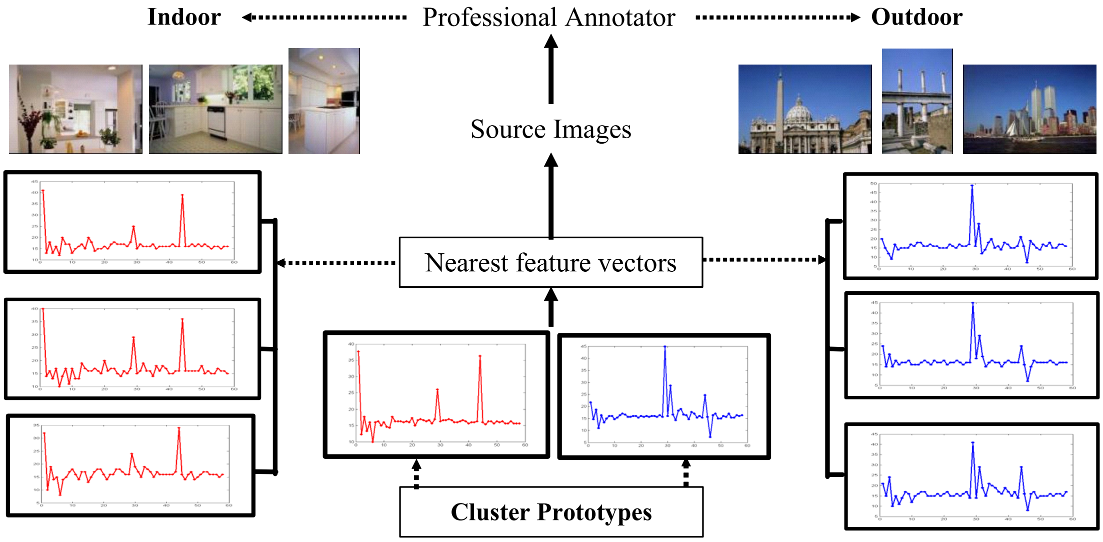


Figure 6.3: Selection of sample images. Training examples are chosen from nearest feature vectors to cluster prototypes. In this case, feature vectors correspond toCLD.

The best-ranked images, nearest patterns to the prototypes, are organized into sets, which are presented to the annotator who selects images that positively or negatively represent the concept. These are training samples used to train the classifier in a first round. A basic classifier model is obtained using these samples.

Relevance Feedback

SVMs show good performance for the generalization task over various pattern recognition problems and with small training data sets (Duin 2000), which makes them appealing for this framework. The adaptive convolution kernel analysed in Chapter 4 with appropriate number of feature spaces, was also used here for relevance feedback and within the binary classifier.The proposed modification uses SVM and employs kernel-learning approaches to optimise the non-linear mapping introduced with kernels for a better correspondence to chosen features. As depicted in Figure 6.1, the system captures hints of domain knowledge related to the classification problem. During the second step of the training process, a professional annotator provides hints indicating to the classifier whether or not its decisions were correct (positive or negative hints). The classifier uses those hints to adjust the boundaries between patterns containing (or not) the concept. These boundaries are defined by the hyperplane based on support vectors. The idea of this supervised learning step is not to estimate distributions of the known/unknown patterns but to learn the optimal non-linear decision hyperplane.

6.2.3 Experimental studies

Experiments were performed on images selected from the Corel stock gallery. Two groups consisting of 1035 and 1200 photographs were organized into a number of high-level semantic categories.

The first group was used to classify indoor (kitchens, bathrooms, office interiors, museums, etc.) and outdoor (contemporary buildings, city architecture: Rome, Chicago, etc.) images. The second group was used to classify animals (dogs, tropical sea life, etc), city views (New York city, Ottawa, etc), landscapes (autumn, Yosemite, etc), and vegetation (perennials, plants, American gardens, etc) images. The indoor/outdoor feature space was built with vectors containing CLD descriptions. On the other side, the animal-city_ view-landscape-vegetation feature space combines CSD, EHD, and HTD.

Cluster analysis

The similarity of best-ranked images after clustering for the indoor/outdoor classification problem partially resembles the expected semantic grouping.

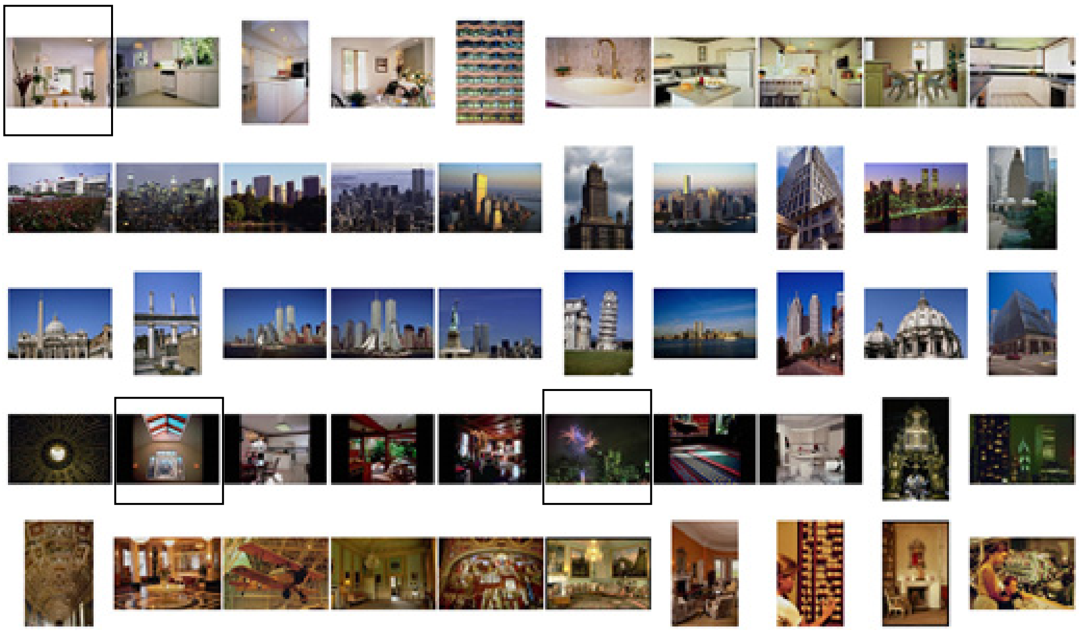


Figure 6.4: Top-ten ranked images according to highest membership in the clusters. Categories: indoor and outdoor. Each row corresponds to a representative set of a cluster.

The number of clusters was experimentally determent to five and images ordered based on membership degrees. Clustering results for the indoor/outdoor classification problem indicate that colour is an appropriate descriptor to create a separable feature space in this domain.

As depicted in Figure 6.4 the first set (row 1) contains samples of indoor images, except the fifth image that corresponds to a building lose-up. Most of the displayed images are good candidates of outdoor (rows 2 and 3) and indoor (rows 4 and 5) concepts. The sixth and tenth images in the fourth set (row 4) are negative examples of indoor category, though their colour distribution is closer to the prototype of this group. Following figures contain the best-ranked images in the clusters for the classification problem of four categories animal, city view, landscape, and vegetation. Low-level similarity is based on colour and texture features. The feature space was partitioned into ten clusters.

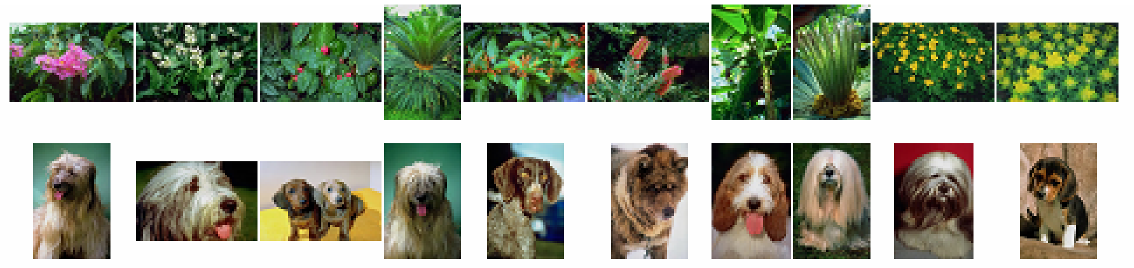


Figure 6.5: Sets satisfying criteria for the semantic categorization: vegetation and animal.

Figure 6.5 show a sample of image sets satisfying criteria for the semantic categorization. Images found in each cluster set can be directly attached to a category, row 1 (cluster 4) to vegetation and row 2 (cluster 7) to animal.

Figure 6.6 gives a sample of image sets with overlapping criteria for the semantic categorization. Each row corresponds to two different clusters. The row 1 (cluster 8) can be qualified as landscape except for the last image (10th column), which is a sample of city view; the second row (cluster 9) satisfies criteria for category vegetation except the image in the 2nd column containing a city view scene. The third and fourth rows show overlapping between categories animal-vegetation and city view-landscape with strong commonalities in their distributions of colour and texture descriptions.

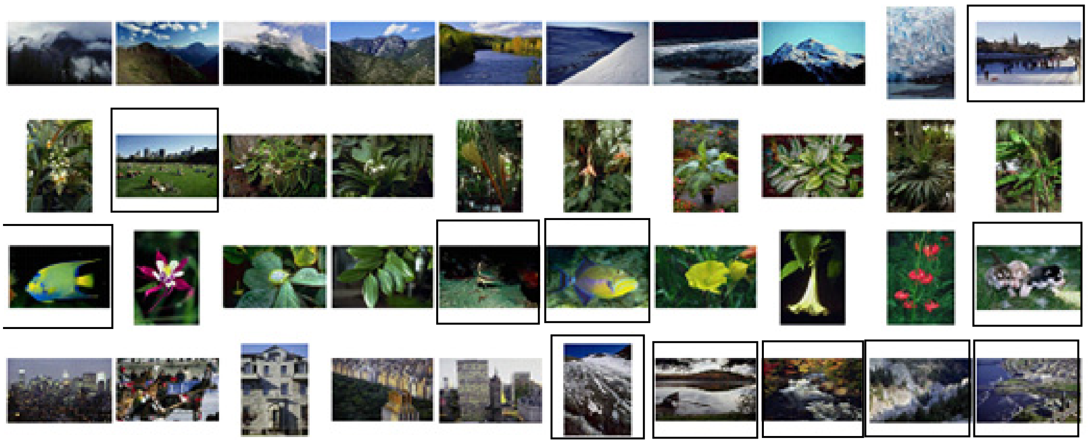


Figure 6.6: Overlapping categories: row 1: landscape-city view, row 2: vegetation-city view, row 3: animal-vegetation, and row 4: city view-landscape

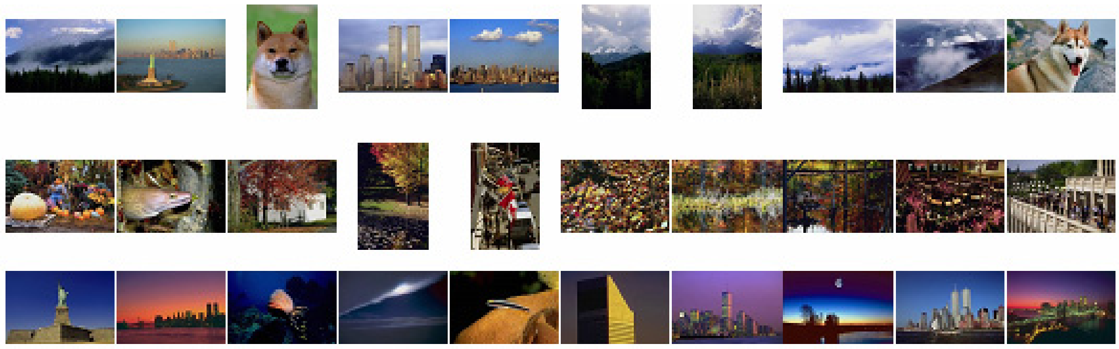


Figure 6.7: Sample of sets containing mixed objects from different categories. row 1, row 2, and row 3 are examples of how low-level similarity can lead to semantically meaningless grouping.

As expected, some sets in the ranked images contain objects from more than two categories. It shows why the clusters cannot be attached to a single category. Consequently, relying just on low-level similarity can lead to semantically meaningless grouping (see Figure 6.7).

Framework assessment

In order to evaluate stability of the classifier model, a set of experiments were carried out using random selection of samples. Conversely, this approach skips the clustering procedure. As can be observed in Figure 6.8, the classification results lack stability. It is due to sample collection based on visual inspection along with subjective criteria of the annotator without taking into account any low-level similarity.

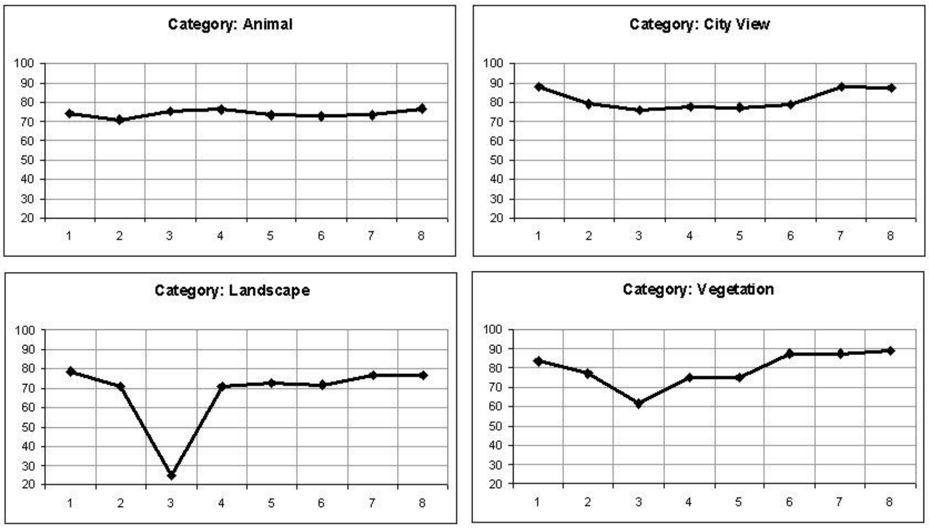


Figure 6.8: Classification results using random selection of images. X-axis indicates the number of iteration in which the annotator provides new samples to the classifier. Y-axis shows the resulting accuracy

The three training approaches summarized in Table 6.1 are used to assess the performance of the classifier within the proposed framework.

Table 6.1: Training approaches used to assess the classifier performance

Training approach	Description
SVM+FCM	SVM classifier assisted with hints provided by a professional annotator governed by clustering (FCM) results during the training phase. Samples are selected from the nearest patterns (see Figure 6.5-Figure 6.7) to the cluster prototypes.
SVM+RF	SVM classifier using only subjective relevance feedback from the professional annotator, obtained by browsing through the database.
SVM+FCM+RF	SVM classifier trained by combining both clustering results and relevance feedback from the professional annotator on the pre-clustered set.

As a result, clustering mechanisms not only assist in the sample selection, but also contribute to the system's stability (see Figure 6.9).

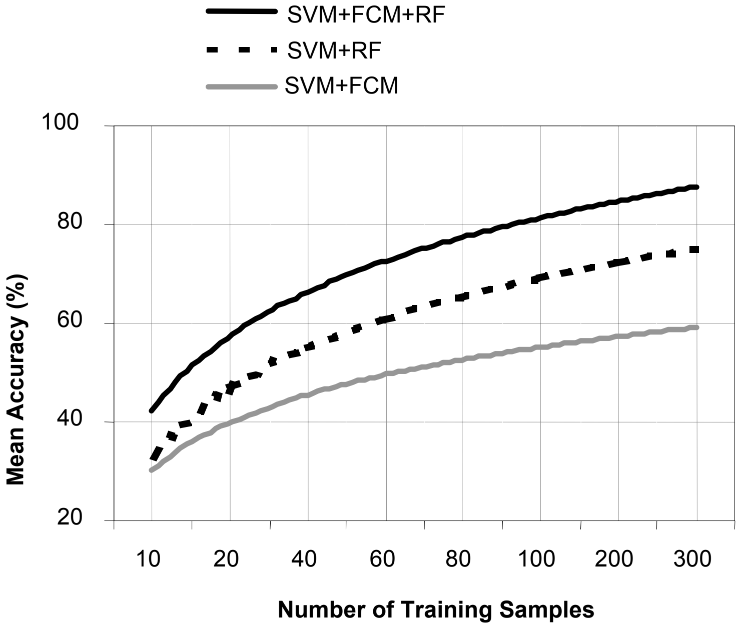


Figure 6.9: Mean accuracies achieved in classification problem using the training approaches detailed in Table 6.1

Mean accuracies obtained in the experimental studies are presented in Figure 6.9. The lowest accuracy is obtained when the support vector classifier learns only from clustering outcomes; the classifier behaves better when using relevance feedback from the professional annotator; the accuracy is further improved when relevance feedback is based on cluster prototypes.

Accuracy in the first approach (SVM+FCM) is lowest though it is expected due to the sensible reduction on the required supervision. The professional annotator needs only to indicate the class label of each cluster. This lightens the burden of annotation while introducing noise at the same time.

The second approach (SVM+RF) depends entirely on the images shown to the user. An inconvenience here is the overall subjectivity due to the fact that selection of sample relies completely on subjective interpretation of images ignoring any low-level similarity between image descriptors.

The third approach (SVM+FCM+RF), corresponding to the proposed method, shows highest performance. It has the advantage of taking into account the underlying low-level structures (revealed by the clusters) while minimizing the required supervision.

The probability of membership for each image to a class is represented through training SVMs and fitting parameters of additional sigmoid function to posterior probability for that class (Platt 1999b). SVMs produce an uncelebrated decision value f, which represents the distance from the separating hyperplane, hence this is not a probability. Based on empirical data, Platt suggest exponential forms for class-conditional densities $P(f|y=1)$ and $,P(f|y=-1)$ between margins of -1 and 1. Where y is the predicted label A parametric form of a sigmoid is used to fit posterior probability $P(y=1|f)$:

$$P(y=1|f)=1/\left(1+\exp(af+b)\right)$$

This approach has two parameters a,b which are trained discriminatively. Figure 6.10 depicts probabilities larger than 0.5 that samples belong to the relevant class. In Table 6. 2 values for accuracies achieved by the two-class classifiers for both datasets, are presented.

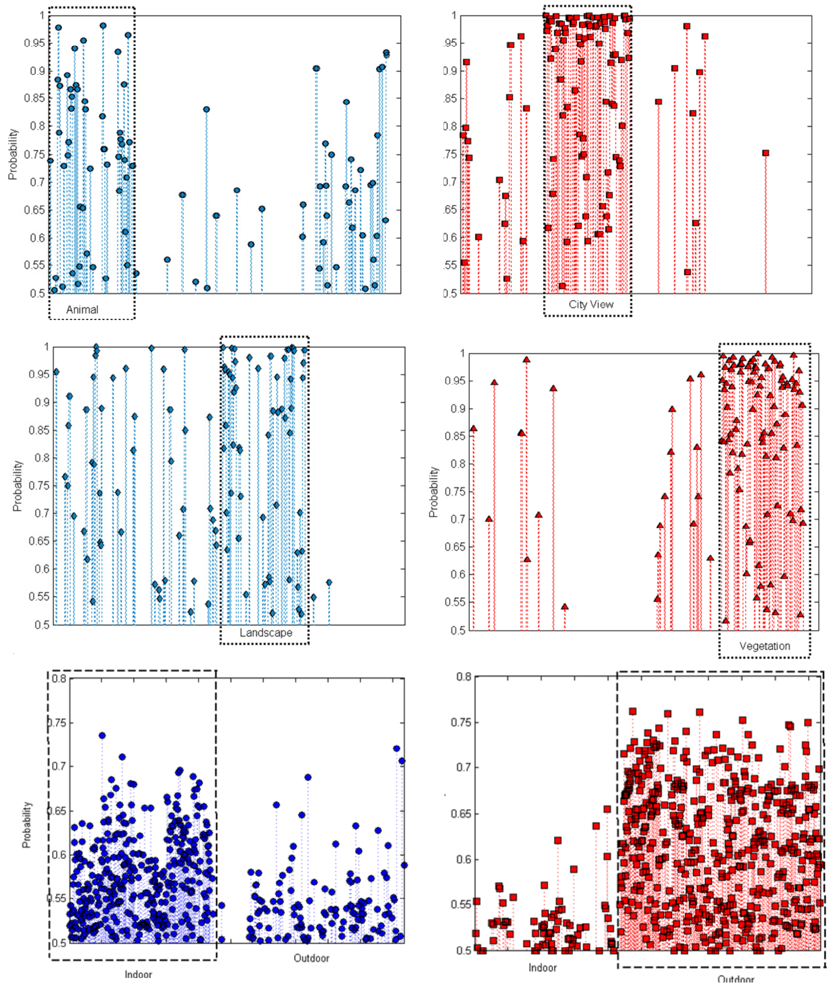


Figure 6.10: Two-class classification outcomes. Input patterns are organized along the X-axis. Y-axis indicates the probability of membership to the corresponding category.

Table 6. 2: Accuracy of two-class classifiers (%)

Animal	City view	Landscape	Vegetation	Indoor	Outdoor
74.38	87.29	74.18	87.29	82.76	79.3

Some samples of correctly classified and misclassified images are given in Figure 6.11 and Figure 6.12.

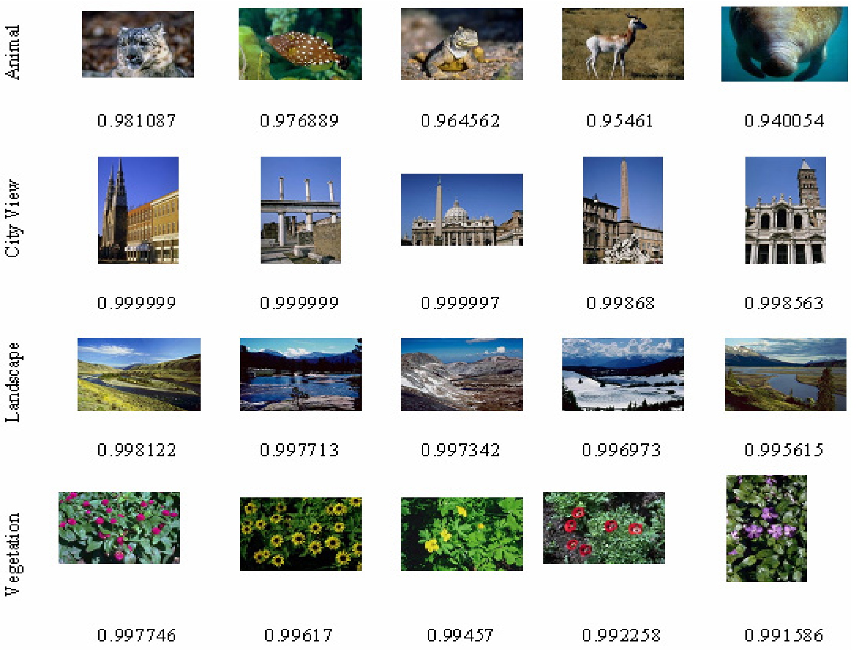


Figure 6.11: Samples of images correctly classified. Probability is indicated below each image.

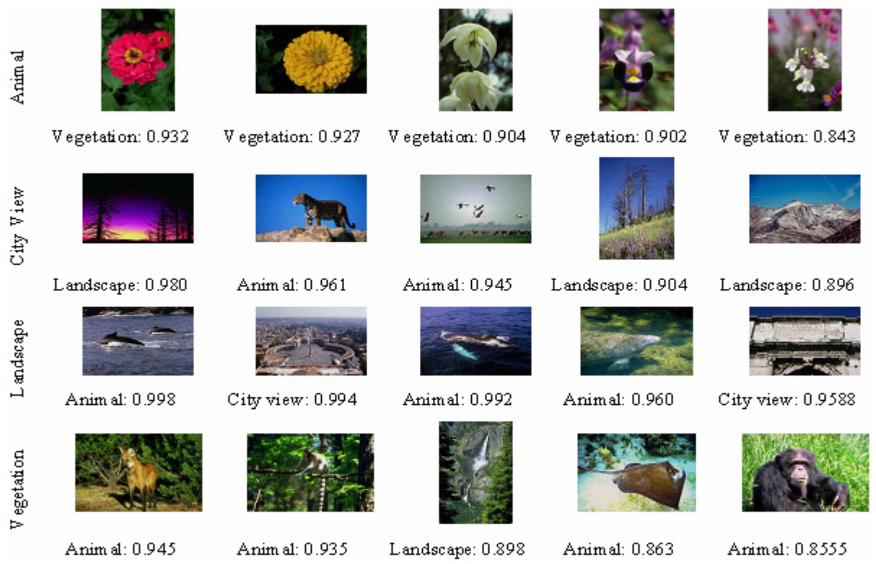


Figure 6.12: Samples of misclassified images. The assigned categories s given on the far left. Misclassification probability and true class are indicated below each image

A framework to assist a professional annotator in choosing image samples to train a semantic classifier was presented. The approach uses clustering mechanisms to reveal the underlying structure in training data in order to shift low-level features towards high-level information. This learning mode reduces the burden of collecting samples by browsing as well as it improves the quality of the chosen samples by taking into account low-level similarity. The applied keyword-oriented classification is useful to describe images with a controlled vocabulary. High precision and accuracy levels are obtained by combining supervised and unsupervised modes of learning.

6.3. The aceMedia project

The relevance feedback module presented in this thesis is a part the European IST aceMedia Integrated Project. aceMedia primarily focuses on development and implementation of a system based on knowledge assisted, adaptive multimedia content management while addressing user needs (Kompatsiaris et al. 2004). The main technological objectives are:

- to discover and exploit knowledge inherent to the content
- to automate annotation at all levels
- and to add functionality to ease content creation, transmission, search, access, consumption and re-use.

6.3.1 aceMedia High Level Objectives

The aceMedia project is centred around the idea of an Autonomous Content Entity (ACE). An ACE has three layers, content, associated metadata, and intelligence layer. The intelligence layer, of which the RF module is part of, consists of distributed functions that enable the content to instantiate itself according to its context including its network environment, the user terminal, and recorded user preferences.

The aceMedia high level objectives are to build a system based on following research areas:

- Knowledge and context-assisted content analysis, based on a multimedia ontology infrastructure to support semantic entity detection and tracking of ACE content.

- High-level semantic reasoning tools for automatic annotation and generation of the ACE metadata layer.
- Query analysis tools and intelligent ACE search, retrieval, ranking and ***relevance feedback*** mechanisms.

The project is developed into two application frameworks, enabling for both home network and mobile communication environments.

6.3.2 aceMedia System Overview and Relevance Feedback Module

The aceMedia integrated project draws together fundamental research in knowledge technologies and multimedia processing, within a user centred design framework. Figure 6.13 shows a system level representation of aceMedia, which depicts the contribution of various research disciplines (aceMedia, 2005).

The content subsystem (denoted as (1) in the Figure 6.13) interacts with the aceMedia content (denoted as (a)) obtained from the content creator. This subsystem handles the essential processing tasks including *pre-processing*, *scalable coding*, *cross-media adaptation* and *visualization*. The user subsystem (denoted as (2) and (3)) interacts with the aceMedia end user (denoted as (b)) and supports the elementary operations required by the service provider and device manufacturer, including: *content search*, *browsing* and *personalisation.* Finally, the knowledge subsystem (denoted as (4) and (5)) interacts with aceMedia knowledge (denoted as (c)) to implement all intermediate steps of *intelligent analysis and reasoning.*

In aceMedia, intelligent search and retrieval is closely related to both visual and textual queries. Advantage is taken of both the metadata obtained by the knowledge-based analysis of content and of the refined visual characterisation of this content. Intelligent search and retrieval modules provide a set of mechanisms for both initialising and refining search, by taking into account natural language queries, queries by visual content and relevance feedback (Heinecke et al. 2005).

The intelligent search and retrieval mechanisms are designed to be very flexible, allowing the users to combine them according to their intentions, with the possibility of influencing the searching session based on subjective interpretations of the content through visual relevance feedback. The RF module in aceMedia framework is a results of the research on non-linear kernels from Chapter 4, specifically for a combination of two descriptors CSD and EHD. This module is used to refine search queries. Using positive and negative user response, new information is fed into the system and a new iteration of the search session enabled.

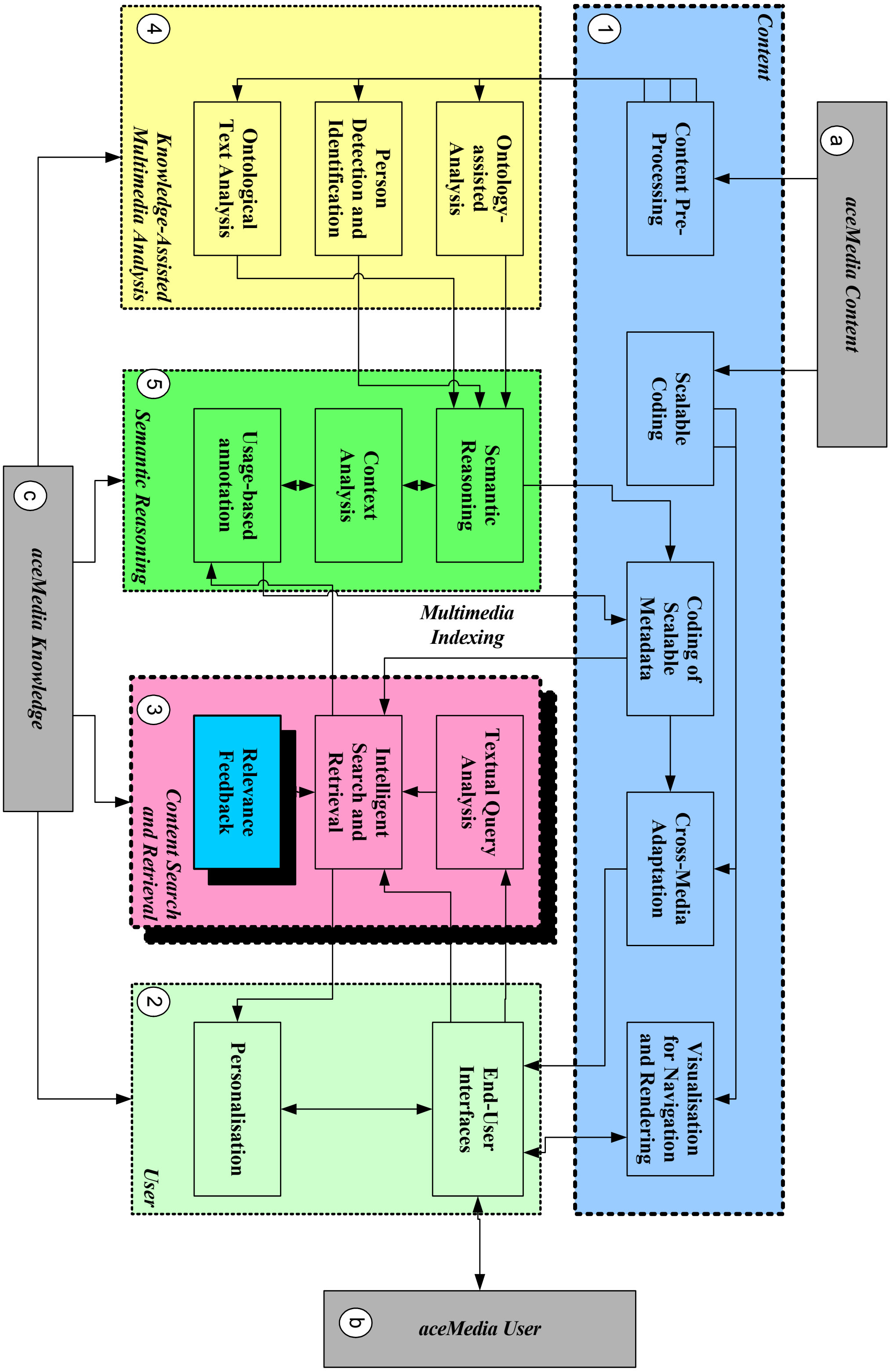
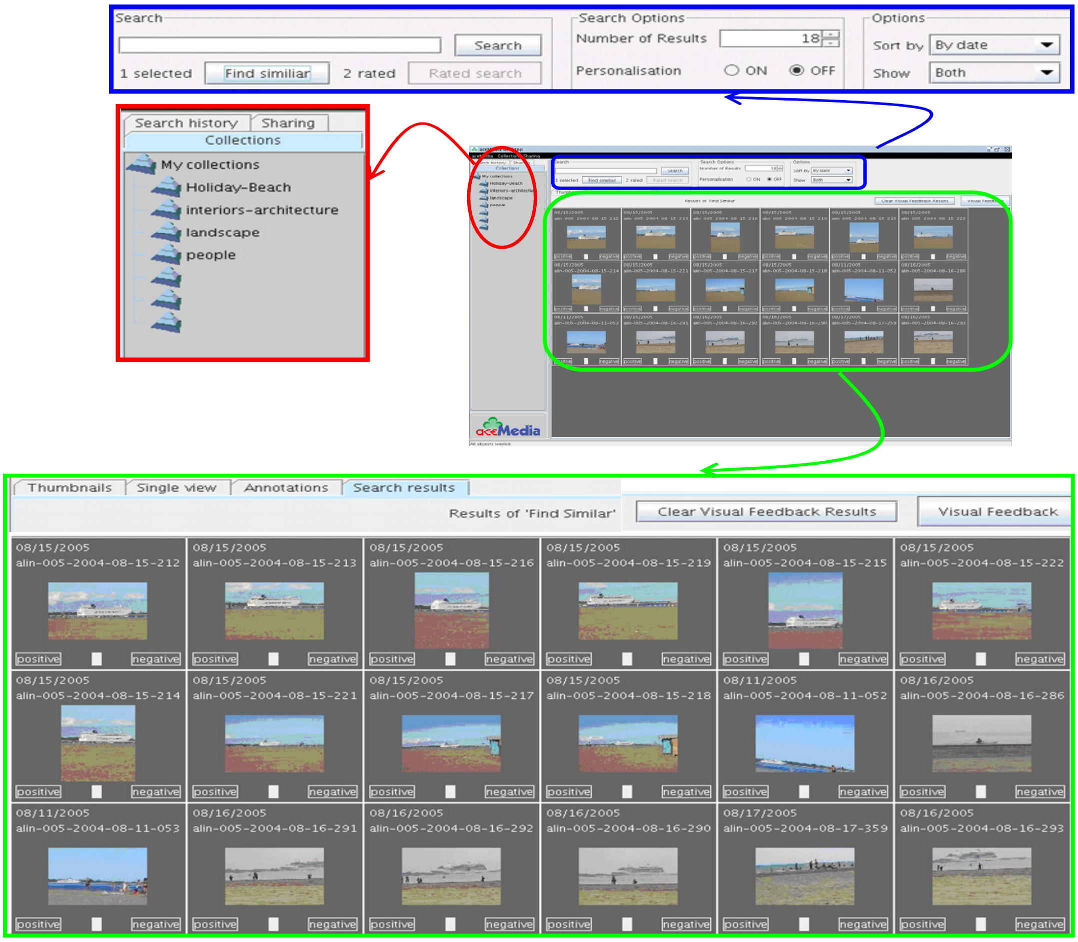


Figure 6.13: The aceMedia system diagram.

Figure 6.14 shows the interface of the visual relevance feedback module within the ace framework. Additional search functionalities are also enabled :

- Textual query: natural language user query is processed by the aceMedia knowledgebase and a first set of resulting images is provided to the user. The user can further refine his search using either a query by visual example or visual relevance feedback.
- Query by visual example: the user can select one image from the repository and ask the system to return the images that are the most similar to this image. Again, such a visual query can be followed by a refinement using visual relevance feedback.

The refinement of the search always relies on the relevance feedback provided by the user on the results of the previous queries



1 selected | 2 rated

Number of Results 18

Personalisation ○ ON ● OFF

Sort by By date

Show Both

08/15/2005 alin-005-2004-08-15-212
08/15/2005 alin-005-2004-08-15-213
08/15/2005 alin-005-2004-08-15-216
08/15/2005 alin-005-2004-08-15-219
08/15/2005 alin-005-2004-08-15-215
08/15/2005 alin-005-2004-08-15-222
08/15/2005 alin-005-2004-08-15-214
08/15/2005 alin-005-2004-08-15-221
08/15/2005 alin-005-2004-08-15-217
08/15/2005 alin-005-2004-08-15-218
08/11/2005 alin-005-2004-08-11-052
08/16/2005 alin-005-2004-08-16-286
08/11/2005 alin-005-2004-08-11-053
08/16/2005 alin-005-2004-08-16-291
08/16/2005 alin-005-2004-08-16-292
08/16/2005 alin-005-2004-08-16-290
08/17/2005 alin-005-2004-08-17-359
08/16/2005 alin-005-2004-08-16-293

Figure 6.14: The Relevance feedback module integrated within the aceMedia project. The interface also supports other functionalities as: generation of collections, search by natural language, search by visual examples etc.

CHAPTER 7 : Conclusions

7.1. Conclusions

One of the most important aspects of today's interactive multimedia systems is the ability to retrieve visual information related to a given query, preferably formulated in semantic terms. This important functionality can be achieved only if the content is well structured and annotated with key-words representing semantic concepts. Unfortunately, the gap between the capabilities of current image understanding algorithms and the richness and subjectivity of semantics in human interpretations of audiovisual media is a formidable obstacle. As means of achieving a step closer to bridging the gap between human and machine driven reasoning, iterative short term and low-effort relevance feedback, has been presented as an obvious step. The thesis presents a thorough study of visual content retrieval using relevance feedback.

The work started with analyzes of low-level features focusing on discriminative properties. As a result of thorough evaluation a combination of features able to effectively capture low-level representations of natural images for a retrieval scenario was obtained (Chapter 2). Different low-level descriptors and similarity measures are not designed to be combined naturally and straightforwardly in a meaningfully manner. Thus, questions related to the definition of multiple feature spaces as well as their similarity functions have been addressed.

Then, possible learners for the RF scenario were instigated. Special attention was paid to the backbone learning theory that might support the use of one machine learning method over another, for the particular RF scenario. As a result of this analyses image relevance feedback framework based on support vector machines as a learning approach was implemented (Chapter 3).

An adaptive convolution kernel dealing with multi-feature spaces has been proposed. This approach enables feature combination and not just concatenation. The kernel facilitates adaptive similarity matching and models the multi-feature space at the same time guaranteeing convergence of the convex support vector optimisation problem (Chapter 4).

It was noticed that labelling complete images as relevant to a given key-word introduces a lot of noise due to the variety of non relevant objects in complex scenes. Hence, a set kernel coupled with clustering approaches, defined in structured space has been proposed. The kernel encloses both multi-feature space and spatial information about localized image structures, enabling a higher transparency between low-level image features and semantic concepts. This approach gives higher performances in a retrieval scenario with RF than approaches dealing with whole images, since the background noise is avoided (Chapter 5).

Application to classification for improved image selection in concept learning has also been proposed. This framework combines unsupervised learning to organize images based on low-level similarity, and reinforcement learning based on relevance feedback, to refine the classifier model. The results show improved performances for the classifier and higher stability, than when just browsing or unsupervised approaches are used to train the classifier model (Chapter 6).

The objectives of this worked outlined in S1.2 have been fully accomplished entirely.

7.2. Future Work

Many issues have still not been solved.

For instance, on-line parameter adaptation of the scale parameter leading to higher resilience to changes in sizes of different image classes could further improve performances.

An idea of convolution kernels can be expanded to different wrapping function (not necessarily exponential ones) that could more precisely define the data and enable combination of features very different in nature as visual, audio and motion information.

Contextual dependencies among structures can be further improved by incorporate prior knowledge into the process for obtaining representative structures and also integrate knowledge into the learning approach for relevance feedback.

One aspect of outmost importance is how to retrieve images to the user, and how to allocate then different levels of relevance Employing fuzzy support vector machines coupled with our structure space can not only capture relevant representatives of relevant and irrelevant images but also use the coherency levels in across image clustering to allocate levels of importance.

APPENDIX A : Ground Truth Image Databases

A number of ground truth image databases are described and presented in section 2.5.1. In this appendix an overview of the visual appearance of the image in every databases is given:

- DColour and VisTex image databases have ground truths based on visual appearance of images belonging to the same class and sharing same color, shape or texture characteristics (see Figure A. 1 and Figure A. 2 ; Table A. 1).
- D25-1800 and D8 databases possess similarity in higher-semantic level, object level, with consistency in visual appearance achieved through uniform background and different views of the same object (see Figure A. 3 and Figure A. 4).
- D7-700 and Caltech 101 databases have a higher level ground truth, based on semantic meaning. Images belonging to the same class illustrate the same concept, but their visual appearance may differ considerably (see Figure A. 5 and Figure A. 6; Table A. 2).

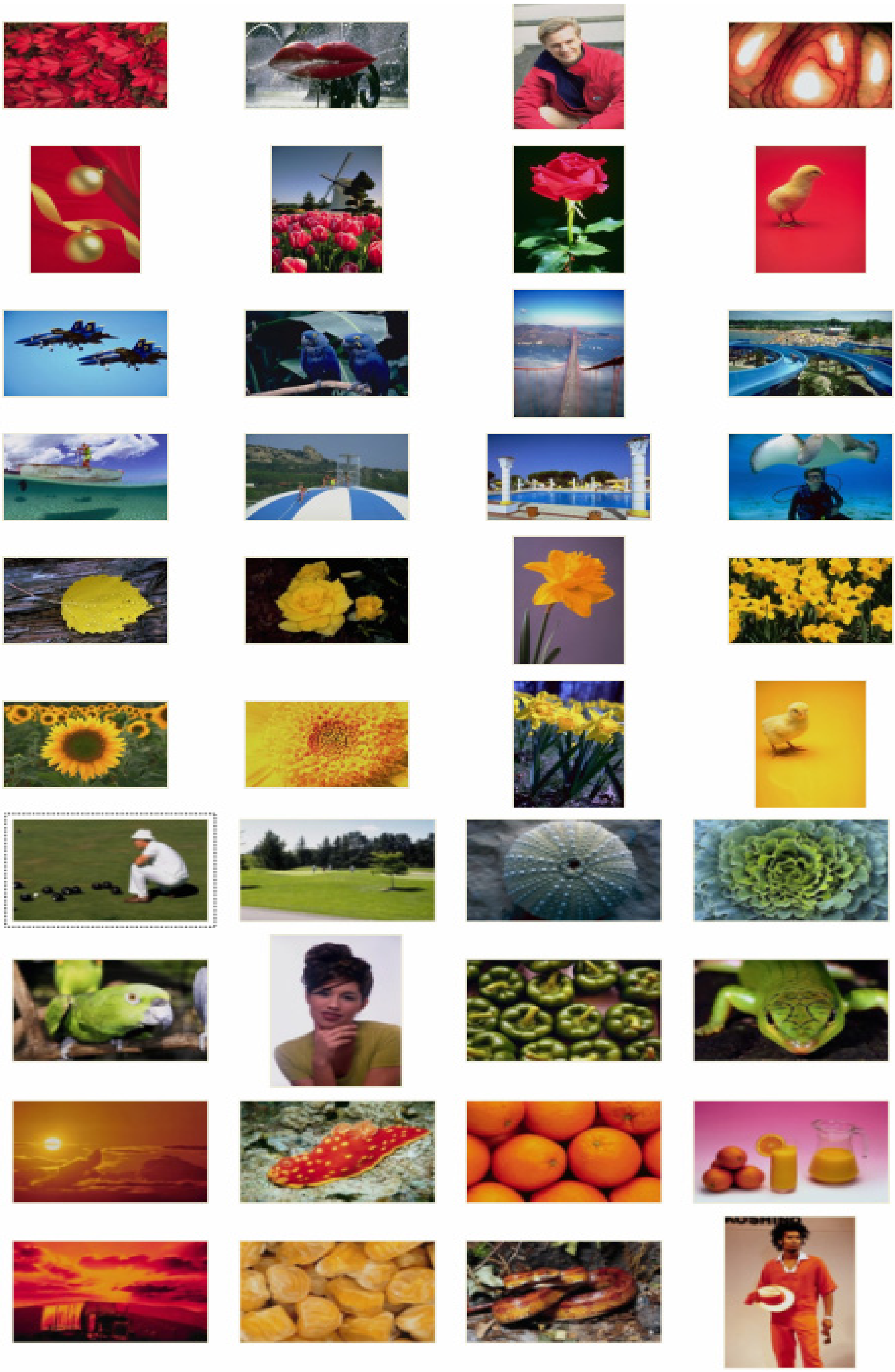


Figure A. 1: Samples of the DColour image dataset. Two rows for each class (top to bottom): red, blue, yellow, green, and orange.

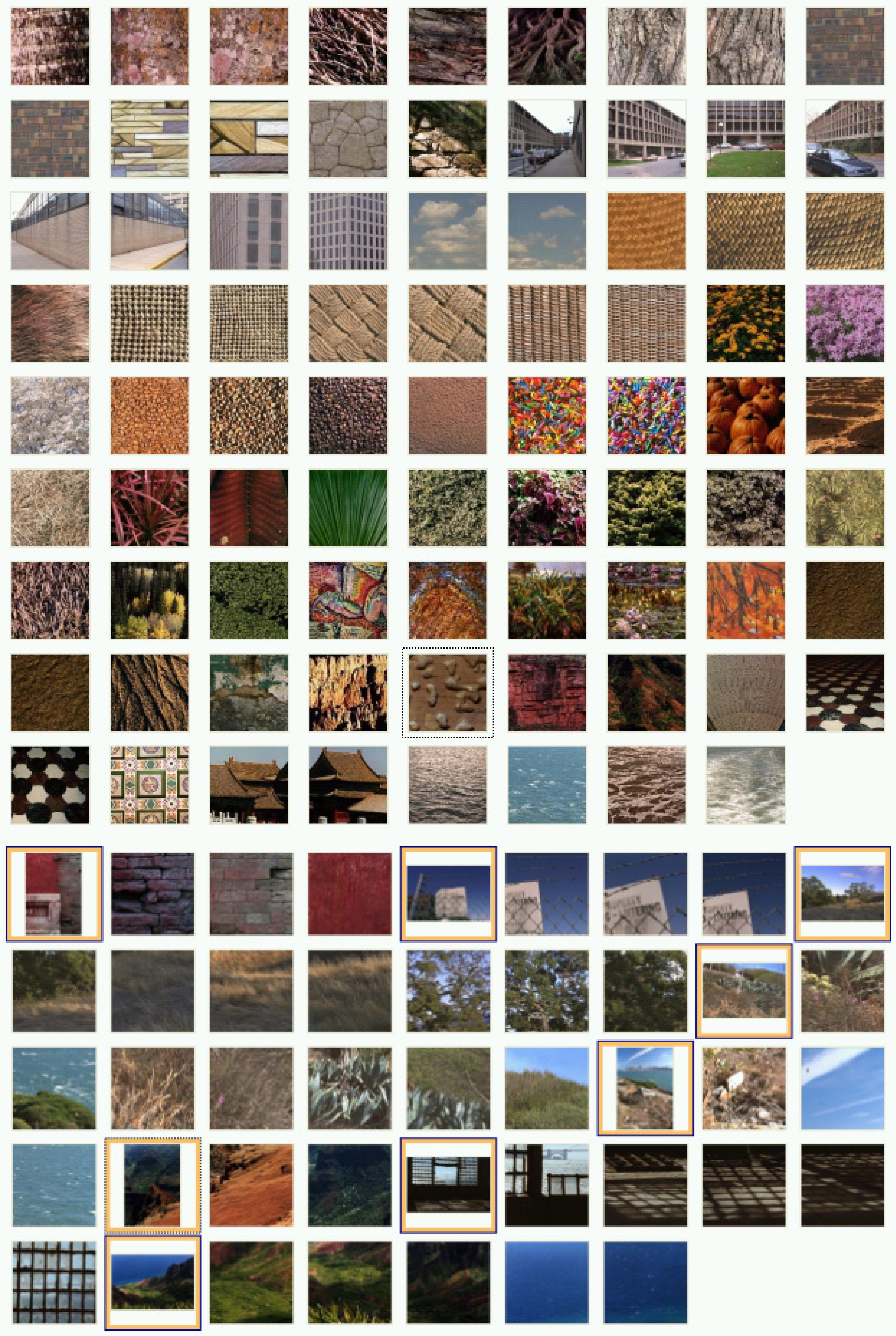


Figure A. 2: Samples of the VisTex image database, reference images (first 9 rows) and scenes (last 5 rows). Highlighted images represent contextual scenes while subsequent images are patches of these scenes (left to right, top to bottom).

Table A. 1: List of categories in the VisTex image database

Reference images		Scene images
Bark	Misc	Brick Paint
Brick	Paintings	Corridor
Buildings	Sand	Doc Cage City
Clouds	Stone	Fence Sign
Fabric	Terrain	Grass Land
Flowers	Tile	Grass Land2
Food	Water	Grass Plants Sky
Grass	Wheres Waldo	GroundWater City
Leaves	Wood	Mt Valley
Metal		Prison Window
		Valley Water

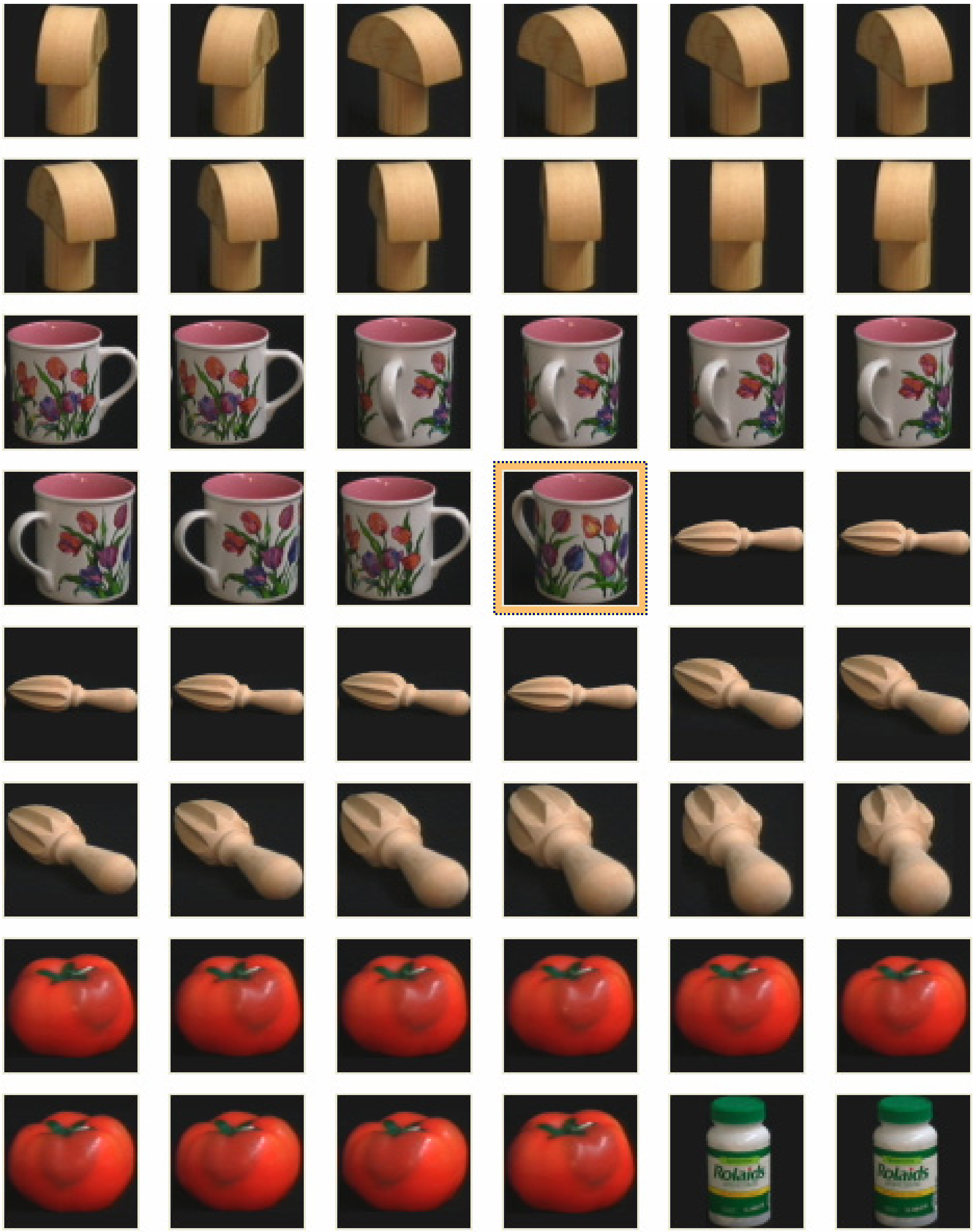


Figure A. 3: Samples of the D25-1800 image database.

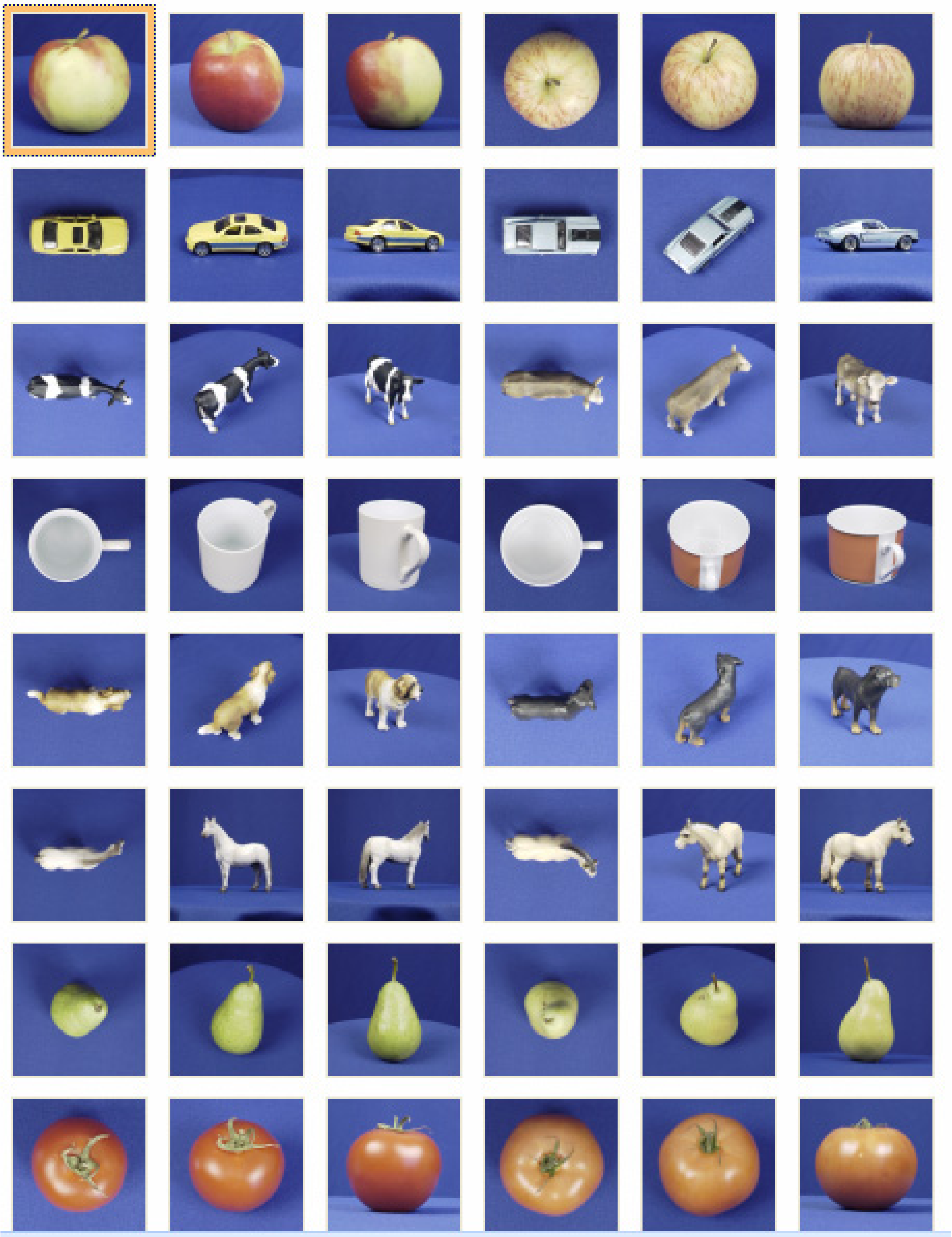


Figure A. 4: Samples of the D8 image database, categories: apple, car, cow, cup, dog, horse, pear and tomato.

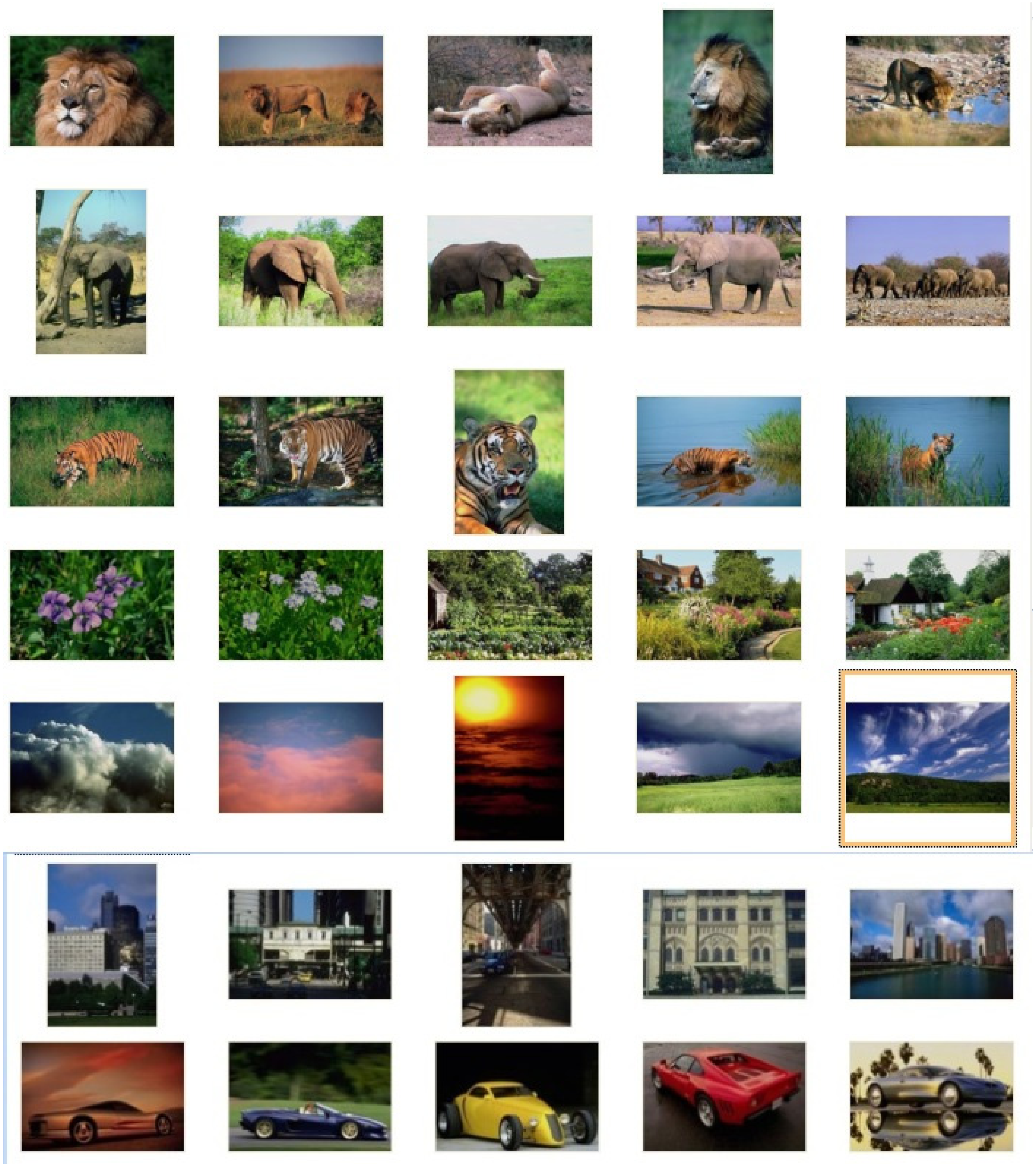


Figure A. 5: Samples of the D7-700 database from the Corel stock, categories: lions, elephants, tigers, grass, clouds, buildings and cars.

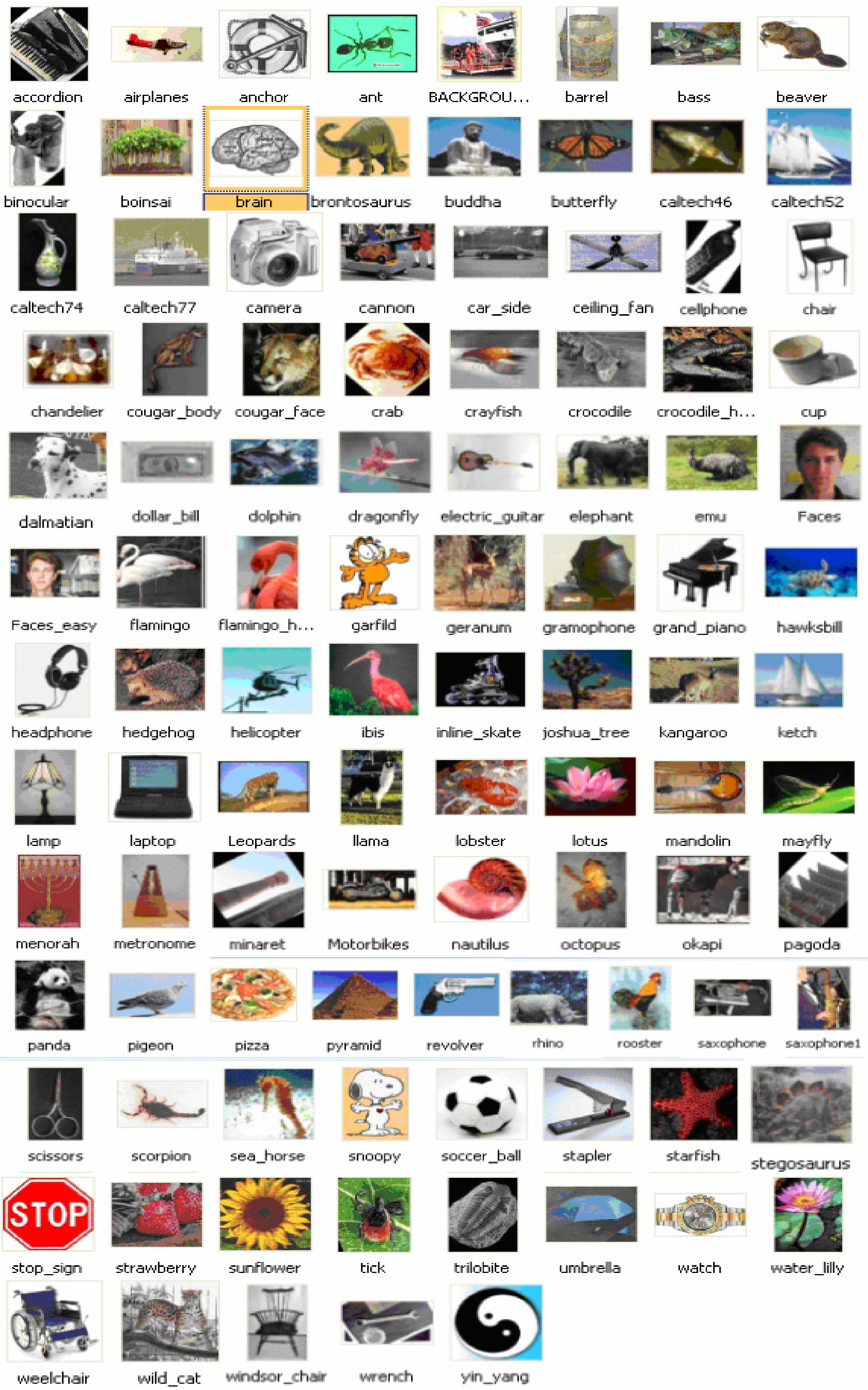


Figure A. 6: One sample for each category out of 101 categories in the Caltech 101 image dataset.

Table A. 2: List of 101 available categories

Accordion
Airplanes
Anchor
Ant
Background
Google
Barrel
Bass
Beaver
Binocular
Bonsai
Brain
Brontosaurus
Buddha
Butterfly
Camera
Cannon
Car side
Ceiling fan
Cellphone
Chair
Chandelier
Cougar body
Cougar face
Crab
Crayfish
Crocodile
Crocodile head
Cup
Dalmatian
Dollar bill
Dolphin
Dragonfly
Electric Guitar
Elephant
Emu
Euphonium
Ewer
Faces
Faces easy
Ferry
Flamingo
Flamingo head
Garfield
Gerenuk
Gramophone
Grand piano
Hawksbill
Headphone
Hedgehog
Helicopter
Ibis
Inline skate
Joshua tree
Kangaroo
Ketch
Lamp
Laptop
Leopards
Llama
Lobster
Lotus
Mandolin
Mayfly
Menorah
Metronome
Minaret
Motorbikes
Snoopy
Nautilus
Octopus
Okapi
Pagoda
Panda
Pigeon
Pizza
Platypus
Pyramid
Revolver
Rhino
Rooster
Saxophone
Schooner
Scissors
Scorpion
Sea horse
Soccer ball
Stapler
Starfish
Stegosaurus
Stop sign
Strawberry
Sunflower
Tick
Trilobite
Umbrella
Watch
Water lilly
Wheelchair
Wild cat
Windsor chair
Wrench
Yin Yang

APPENDIX B : Mathematical prerequisites

In this chapter, some necessary mathematical results are introduced. They are sufficiently standard not to be put in the actual chapters.

Definition B.1: A metric space is a set X together with a function $d: X \times X \to [0, \infty)$ which satisfies the following axioms:

(1) Self-identity: $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$;

(2) Symmetry: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$, $\forall \mathbf{x}, \mathbf{y} \in X$

(3) Triangular inequality: $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$, $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in X$

Then d is a metric on X, the pair (X, d) is called metric space and $d(\mathbf{x}, \mathbf{y})$ is a distance between.

Definition B.2: Consider a finite number of metric spaces $(X_i, d_i), 1 \leq i \leq n$, and let set X be a Cartesian product of individual sets X_i, $\prod_{i=1}^{n} X_i$. Points in the set X are denoted as $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ and $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$ with $\mathbf{x}_i, \mathbf{y}_i \in X_i$, then a new metric on the set X is defined as:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} d_i(\mathbf{x}_i, \mathbf{y}_i)$$

Definition B.3: Let X be a vector space over $\mathbb{R}$. A norm is a function $\|\cdot\| : X \to \mathbb{R}$ having the following properties:

(1) $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$;

(2) $\|\alpha\cdot\mathbf{x}\|=|\alpha|\cdot\|\mathbf{x}\|$ for all $\mathbf{x}\in X$ and $\alpha\in\mathbb{R}$,

(3) $\|\mathbf{x}+\mathbf{y}\|\le\|\mathbf{x}\|+\|\mathbf{y}\|$ for all $\mathbf{x},\mathbf{y}\in X$.

The pair $(\mathbf{x},\|\cdot\|)$ is called a normed vector space.

Proposition B.4: Let X be a normed space, any norm defines a metric d on X as $d(\mathbf{x},\mathbf{y})=\|\mathbf{x}-\mathbf{y}\|$.

Definition B.5 (Convex set): A set X in a vector space is convex if for any $\mathbf{x},\mathbf{x}'\in X$ and any $\lambda\in[0,1]$, $\lambda\mathbf{x}+(1-\lambda)\mathbf{x}'\in X$.

Definition B.6 (Convex Functions): A function f defined on a set X is convex for any $\mathbf{x},\mathbf{x}'\in X$ and any $\lambda\in[0,1]$ such that $\lambda\mathbf{x}+(1-\lambda)\mathbf{x}'\in X$ and

$$f(\lambda\mathbf{x}+(1-\lambda)\mathbf{x}')\le\lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{x}')$$

A function is strictly convex if for $\mathbf{x}\neq\mathbf{x}'$ and $\lambda\in(0,1)$ strict inequality holds.

Theorem B.7 (Minima on the Convex set): If a convex function $f:\mathbb{R}\to\mathbf{X}$ has a minimum on a convex set $X\subset\mathbf{X}$ then arguments $\mathbf{x}$ for which the minimal values are obtained, form a convex set. If f is strictly convex, than the set will contain only one element.

Corollary B.8 (Constrained Convex Minimization): Given the set of convex functions $f,c_1,\ldots,c_m$ on convex set $\mathbf{X}$, the problem

$$\min_{\mathbf{x}} f(\mathbf{x})\,\text{subject to } c_i(\mathbf{x})\le 0, i=1,...,m$$

has as its solution a convex set, if a solution exists. This solution is unique if f is strictly convex.

Definition B.9 (KKT conditions): For the following optimisation problem with inequality constraints:

$$\max_{\mathbf{x}} f(\mathbf{x}) \text{ subject to } g_i(\mathbf{x})\le c_i \text{ for } i=1,...,m .$$

The Karush-Kuhn-Tucker conditions are $L_j'(\mathbf{x})=0,$ for $j=1,...,n$, $\lambda_i\ge 0$, $g_i(\mathbf{x})\le c_i$ and $\lambda_i[g_i(\mathbf{x})-c_i]=0$ for $i=1,...,m$. Where $L(\mathbf{x})=f(\mathbf{x})-\sum_{i=1}^{m}\lambda_i(g_i(\mathbf{x})-c_i).$ is a Lagrangian.

Definition B.10: Inner product space is a vector space X over real numbers $\mathbb{R}$ if there exist a bilinear (linear in each argument), symmetric, positive definite scalar value product $\langle \cdot,\cdot \rangle$, for vectors $\mathbf{x},\mathbf{y},\mathbf{z} \in X$ and scalars $\alpha,\beta \in \mathbb{R}$:

(1) $\langle \alpha\mathbf{x}+\beta\mathbf{y},\mathbf{z} \rangle = \alpha\langle \mathbf{x},\mathbf{y} \rangle + \beta\langle \mathbf{y},\mathbf{z} \rangle$

(2) $\langle \mathbf{x},\mathbf{y} \rangle = \langle \mathbf{y},\mathbf{x} \rangle$

(3) $\langle \mathbf{x},\mathbf{x} \rangle \geq 0$

Additionally for $\langle \mathbf{x},\mathbf{x} \rangle = 0$ if $\mathbf{x}=0$ we have a strict inner product, dot or scalar product.

Definition B.11 (Hilbert Space) : A Hilbert space H is any linear space with an inner product being separable and complete with corresponding norm.

Completeness is expressed as convergence of any Cauchy sequence of elements $\{\mathbf{x}_i\}_{i\in\mathbb{N}} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)$ in a normed space, where Cauchy sequence is the one satisfying the following property:

$$\sup_{j>i} \left\| \mathbf{x}_i - \mathbf{x}_j \right\| \to 0, \text{ for } n \to \infty .$$

Separable space H assumes a finite set of elements $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N$ for $\varepsilon > 0$, so that for all $\mathbf{x} \in H$ $\min_i \left\| \mathbf{x}_i - \mathbf{x} \right\| < \varepsilon$.

Example B.12: Let $L_2(X)$ be a Hilbert space of square integrable functions on $X \subset \mathbb{R}^N$, with completeness defined as $L_2(X) = \left\{ f : \int_X f(x)^2 dx < \infty \right\}$. For $f,g \in X$ the inner product is defined as $\langle f,g \rangle = \int_X f(x)g(x)dx$.

Definition B.13 (σ-Algebra): A σ-algebra over a non-empty set $\boldsymbol{\mathcal{X}}$ is a non-empty collection of subsets $\boldsymbol{\mathcal{C}}$, closed under complements and countable infinite unions:

(1) Set $\boldsymbol{\mathcal{X}}$ and an empty-set are elements of any σ-algebra $\boldsymbol{\mathcal{C}}$ over $\boldsymbol{\mathcal{X}}$.

(2) If set $C \in \boldsymbol{\mathcal{C}}$ so is $\overline{C}$, compliment of C.

(3) A union (intersection) of countable many sets in C is also in C.

Definition B.14: A measure μ, is a function from σ-algebra $\mathcal{C}$ on $\mathcal{X}$ which assigns a real number to subsets of $\mathcal{X}$, $\mu : \mathcal{C} \to [0, \infty]$, such that the two following properties hold:

(1) Measure of an empty set is zero, $\mu(\varnothing) = 0$.

(2) Measure of any finite or countable infinite union of all mutually disjoint sets $C_1, C_2, \ldots \in \mathcal{C}$ is equal to the sum of the measures on these sets, σ-additivity: $\mu(\bigcup_{i=1}^{\infty} C_i) = \sum_{i=1}^{\infty} \mu(C_i)$

Definition B.15: A function f is measurable if the inverse function (pre-image) on any closed real number interval is in σ-algebra $\mathcal{C}$ on $\mathcal{X}$.

Definition B.16: The integral of a measurable function f on a set $\mathcal{C}$ is denoted as

$$\int_{\mathcal{X}} f(x) d\mu(x) .$$

APPENDIX C : Algorithm for Sequential Minimal Optimization

SVM optimality conditions

As mentioned in Chapter 3, the goal of a SVM is to maximize the objective function from (3.24):

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

subject to $y_i((\mathbf{w}\cdot\Phi(\mathbf{x}_i)+b) \geq 1-\xi_i,\, \xi_i \geq 0\ i=1,..,m$.

Sequential Minimal optimisation is a method proposed by Platt (1999a). In each iteration a quadratic problem of size two is solved, since this can be done analytically there is no need for a quadratic optimizer. However the problem is how to choose a good pair of variable to optimize in each iteration. The algorithm explained in detail in this appendix is an improvement from Keerthi et al. (2001) used for solving the SVM optimisation problem, and implemented for the RF approach in this thesis.

As mentioned the Lagrangian of this problem is:

$$L(\mathbf{w},b,\xi_i,\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i(y_i\cdot((\mathbf{w}\cdot\Phi(\mathbf{x}_i))+b)-1-\xi_i) - \sum_{i=1}^{m}\beta_i\xi_i\ ,$$

Based on the KKT conditions (Appendix B, Definition B.9):

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \Phi(\mathbf{x}_i)\,,\ \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{m}\alpha_i y_i = 0\,,$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \beta_i = 0\,,\; i = 1,...m$$

$$\alpha_i \geq 0\,,\; \sum_{i=1}^{m} \alpha_i (y_i \cdot ((\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) - 1 - \xi_i) \text{ and} \beta_i \geq 0\,,\; \beta_i \xi_i = 0$$

Hence the dual problem is now:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\mathbf{w}^T\mathbf{w} = \max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{C.1}$$

$$0 \leq \alpha_i \leq C,\; i = 1,...,m \tag{C.2}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{C.3}$$

The numerical approach is to solve the dual instead of the primal problem since it is a finite dimensional optimisation problem. In order to obtain proper stopping criteria for the dual optimisation problem an optimisation of the dual was proposed by Keerthi et al. (2001). The Lagrangian for the dual problem is:

$$\tilde{L} = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \delta_i \alpha_i - \sum_{i=1}^{m} \mu_i (C - \alpha_i) - \rho \sum_{i=1}^{m} \alpha_i y_i\,,$$

$$F_i = \mathbf{w} \cdot \Phi(\mathbf{x}_i) - y_i = \sum_{j=1}^{m} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i) - y_i \tag{C.4}$$

The KKT conditions are now:

$$\frac{d\tilde{L}}{d\alpha_i} = (F_i - \rho) y_i - \delta_i + \mu_i = 0\,,\; \delta_i \geq 0\,,\; \delta_i \alpha_i = 0\,, \mu_i \geq 0\,,\; \mu_i (C - \alpha_i) = 0\,,\; i = 1,...,m\,.$$

Three cases can be distinguished:

(1) $\alpha_i = 0$ and $\delta_i > 0, \mu_i = 0 \Rightarrow y_i(F_i - \rho) \geq 0$

(2) $0 < \alpha_i < C$ and $\delta_i > 0, \mu_i = 0 \Rightarrow y_i(F_i - \rho) = 0$

(3) $\alpha_i = C$ and $\delta_i = 0$, $\mu_i \geq 0 \Rightarrow y_i(F_i - \rho) \leq 0$

A number of index sets can be defined for α: $I_0 = \{i : 0 < \alpha_i < C\}$, $I_1 = \{i : y_i = 1, \alpha_i = 0\}$, $I_2 = \{i : y_i = -1, \alpha_i = C\}$, $I_3 = \{i : y_i = 1, \alpha_i = C\}$ and $I_4 = \{i : y_i = -1, \alpha_i = 0\}$.

In case KKT conditions are satisfied the three conditions above can now be re- written as:

$$\forall i \in I_0 \cup I_1 \cup I_2, F_i \geq \rho$$
$$\forall i \in I_0 \cup I_3 \cup I_4, F_i \leq \rho$$

Hence an upper and lower limit on value F_i can be defined through:

$$b_{up} = \min\{F_i : i \in I_0 \cup I_1 \cup I_2\}$$
$$b_{low} = \max\{F_i : i \in I_0 \cup I_3 \cup I_4\}$$

The KKT conditions are now simply $b_{up} \geq b_{low}$, and the optimality multiplication ρ equals to bias b, and can be placed halfway between b_{low} and b_{up}. Violation of the conditions is defined as:

$$i \in I_0 \cup I_1 \cup I_2,\ j \in I_0 \cup I_3 \cup I_4 \text{ and } F_i < F_j$$
$$i \in I_0 \cup I_3 \cup I_4,\ j \in I_0 \cup I_1 \cup I_2 \text{ and } F_i > F_j$$

Optimizing a quadratic problem of size two

It is assumed that the two coefficients that are result of the current optimisation step are denoted as $\alpha_1^{new}, \alpha_2^{new}$ their previous values are $\alpha_1^{old}, \alpha_2^{old}$ with the rest of the coefficients $\alpha_3, \ldots, \alpha_m$ fixed.

The following equation $\sum_{i=1}^{m} \alpha_i y_i = 0$ implies that :

$$y_1\alpha_1^{new} + y_2\alpha_2^{new} = y_1\alpha_1^{old} + y_2\alpha_2^{old} = const. \tag{C.5}$$

Leading to line optimisation (Platt, 1999a):

$$s = y_1 \cdot y_2 = \begin{cases} 1, y_1 = y_2 \\ -1, y_1 \neq y_2 \end{cases},\ \alpha_1^{new} + s\alpha_2^{new} = \alpha_1^{old} + s\alpha_2^{old} = \gamma = const. \tag{C.6}$$

Since the rest of the coefficients are fixed, in this current iteration they are constant hence (C.1) is now:

$$\begin{aligned} &L(\boldsymbol{\alpha}) = \alpha_1^{new} + \alpha_2^{new} + const. \\ &-\frac{1}{2}[y_1y_1\Phi(\mathbf{x}_1)\Phi(\mathbf{x}_1)(\alpha_1^{new})^2 + y_2y_2\Phi(\mathbf{x}_2)\Phi(\mathbf{x}_2)(\alpha_2^{new})^2 + 2y_1y_2\Phi(\mathbf{x}_1)\Phi(\mathbf{x}_2)\alpha_1^{new}\alpha_2^{new} \\ &+2\left(\sum_{i=3}^{m} \alpha_i y_i \Phi(\mathbf{x}_i)\right) \cdot (y_1\Phi(\mathbf{x}_1)\alpha_1^{new} + y_2\Phi(\mathbf{x}_2)\alpha_2^{new}) + const.] \end{aligned}$$

Appropriate kernels are denoted as $K_{11} = \Phi(\mathbf{x}_1)\Phi(\mathbf{x}_1)$, $K_{22} = \Phi(\mathbf{x}_2)\Phi(\mathbf{x}_2)$ and $K_{12} = \Phi(\mathbf{x}_1)\Phi(\mathbf{x}_2)$. Based on $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \Phi(\mathbf{x}_i)$ part of the above expression can be formulated in the following way:

$$\begin{aligned} v_j &\equiv (\sum_{i=3}^{m} \alpha_i y_i \Phi(\mathbf{x}_i))\Phi(\mathbf{x}_j) = (\sum_{i=1}^{m} \alpha_i y_i \Phi(\mathbf{x}_i) - \alpha_1 y_1 \Phi(\mathbf{x}_1) - \alpha_2 y_2 \Phi(\mathbf{x}_2))\Phi(\mathbf{x}_j) \\ &= (\Phi(\mathbf{x}_j)\mathbf{w}^{old} + b^{old}) - b^{old} - \alpha_1^{old} y_1 \Phi(\mathbf{x}_1)\Phi(\mathbf{x}_j) - \alpha_2^{old} y_2 \Phi(\mathbf{x}_2)\Phi(\mathbf{x}_j) \\ &= u_j^{old} - b^{old} - \alpha_1^{old} y_1 \Phi(\mathbf{x}_1)\Phi(\mathbf{x}_j) - \alpha_2^{old} y_2 \Phi(\mathbf{x}_2)\Phi(\mathbf{x}_j) \end{aligned}$$

The predicted output for input pattern $\mathbf{x}_j$ with respect to the values from the previous iteration can be expressed as:

$$u_j^{old} = \Phi(\mathbf{x}_j)\mathbf{w}^{old} + b^{old} \tag{C.7}$$

Incorporating this relation into the expression for Lagrange leads to:

$$\begin{aligned} L(\boldsymbol{\alpha}) = \alpha_1^{new} + \alpha_2^{new} - \frac{1}{2}[K_{11}(\alpha_1^{new})^2 + K_{22}(\alpha_2^{new})^2 + 2sK_{12}\alpha_1^{new}\alpha_2^{new} \\ + 2y_1 v_1 \alpha_1^{new} + y_2 v_2 \alpha_2^{new}) + const. \end{aligned}$$

Now the expression in (C.6) can be used to further express the Lagrange as a function of only α_2^{new} coefficient:

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \gamma - s\alpha_2^{new} + \alpha_2^{new} - \frac{1}{2}[K_{11}(\gamma - s\alpha_2^{new})^2 + K_{22}(\alpha_2^{new})^2 + 2sK_{12}(\gamma - s\alpha_2^{new})\alpha_2^{new} \\ &\quad + 2y_1 v_1 (\gamma - s\alpha_2^{new}) + y_2 v_2 \alpha_2^{new}] + const. \\ &\vdots \\ L(\boldsymbol{\alpha}) &= \frac{1}{2}(2K_{12} - K_{11} - K_{22}) \cdot (\alpha_2^{new})^2 \\ &\quad + (1 - s + sK_{11}\gamma - sK_{12}\gamma + y_2 v_1 + y_1 v_2)\alpha_2^{new} + const. \end{aligned}$$

Multiplier of α_2^{new} is now:

$$
\begin{aligned}
&1-s+sK_{11}\gamma-sK_{12}\gamma+y_2v_1-y_2v\\
&=1-s+sK_{11}(\alpha_1^{old}+s\alpha_2^{old})-sK_{12}(\alpha_1^{old}+s\alpha_2^{old})\\
&+y_2(u_1^{old}-b^{old}-\alpha_1^{old}y_1K_{11}-\alpha_2^{old}y_2K_{12})-y_2(u_2^{old}-b^{old}-\alpha_1^{old}y_1K_{12}-\alpha_2^{old}y_2K_{22})\\
&\vdots\\
&=y_2^2-y_1y_2+\underbrace{(K_{11}-2K_{12}+K_{22})}_{-\eta}\alpha_2^{old}+y_2(u_1^{old}-u_2^{old})\\
&=y_2((u_1^{old}-y_1+b^{old})-(u_2^{old}-y_2+b^{old}))-\eta\alpha_2^{old}
\end{aligned}
$$

Predicted output u_j^{old} by the SVM for pattern $\mathbf{x}_j$ based on values in the previous iteration is given with (C.7), y_i represents the real label and the prediction error is:

$$u_i^{old}-y_i=\Phi(\mathbf{x}_j)\mathbf{w}^{old}+b^{old}-y_i=F_i^{old}+b^{old}, i=1,2. \qquad \text{(C.8)}$$

Hence the objective function can be expressed in the following form:

$$L(\boldsymbol{\alpha})=\frac{1}{2}\eta(\alpha_2^{new})^2+(y_2(F_1^{old}-F_2^{old})-\eta\alpha_2^{old})\alpha_2^{new}+const.$$

The first and second derivatives are:

$$
\begin{aligned}
&\frac{dL(\boldsymbol{\alpha})}{d\alpha_2^{new}}=\eta\alpha_2^{new}+(y_2(F_1^{old}-F_2^{old})-\eta\alpha_2^{old}),\\
&\frac{dL(\boldsymbol{\alpha})}{d\alpha_2^{new}}=0\Rightarrow\alpha_2^{new}=\alpha_2^{old}-\frac{y_2(F_2^{old}-F_1^{old})}{\eta}
\end{aligned}
\qquad \text{(C.9)}
$$

$$\frac{d^2L(\boldsymbol{\alpha})}{d(\alpha_2^{new})^2}=\eta, \quad \eta=2K_{12}-K_{11}-K_{22}\le 0 \qquad \text{(C.10)}$$

In case $\eta<0$ the equation for α_2^{new} gives the unconstrained maximum point and this point must be checked if it belongs to the feasibility range.

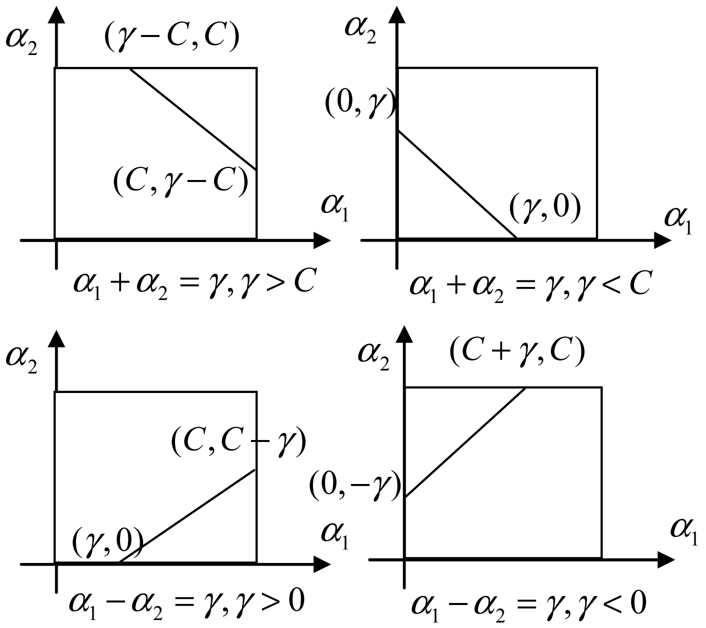


Figure C. 1: The two Lagrangian multiplier must lie within a box $0 \le \alpha_i \le C$, and at the same time fulfil the linear equality constraint (C.3) and lie on the diagonal line (Platt, 1999a).

Based on (C.6) and Figure C. 1 the range for α_2^{new} is determined as:

$$s=1,\ \alpha_1^{new}+\alpha_2^{new}=\gamma \Rightarrow \begin{cases} \gamma>C, \max\alpha_2^{new}=C \wedge \min\alpha_2^{new}=\gamma-C \\ \gamma<C, \quad \min\alpha_2^{new}=0 \wedge \max\alpha_2^{new}=\gamma \end{cases}$$

$$s=-1,\ \alpha_1^{new}-\alpha_2^{new}=\gamma \Rightarrow \begin{cases} \gamma>0, \min\alpha_2^{new}=0 \wedge \max\alpha_2^{new}=C-\gamma \\ \gamma<0, \quad \min\alpha_2^{new}=-\gamma \wedge \max\alpha_2^{new}=C \end{cases}$$

The optimisation step

Given $\alpha_1^{old},\alpha_2^{old}$ and appropriate $y_1,y_2,K_{11},K_{12},K_{22}$, $F_2^{old}-F_1^{old}$, the two Lagrangian coefficients are optimized based on values of η from (C.10):

- If $\eta<0$, $\alpha_2^{old}-\alpha_2^{new}=\Delta\alpha_2=\dfrac{y_2(F_2^{old}-F_1^{old})}{\eta}$, the solution α_2^{new} is then clipped based on boundaries in Figure C. 1, and values for α_1^{new} obtained as: $\alpha_1^{old}-\alpha_1^{new}=\Delta\alpha_1=-s\Delta\alpha_2$.

- If $\eta = 0$ the objective function needs to be evaluated at the two endpoints for α_2 that are L, H, and α_2^{new} set to be the one with larger objective function value.

The final form of the objective function is denoted as follows:

$$L(\boldsymbol{\alpha}) = \frac{1}{2}\eta(\alpha_2^{new})^2 + (y_2(F_1^{old} - F_2^{old}) - \eta\alpha_2^{old})\alpha_2^{new} + const.$$

Updating after the optimisation step

After value for $\alpha_1^{new}, \alpha_2^{new}$ are obtained the improvement by Keerthi et al. (2001) avoids using the threshold b and compares two F_i's given with (C.4) while automatically selecting the second α_i for joint optimisation. The following values for F_i are updated and cashed for Lagrangian multipliers falling into the set I_0:

$$F_1^{new} = F_1^{old} + \Delta\alpha_1 y_1 K_{11} + \Delta\alpha_2 y_2 K_{12}$$
$$F_2^{new} = F_2^{old} + \Delta\alpha_1 y_1 K_{12} + \Delta\alpha_2 y_2 K_{22}.$$

Choosing examples violating KKT conditions

The first α_i is selected sequentially from all non boundary examples from index set I_0, $0 < \alpha_i < C$. If the first α_i violates KKT conditions it is compared with α_j values for $F_j = b_{low}$ or $F_j = b_{up}$. The selected α_j is the second needed Lagrangian coefficient. The second version of this approach assumes choosing the worst KKT condition violating pair, as Lagrangian coefficients, that is coefficients with one or both F_i set to b_{low} or b_{up}.

Authors Publications

1. Djordjevic, D., and Izquierdo, E. (2006). Relevance Feedback for Image Retrieval in a Structured Multi-Feature Space. *Submitted to IEEE Transactions on Circuits and Systems for Video Technology.*

2. Djordjevic, D., and Izquierdo, E. (2006). Kernel in structured multi-feature spaces for image retrieval. *Electronics Letter*s, 42(15), 856-857.

3. Dorado, A., Djordjevic, D., Pedrycz, W., and Izquierdo, E. (2006). Efficient image selection for concept learning. *IEE Proceedings on Vision, Image and Signal Processing* 153(3), 263-273

4. Izquierdo. E., and Djordjevic, D., (2006). Using Relevance Feedback to Bridge the Semantic Gap. *Lecture Notes in Computer Science*, Springer-Verlag 3877, February 2006, pp. 19-34.

5. Djordjevic, D., and Izquierdo, E. (2006). Empirical Analysis of Descriptor Spaces and Metrics for Image Classification. In: *Proceedings of the 5th European workshop on Image Analysis for Multimedia Interactive Services*, March 2006, Korea. pp. 13-16.

6. Djordjevic, D., and Izquierdo, E. (2006). Relevance Feedback for Image Retrieval in Structured Multi-Feature Spaces. In: *Proceeding of 2nd International Mobile Multimedia Communications Conference*, September 2006, Alghero, Sardinia.

7. Chandramouli, K., Djordjevic, D., and Izquierdo, E. (2006). Binary Particle Swarm and Fuzzy Inference for Image Classification. In: *Proceeding 3rd International Conference on Visual Information Engineering*, Sep. 2006, Bangalore, India.

8. Djordjevic, D., Dorado, A., Izquierdo, E., and Pedrycz, W. (2005) Concept-Oriented Sample Images Selection. In: *Proceedings of 6th European workshop on Image Analysis for Multimedia Interactive Services.*

9. Djordjevic, D., and Izquierdo, E. (2005) Relevance Feedback based on Content-based Image Retrieval Systems, an Overview and Analysis. In: *Proceedings of the International Conference on Computer as a Tool, Serbia & Montenegro*, Belgrade, November 22-24.

10. Djordjevic, D., Zhang, Q., and Izquierdo, E. (2005). Fusion of semantic and visual information for automatic indexing and annotation of key frame images. In: *IEE Proceeding of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, pp. 293-299.

11. Dorado, A., Djordjevic, D., Izquierdo, E., and Pedrycz, W., (2004). Supervised semantic scene classification based on low-level clustering and relevance feedback. In *Proceedings of the European Workshop on the Integration of Knowledge*, Semantics and Digital Media Technology, November 2004, pp. 181-188, 2004.

12. Sprljan, N., Djordjevic, D., and Izquierdo, E., (2004). Scalability Evaluation of Still Image Coders. In: *Proceeding of 5th European workshop on Image Analysis for Multimedia Interactive Services*, April 2004, Lisbon, p.88.

13. Djordjevic, D., and Izquierdo, E., (2004). Assessing Scalability Features in Still Image Coders. In: *Proceedings of Postgraduate Research Conference in Electronics, Photonics, Communications & Networks, and Computing Science*, pp. 220-221, University of Hertfordshire, UK, April, 2004.

References

aceMedia (2005). Annual public report. External Deliverable D.7.8. [online]. Available: URL http://www.acemedia.org/aceMedia [November 2005].

Aksoy, S., and Haralick, R. M. (2000). Probabilistic vs. geometric similarity measure for image retrieval. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, South Carolina.

Agarwal, S., and Roth, D. (2002). Learning a sparse representation for object detection. In: *Proceedings European Conference on Computer Vision* 4, pp. 113–130.

Aggarwal, C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behaviour of distance metrics in high dimensional space, In: *Proceedings of the 8th International Conference on Database Theory*, pp. 420–434.

Bhanu, B., and Dong, A. (2002). Concepts learning with fuzzy clustering and relevance feedback. *Engineering Applications Artificial Intelligence* 15, 123-138

Bartolini, I., Ciaccia, P., and Waas, F. (2001). Feedbackbypass: A new approach to interactive similarity query processing. In: *Proceedings of the 27th International Conference on Very Large Data Bases*, Italy, pp. 201–210.

Baxter, B. (1991). Conditionally positive functions and p-norm distance matrices. *Constructive Approximation* 7 (4), 427-440.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton U. Press.

Berg, C., Christensen, J.P.R., and Ressel, P. (1984). *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer-Verlag.

Bober, M. (2001). MPEG-7 Visual Shape Descriptors. *IEEE Transcriptions on Circuits Systems and Video Technology* 11(6), 716-719.

Borges P., Mayer J., Izquierdo E. (2006). Robust and Transparent Color Modulation for Text Data Hiding, *IEEE Transactions on Multimedia*, Volume 10, Issue 8, pp 1479-1489.

Boughhorbel, S., Tarel, J.-P., and Fleuret, F. (2004). Non-Mercer Kernels for SVM Object Recognition. In: *Proceedings of British Machine Vision Conference,* September 7-9, London, pp. 137-146.

Boughorbel, S., Tarel, J.-P., and Boujemaa, N. (2005). Generalized Histogram Intersection Kernel for Image Recognition. In: *Proceedings of IEEE International Conference on Image Processing*, September 11-14, Italy, 3, pp. 161-164.

Boutell, M., Luo, J., Shen, X., and Brown, C. (2004). Learning multi-label scene classification. *Pattern Recognition* 37(9), 1757-1771.

Burges, C.J.C (1996). Simplified support vector decision rules. In: *The 13th International Conference on Machine Learning*.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2, 121 – 167.

Burges, C. J. C., and Schoelkopf, B. (1997). Improving speed and accuracy of support vector learning machines. *In Advances in Neural Information Processing Systems*. MIT Press, pp. 375–381.

Calic J., Izquierdo E. (2001). Towards Real-Time Shot Detection in the MPEG Compressed Domain, *3rd Workshop on Image Analysis for Multimedia Interactive Services* (WIAMIS 2001), Tampere, 16-17 May 2001, pp 1-5.

Calic J., Izquierdo E. (2002). Temporal Segmentation of MPEG Video Streams, *EURASIP Journal on Applied Signal Processing*, Issue 6, pp 561-565.

Calic J., Sav S., Izquierdo E., Marlow S., Murphy N., O'Connor N. (2002). Temporal Video Segmentation for Real-Time Key Frame Extraction, *27th IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2002). Orlando, FL, 13-17 May 2002, Volume 4, pp 3632-3635.

Calic J., Izquierdo E. (2002). A Multiresolution Technique for Video Indexing and Retrieval, *2002 International Conference on Image Processing* (ICIP 2002), Rochester, NY, September 22-25, 2002, Volume 1, pp 952-955.

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

Chandramouli K., Izquierdo E. (2006). Image Classification using Self-Organising Feature Map and Particle Swarm Optimisation, *7th International Workshop on Image Analysis for Multimedia Interactive Services* (WIAMIS 2006). Incheon, Korea, 19-21 April 2006, pp 313-316.

Chandramouli K., Izquierdo E. (2006). Image Classification using Chaotic Particle Swarm Optimization, *13th IEEE International Conference on Image Processing* (ICIP 2006), Atlanta, Georgia, 8-11 October 2006, pp 3001-3004.

Chang, S.-F., Sikora, T., and Puri, A. (2001). Overview of the MPEG-7 Standard. *IEEE Transactions on Circuits Systems Video Technology* 11, 688-695.

Chang, E. Y., Li, B., Wu, G., and Goh, K. (2003). Statistical Learning for Effective visual image retrieval. In: *Proceedings of the IEEE International Conference on Image Processing*, pp. 609–612.

Chapelle, O., Haffner, P. and Vapnik, V. (1999). Support-vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks* 10(5), 1055–1064

Chen, Y., and Zhou, X.S., and Huang, T.S. (2001). One-class SVM for Learning in Image Retrieval. In: *Proceedings of the IEEE International Conference on Image Processing*, Thessaloniki, Greece, 7-10 October.

Ciocca, G., and Schettini, R. (1999). A relevance feedback mechanism for content-based image retrieval. *Information Processing and Management* 35, 605-632.

Corel Corporation (1990). Corel stock photo images. URL: http://www.corel.com

Cortes, C, and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning* 20(3), 273-297.

Cox, I. J., Miller, M. L., Omohundro, S. M., and Yianilos, P. N. (1998). An Optimized Interaction Strategy for Bayesian Relevance Feedback. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, pp. 553-558.

Cox, I. J., Miller, M. L., Minka, T. P., Papathomas, T., and Yianilos, P. N. (2000). The Bayesian image retrieval system, PicHunter: theory, implementation and psychophysical experiments. *IEEE Trans. on Image Processing* 9(1), 20- 37.

Csurka, G., Bray, C., Dance, C., and Fan, L. (2004). Visual Categorization with Bags of Keypoints. *Proceedings of 8th European Conf on Computer Vision*, pp. 59-74.

Devalois, R.L., Albrecht, D.G. and Thorell, L.G. (1982). Spatial -frequency selectivity of cells in macaque visual cortex. *Vision Research* 22, 545-559.

Djordjevic, D., Dorado, A., Izquierdo, E., and Pedrycz, W. (2005) Concept-Oriented Sample Images Selection. In: *Proceedings of 6th European workshop on Image Analysis for Multimedia Interactive Services*.

Djordjevic, D., and Izquierdo, E. (2005). Relevance Feedback based on Content-based Image Retrieval Systems, an Overview and Analysis. In: P*roceedings of the International Conference on Computer as a Tool*, Serbia & Montenegro, Belgrade, November 22-24.

Djordjevic, D., Zhang, Q., and Izquierdo, E. (2005). Fusion of semantic and visual information for automatic indexing and annotation of key frame images. In: *IEE Proceeding of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, pp. 293-299.

Djordjevic, D., and Izquierdo, E. (2006a). Empirical Analysis of Descriptor Spaces and Metrics for Image Classification. In: *Proceedings of the 5th European workshop on Image Analysis for Multimedia Interactive Services*, March 2006, Korea, pp. 13-16.

Djordjevic, D., and Izquierdo, E. (2006b). Kernel in structured multi-feature spaces for image retrieval. *Electronics Letters*, 42(15), 856-857.

Djordjevic, D., and Izquierdo, E. (2006c). Relevance Feedback for Image Retrieval in Structured Multi-Feature Spaces. In: *Proceeding of 2nd International Mobile Multimedia Communications Conference*, September 2006, Alghero, Sardinia.

Djordjevic, D., and Izquierdo, E. (2006d). Relevance Feedback for Image Retrieval in a Structured Multi-Feature Space. *Submitted to IEEE Transactions on Circuits and Systems for Video Technology*.

Dorado A., Izquierdo E. (2002). Fuzzy Color Signatures, *IEEE Proceedings International Conference on Image Processing* (ICIP 2002). Rochester, New York, 22-25 September 2002, Volume 1, pp 433-436.

Dorado A., Izquierdo E. (2003). Semi-Automatic Image Annotation Using Frequent Keyword Mining, *IEEE Proceedings 7th International Conference on Information Visualisation* (IV 2003). London, 16-18 July 2003, pp 532-535.

Dorado A., Izquierdo E. (2003). Semantic Labeling of Images Combining Color, Texture and Keywords, *IEEE Proceedings 10th International Conference on Image Processing* (ICIP 2003). Barcelona, 14-18 September 2003, pp 9-12.

Dorado, A., Djordjevic, D., Pedrycz, W., and Izquierdo, E. (2006). Efficient image selection for concept learning. I*EE Proceedings on Vision, Image and Signal Processing* 153(3), 263-273.

Duin, R., (2000). Classifiers in almost empty spaces. In: *Proceedings. of 15th International Conference on Pattern Recognition* 2, pp. 1-7.

Eidenberger, H. (2003). Distance measures for MPEG-7-based retrieval. In: *Proceedings of the 5th ACM SIGMM international Workshop on Multimedia information Retrieva*l , Berkeley, New York, ACM Press, pp. 130-137.

Eidenberger, H. (2004). Statistical analysis of content-based MPEG-7 descriptors for image retrieval. *Multimedia Systems* 10(2), 84-97.

Fei-Fei, L., Fergus, R., and Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: *Proceedings of IEEE Workshop on Generative-Model Based Vision*. URL: http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

Ferecatu, M., Crucianu, M. and, Boujemaa, N. (2004). Sample selection strategies for relevance feedback in region-based image retrieval. In: K. Aizawa, Y. Nakamura and S. Satoh, eds, *Proc. of the Pacific-Rim Conference on Multimedi*a, pp. 497 – 504.

Fergus, R., Perona, P., and Zisserman, A. (2003). Object Class Recognition by Unsupervised Scale-Invariant Learning. In: Pr*oceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2, pp. 264-271.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., and Yanker, P. (1995). Query by image and video content: The QBIC system. *IEEE Computer Magazine* 28(9), 23-32.

Forsyth, D., and Fleck, M. (1997). Finding people and animals by guided assembly. In: *Proceedings of the IEEE International Conference on Image Processing* 3, Santa Barbara, pp. 5-8.

Fournier, J., and Cord, M. (2002). Long-term similarity learning in content-based image retrieval. In: *Proceedings of the International Conference on Image Processing*, pp. 441-444.

Gevers, T., (2001). Color in Image Search Engines. In M.S. Lew (ed.) *Principles of Visual Information Retrieval*, Springer-Verlag.

Gevers, T., and Smeulders, A. (2004). Content-based image retrieval: An overview. In G. Medioni and S. B. Kang (eds.) *Emerging Topics in Computer Vision*. Prentice Hall.

Gevers, T., and Stokman, H. (2003). Robust histogram construction from color invariants for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(10)

Grauman, K., and Darrell, T. (2005). The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. In: P*roceedings of IEEE International Conference on Computer Vision* 2, pp. 1458-1464.

Guo, G.-D., Jain, A.K., Ma, W.-Y., and Zhang, H.-J. (2002). Learning Similarity Measure for Natural Image Retrieval with Relevance Feedback. *IEEE Transactions on Neural Networks* 13, 811- 820

Gupta, A., and Jain, R. (1997). Visual information retrieval. Communications on the ACM 40(5), 70–79.

Hanke M., Izquierdo E., März R. (1998). On Asymptotics in Case of Linear Index-2 Differential-Algebraic Equations, *SIAM Journal on Numerical Analysis 1998*, Volume 35, Issue 4, pp 1326-1346.

Haralick, R. (1979) Statistical and Structural Approach to Texture. *Proceedings of the IEEE* 67(5), pp. 786-804.

Haussler, D. (1999). Convolution kernels on discrete structures. In Technical Report UCS-CRL-99-10. University of California at Santa Cruz, Department of Computer Science.

Heesch, D., and Ruger, S. (2003) Performance boosting with three mouse clicks - Relevance Feedback for CBIR. In: *Proceedings of the 25th European Conference on Information Retrieval Research*, Springer-Verlag LNCS 2633, Pisa, Italy, 14-16 April 2003, pp 363-376.

Heinecke, J., Boujemaa, N., Crucianu, M., Herve, N., Houissa., H., Djordjevic, D., and Izquierdo., E. (2005). User query analysis and intelligent search and retrieval modules. aceMedia technical report, External Deliverable D.6.7, 06 Apr 2005.

Hong, P., and Huang, T.S. (2001). Spatial pattern discovering by learning the isomorphic subgraph from multiple attributed relation graphs. In S. Fourey, G. T. Herman and T. Y. Kong (eds.) *Electronic Notes in Theoretical Computer Science* 46. Elsevier.

Hong, P., Tian, Q. and Huang, T. S. (2000) Incorporate support vector machines to content-based image retrieval with relevant feedback. In: *Proceedings of the 7th IEEE International Conference on Image Processing*, pp. 750-753.

Huang, T.S., and Zhou, X. S. (2001) Image Retrieval with Relevance Feedback: From heuristic weight adjustment to optimal learning methods, In: *Proceedings of International Conference in Image Processing*, Thessaloniki, Greece.

Huijmans, D., and Sebe, N. (2003). Content based indexing performance: a class size normalized precision, recall, generality evaluation, In: *Proceedings of International Conference on Image Processing*, Barcelona, pp. 733 – 736.

Ishikawa, Y., Subramanya, R. and Faloutsos, C. (1998). Mind-Reader: Query databases through multiple examples. In: *Proceedings of the 24th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., pp. 218–227.

Izquierdo E., Ernst M. (1995). Motion/Disparity analysis for 3D-Video-Conference Applications, *1995 International Workshop on Stereoscopy and 3-Dimensional Imaging* (IWS3DI 1995). Santorini, Greece, September 1995.

Izquierdo E., Kruse S. (1998). Image Analysis for 3D Modeling, Rendering, and Virtual View Generation, *Elsevier Journal Computer Vision and Image Understanding*, 1998, Volume 71, Issue 2, pp 231-253.

Izquierdo E., Ohm J. (2000). Image-based rendering and 3D modeling: a complete framework, *Signal Processing: Image Communication*, Volume 15, Issue 10, 2000, pp 817-858.

Izquierdo E., Ghanbari M. (2002). Key Components for an Advanced Segmentation System, *IEEE Transactions on Multimedia*, Volume 4, Issue 1, pp 97-113.

Izquierdo E., Casas J., Leonardi R., Migliorati P., O'Connor N., Kompatsiaris I., Strintzis M. (2003). Advanced Content-Based Semantic Scene Analysis and Information Retrieval: The Schema Project, *4th European Workshop on Image Analysis for Multimedia Interactive Services* (WIAMIS 2003), World Scientific

Publishing, London, England, 9-11 April 2003, pp 519-528.

Izquierdo E., Guerra V. (2003). An Ill-Posed Operator for Secure Image Authentication, *IEEE Transactions on Circuits and Systems for Video Technology,* Volume 13, Issue 8, pp 842-852

Izquierdo., E., and Djordjevic, D., (2006). Using Relevance Feedback to Bridge the Semantic Gap. *Lecture Notes in Computer Science*, Springer-Verlag 3877, February 2006, pp. 19 – 34.

Jain, A., Murty, M., and Flynn, P. (1999). Data Clustering: A Review. *ACM Computing Surveys* 31(3), 264-323.

Jain, A., Duin, P., and Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(1), 4-37.

Jing, F., Li, M., Zhang, H.-J., and Zhang, B. (2004) Relevance Feedback in Region-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 14(5), May 2004.

Jing, F., Li, M., Zhang, H.-J., and Zhang, B. (2004). An Efficient and Effective Region-Based Image Retrieval Framework. *IEEE Transactions on Image Processing* 13(5), May 2004.

Kaufman, L., and Rousseeuw, P.J. (1987). Clustering by means of medoids. In Y. Dodge (ed.) *Statistical Data Analysis Based on the L1-Norm and Related Methods*. North-Holland, pp. 405-416.

Kay S., Izquierdo E. (2001). Robust Content Based Image Watermarking, *3rd Workshop on Image Analysis for Multimedia Interactive Services* (WIAMIS 2001), Tampere, Finland, 16-17 May 2001, pp 53-56.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., and Murthy, K.R.K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. N*eural Computation* 13, 637-649.

Kliegr T., Chandramouli K., Nemrava J., Svatek V., Izquierdo E. (2006) “Combining Image Captions and Visual Analysis for Image Concept Classification”, *9th International Workshop on Multimedia Data Mining,* Las Vegas, 24-27, pp 817.

Kompatsiaris, I., Avrithis Y., Hobson, P., and Strintzis, M.G. (2004). Integrating Knowledge, Semantics and Content for User-Centred Intelligent Media

Services: The aceMedia Project. In: *Proceedings of the European workshop on Image Analysis for Multimedia Interactive Services,* 23 April 2004, Lisboa, Portugal.

Kondor, R., and Jebara, T. (2003). A Kernel between Sets of Vectors. In: P*roceedings of International Conference on Machine Learning*, Washington DC, USA.

Koskela, M., Laaksonen, J., and Oja, E. (2004). Use of image Subsets in Image Retrieval with Self-Organizing Maps. In: *Proceedings for International Conference on Image and Video Retrieval,* Springer-Verlag LNCS series, 3115, pp. 508-516.

Kurita, T., and Kato, T. (1993). Learning of personal visual impression for image database systems. In: S*econd International Conference on Document Analysis and Recognition*, pp. 547–552.

Laaksonen, J., Koskela, M., and Oja, E. (1999). PicSOM: Self-Organizing Maps for Content-Based Image Retrieval. In: *Proceedings of INNS-IEEE International Joint Conference on Neural Networks*, Washington DC.

Leibe, B., and Schiele, B. (2003). Analyzing Appearance and Contour Based Methods for Object Categorization. In: *Proceedings, International Conference on Computer Vision and Pattern Recognition*, Madison, June 2003. URL: http://www.mis.informatik.tu-mstadt.de/Research/Projects/categorization/eth80-db.html

Li, B., Chang, E., and Wu, Y., (2003). Discovery of a perceptual distance function for measuring image similarity. *MultiMedia Systems* 8, 512–522.

Lim, J.H. (1999). Learnable visual keywords for image classification. In: *Proceedings of the Fourth ACM Conference on Digital Libraries* (Berkeley, California, United States, August 11 - 14, 1999). ACM Press, NY, pp. 139-145.

Lim, J.H., and Jesse, S. J. (2005). Combining intra-image and inter-class semantics for consumer image retrieval. *Pattern Recognition* 38(6), pp. 847-864.

Lim, J.H., Jin, J.S. (2003). Support regions and images for photo event retrieval. In: *Proceedings of the IEEE International Conference on Image Processing,* 2, pp. 515–518.

Long, F., Zhang, H.J., and Feng, D. (2002). *Fundamentals of Content-based Image* *retrieval, in Multimedia Information Retrieval and Management-Technological Fundamentals and Applications*. D. Feng, W.C. Siu, and H.J. Zhang. (ed.), Springer.

Lowe, D. (1999) Object recognition from local scale-invariant features. In: P*roceedings of the International Conference on Computer Vision*, pp. 1150–1157.

Lyu, S. (2005). Mercer Kernels for Object Recognition with Local Features. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 20-26, IEEE Computer Society, Washington DC, 2, pp. 223-229.

Lu, Y., Hu, C., Zhu, X., Zhang, H.-J., and Yang,Q. (2000). A unified framework for semantics and feature based relevance feedback in image retrieval systems. In: *Proceedings of the 8th ACM International Conference on Multimedia*, ACM Press, pp. 31–37.

MPEG (2001). MPEG-7 Visual Experimentation Model (XM), Version 10.0. ISO/IEC/JTC1/SC29/WG11, Doc. N4063, Mar. 2001.

Ma, W. Y., and Manjunath, B. (1997). NeTra: A toolbox for navigating large image databases. In: *Proceedings of IEEE International Conference on Image Processing*, pp. 568–571.

Manjunath, B.S. and Ma, W.Y. (1996). Texture Features for Browsing and Retrieval of Image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(8), 837-842.

Manjunath, B.S., Ohm, J.-R., Vasudevan, V.V., and Yamada, A. (2001) MPEG-7 Color and Texture Descriptors. *IEEE Trans. Circuits Systems and Video Technology* 11, 703-715.

Manjunath, B.S., Salembier, P., Sikora, T. (2003). I*ntroduction to MPEG-7, Multimedia Content Description Interface*. John Wiley and Sons.

Martinez, J.M. (2001). *Overview of the MPEG-7 Standard*. ISO/IEC JTC1/SC29/WG1.

Meilhac, C., and Nastar, C. (1999). Relevance feedback and category search in image databases. *Proceedings IEEE Int. Conference on Multimedia Computing and Systems*, Florence, Italy, 7-11 June, IEEE Computer Society, pp. 512–517.

Michalski, R., Stepp, R., and Diday, E. (1981). A Recent Advance in Data Analysis:

Clustering Objects into Classes characterized by Conjunctive Concepts. In Laveen N. Kanal and Azriel Rosenfeld (eds.) *Progress in Pattern Recognition 1*, New York: North-Holland, pp. 33-56.

Mikolajczyk, K., and Schmid, C. (2005). A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(10), 1615-1630.

Mokhtarian, F., Abbasi, S., and Kittler, J. (1996). Robust and Efficient Shape Indexing through Curvature Scale Space. In: *Proceedings of International Workshop on Image DataBases and MultiMedia Search*, Amsterdam, pp. 35-42.

Moreno, P., Ho, P., and Vasconcelos, N. (2003). A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. In *Advances in Neural Information Processing Systems* 16, Cambridge, MA, MIT Press.

Mrak M., Abhayaratne C. K., Izquierdo E. (2004). On the influence of motion vector precision limiting in scalable video coding, *7th International Conference on Signal Processing* (ICSP 2004). Beijing, China, 31 August - 4 September 2004, Volume 2, pp 1143-1146.

Mrak M., Sprljan N., Izquierdo E. (2006). Motion estimation in temporal subbands for quality scalable motion coding, *IET Electronics Letters*, Volume 41, Issue 19, pp 1050-1051.

Mrak M., Sprljan N., Zgaljic T., Ramzan N., Wan S., Izquierdo E. (2006). Performance evidence of software proposal for Wavelet Video Coding Exploration group, *Conference 76th MPEG Meeting*, 2006/4/7.

Mrówka E., Dorado A., Pedrycz W., Izquierdo E. (2004). Dimensionality Reduction for Content-Based Image Classification, *IEEE Proceedings 8th International Conference on Information Visualisation* (IV 2004). London, England, 14-16 July 2004, pp 435-438.

Muller, K., Mika, S., Ratsch, G., Tsuda, K., and Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks* 12(2), 181–202.

Nakazato, M., Dagli, C, and Huang, T.S. (2003). Evaluating group-based relevance feedback for content-based image retrieval. In: *Proceedings of the IEEE International Conference on Image Processing* 2, pp. 599-602.

Nastar, C., Mitschke, M., and Meilhac, C. (1998). Efficient query refinement for image retrieval. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Comp. Soc., pp. 547– 552.

Nene, S.A., Nayar, S.K, and Murase, H. (1996). Columbia Object Image Library (COIL-100). Technical Report CUCS-006-96, [online]. Available: URL http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php [Accessed May 2006].

Newsam, S., Sumengen, B., and Manjunath, B.S. (2001). Category-based Image Retrieval. In: *Proceedings of the IEEE International Conference on Image Processing, Special Session on Multimedia Indexing, Browsing and Retrieval* 3, pp. 596-599.

Newsam, S., and Kamath, C. (2005). Comparing shape and texture features for pattern recognition in simulation data. In: *Image Processing: Algorithms and Systems* IV, SPIE Electronic Imaging.

Nguyen, H., and Smeulders, A. (2004). Active learning using pre-clustering. In: *Proceedings of 21st International Conference on Machine Learning*.

O'Connor N., Sav S., Adamek T., Mezaris V., Kompatsiaris I., Lui Z., Izquierdo E., Bennström C., Casas J. (2003). Region and Object Segmentation Algorithms in the Qimera Segmentation Platform, *3rd Int. Workshop on Content-Based Multimedia Indexing* (CBMI 2003), Rennes, 22-24 September 2003, pp 1-8.

Ojala, T, Aittola, M., and Matinmikko, E. (2002). Empirical evaluation of MPEG-7 XM color descriptors in content-based retrieval of semantic image categories. In: *Proceedings of International Conference on Pattern Recognition*, pp. 1021-1024.

Osuna, E., Freund, R., and Girosi, F. (1997). An improved training algorithm for support vector machines. In: *Proceedings of Neural Networks for Signal Processing Conference*, pp. 276–284.

Pedrycz, W. (1990). Fuzzy sets pattern recognition: methodology and methods. *Pattern Recognition* 23 (1/2), pp. 121-146.

Pentland, A., Picard, R. W., and Sclaroff, S. (1996). Photobook: Content-based manipulation for image databases. *International Journal of Computer Vision* 18(3), 233–254.

Persoon, E., and Fu, K.S. (1977). Shape Discrimination Using Fourier Descriptors. *IEEE Transactions on Systems, Man, and Cybernetics* 7(3), 170-179.

Platt, J. (1999a). Fast training of support vector machines using sequential minimal optimization. In B. Scholkopf, C. J. C. Burges and A. J. Smola (eds.), *Advance in kernel methods -Support Vector Learning*, MIT Press, pp. 185-208.

Platt, J. (1999b). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (eds.) *Advances in Large Margin Classifiers*. MIT Press, pp. 61-74.

Picard, R.W., Minka, T.P., and Szummer, M. (1996). Modeling user subjectivity in image libraries. In: *IEEE International Conference on Image Processing*, pp. 777–780.

Pinheiro A., Izquierdo E., Ghanhari M. (2000) Shape Matching using a Curvature Based Polygonal Approximation in Scale-Space, *IEEE Proceedings International Conference on Image Processing* (ICIP 2000). Vancouver, BC, 10-13 September 2000, Volume 2, pp 538-541.

Porkaew, K., and Chakrabarti, K. (1999). Query refinement for multimedia similarity retrieval in MARS. In: *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, ACM Press, pp. 235–238.

Qian, F., Li, M., Zhang, L., Zhang, H., and Zhang, B. (2002). Gaussian mixture model for relevance feedback in image retrieval. In: *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 229- 232.

Quack, T., Monich, U., Thiele, L., and Manjunath, B.S. (2004). Cortina: a system for large-scale, content-based web image retrieval. In: *Proceedings of the 12th Annual ACM international Conference on Multimedia*.

Ramos J., Guil N., González J., Zapata E., Izquierdo E. (2006). Logotype detection to support semantic-based video annotation, *Journal Signal Processing: Image Communication*, Volume 22, Issue 7-8, pp 669-679

Ramzan N., Wan S., Izquierdo E., “Joint Source-Channel Coding for Wavelet-Based Scalable Video Transmission Using an Adaptive Turbo Code”, EURASIP Journal on Image and Video Processing, Volume 2007, Pages 1-12.

Ratan, A., Maron, O., Grimson, W.E.L., and Lozano-Perez, T. (1999). A framework for learning query concepts in image classification. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1423–1429.

Rijsbergen, C.J. van (1979). *Information retrieval*. 2nd ed., London, Butterworths.

Royden, H.L. (1988). *Real Analysis*. 3rd ed., NY., Macmillan Publishing Company.

Rui, Y., Huang, T.S., and Mehrotra, S. (1997). Content-based Image Retrieval with Relevance Feedback in MARS. In: *Proceedings of IEEE International Conference on Image Processing*, 26-29 October.

Rui, Y., Huang, T. S., Ortega, M. and Mehrotra, S. (1998). Relevance feedback: a power tool in interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 8(5), 644–654.

Rui, Y., and Huang, T.S. (2000). Optimizing learning in image retrieval. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* pp. 236-243.

Rubner, Y., Tomasi, C., and Guibas, L.J. (1998). A metric for distributions with applications to image databases. In: *Proceedings of IEEE International Conference on Computer Vision,* pp. 59-66.

Rubner, Y., Puzicha, J., Tomasi, C., and Buhmann, J.M. (2001). Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding* 84(1), 25-43.

Santini, S., and Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(9), 871-883

Schettini, R., Ciocca, G., and Gagliardi, I. (1999). Content-based color image retrieval with relevance feedback. In: *Proceedings of International Conference on Image Processing*, Kobe, Japan, pp. 75-79.

Sclaroff, S., Taycher, L., and Cascia, M.L. (1997). Imagerover: A content-based mage browser for the world wide web. In: *Proceedings of the Workshop on Content-Based Access of Image and Video Libraries*, IEEE computer Society, p. 2.

Scholkopf, B. (2000). The kernel trick for distances. In: *Advances in Neural Information Processing Systems* 12, MIT Press, pp. 301–307.

Scholkopf, B., Burges, C.J.C., and Smola, A.J. (1999) *Advances in kernel methods*

Support Vector Machines. MIT Press.

Schokopf, B., and Smola, A.J. (2002). *Learning with Kernels. Cambridge*, MA, MIT Press.

Scholkopf, B. (2000). Statistical Learning and Kernel Methods. Data Fusion and Perception Technical report No.(23), 3-24. Della Riccia, Lenz and Kruse, Springer, Redmond eds. WA.

Schmid, C., and Mohr, R. (1997). Local Greyvalue Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(5), 530–534

Shashua, A., and Hazan, T. (2005). Algebraic Set Kernels with Application to Inference over Local Image Representations. In Lawrence K. Saul, Yair Weiss and Leon Bottou (ed.). Advances in Neural Information Processing Systems. MIT Press, 17, pp. 1257-1264.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis.* Cambridge University Press, Published June 2004.

Sikora, T. (2001). The MPEG-7 Visual Standard for Content Description-An Overview. *IEEE Transactions on Circuits Systems and Video Technology* 11, 696-702.

Simou, N., Saathoff, C., Dasiopoulou, S., Spyrou, E., Voisine, N., Tzouvaras, V., Kompatsiaris, I., Avrithis, Y., and Staab., S. (2005). An Ontology Infrastructure for Multimedia Reasoning. In: *International Workshop* VLBV 2005, Sardinia, Italy, 15-16 September.

Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12), 1349–1380

Smith, J. R., and Chang, S.-F. (1996). VisualSEEk: A fully automated content-based query system, In: *Proceedings of the 4th ACM Conference on Multimedia*, pp. 87–98.

Smith, J. R., and Chang, S.-F. (1997). An Image and Video Search Engine for the World Wide Web. In: *Storage and Retrieval for Image and Video Databases* (SPIE), pp 84-94.

Sprljan N., Mrak M., Abhayaratne C., Izquierdo E. (2005). A Scalable Coding Framework for Efficient Video Adaptation, *6th International Workshop on*

Image Analysis for Multimedia Interactive Services (WIAMIS 2005). Montreux, Switzerland, 2005, 13-15 April 2005, pp 1-4.

Swain, M., and Ballard, D. (1991). Color indexing. *International Journal of Computer Vision* 7(1), 11–32

Tamura, H., Mori, S., and Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics* 8 (6), 460-473

Tian, Q., Hong, P., and Huang, T. S. (2000). Update relevant image weights for content-based image retrieval using support vector machines. In: *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, July 30 - Aug. 2, 2, pp.1199-1202.

Tong, S., and Chang, E. (2001). Support vector machine active learning for image retrieval. In: *Proceedings of the 9th ACM international conference on Multimedia*, ACM Press, pp. 107–118.

Tong, S., and Koller, D. (2000). Support vector machine active learning with applications to text classification. In: *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann, pp. 999-1006.

Town, C., and Sinclair, D. (2001). Content Based Image Retrieval using Semantic Visual Categories. Society for Manufacturing Engineers, *Technical Report MV01-211* (previously published as AT&T Laboratories Cambridge Technical Report TR2000-14)

Tsomko E., Kim H., Izquierdo E. (2006). Linear Gaussian blur evolution for detection of blurry images, *IET Image Processing*, Volume 4, Issue 4, pp 302-312.

Tuceryan, M., and Jain, A.K. (1988). *The Handbook of Pattern Recognition and Computer Vision: Texture Analysis*. 2nd ed. World Scientific Publishing Co., pp. 207-248

Vailaya, A., Jain, A., and Zhang, H.-J. (1998). On image classification: city vs. landscape. *Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara, CA, 21 June, pp. 3-8.

Vailaya, A., Figueiredo, M., Jain, A., and Zhang, H.-J. (2001). Image classification for content based indexing. *IEEE Transactions on Image Processing* 10(1), 117-130.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley.

Vapnik, V. (1999). An overview of statistical learning theory. I*EEE Transactions on Neural Networks* 10(5), 988-1000.

Vasconcelos, N., and Lippman, A. (2000). Bayesian Relevance Feedback for Content-Based Image Retrieval. In: *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, 12 June, IEEE Computer Society, Washington, DC, pp. 63 - 67.

VisText Database. (2002), [online]. Available: URL http://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html [Accessed May 2006].

Wallraven, C., Caputo, B., and Graf, A. (2003). Recognition with Local Features: the Kernel Recipe. In: *IEEE International Conference on Computer Vision*, pp. 257–264.

Wang Y., Izquierdo E. (2002). High-Capacity Data Hiding in MPEG-2 Compressed Video, *9th International Workshop on Systems, Signals and Image Processing* World Scientific, Manchester, England, 7-8 November 2002, pp 212-218.

Wang, Z., Li, J., and Wiederhold, G. (2001). Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(9), 947-963.

Wilkins P., Izquierdo E., et al. (2006). K-Space at TRECVid 2006, *4th TRECVID Workshop* (TRECVID 2006). Gaithersburg, Maryland, 13-14 November 2006

Wolf, L., and Shashua, A. (2003). Learning Over Sets Using Kernel Principal Angles. *Journal of Machine Learning Research* 4, 913–931.

Wilson, D.R., and Martnez, T.R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research* 6(1), 1-34.

Wu, Y., Tian, Q., and Huang, T. S. (2000). Integrating Unlabeled Images for Image Retrieval Based on Relevance Feedback. In: *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, pp. 21-24.

Yap, K.-H., and Wu, K. (2003). Fuzzy relevance feedback in content-based image retrieval. In: *Proceedings of International Conference on Information, and Signal Processing and Pacific-Rim Conference on Multimedia*, 3, pp. 1595-1599.

Yoshizawa, T., and Schweitzer, H. (2004). Long-term learning of semantic grouping from relevance-feedback. In: *Proceedings of. 6th ACM SIGMM International Workshop Multimedia Information*, pp. 165-172.

Zgaljic T., Sprljan N., Izquierdo E. (2005). Bitstream Syntax Description based Adaptation of Scalable Video, *2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology* (EWIMT 2005). London, England, 30 November - 1 December 2005.

Zgaljic T., Sprljan N., Izquierdo E. (2005). Bit-stream allocation methods for scalable video coding supporting wireless communications, *Signal Processing: Image Communication (SS on Mobile Video)*, Volume 22, Issue 3, pp 298-316.

Zhang, C. and Chen, T. (2002). An active learning framework for content based information retrieval. *IEEE Trans. Multimedia, Special Issue on Multimedia Database* 4(2), 260-268.

Zhang Q., Izquierdo E. (2006). A Multi-Feature Optimization Approach to Object-Based Image Classification", *5th International Conference Image and Video Retrieval* (CIVR 2006). Tempe, Arizona, 13-15 July 2006, pp 310-319.

Zhang, Q., and Izquierdo, E. (2006). Optimizing metrics combining low-level visual descriptors for image annotation and retrieval. In: *Proceedings on International Conference on Acoustics, Speech, and Signal Processin*g, May 14-19, Toulouse, France.

Zhou, X.S., and Huang, T.S. (2003). Relevance feedback for image retrieval: a comprehensive review. *ACM Multimedia Systems Journal*, special issue on CBIR 8(6), April 2003, 536-544.

Zhou, X.S., and Huang, T.S (2001a). Comparing discriminating transformations and SVM for learning during multimedia retrieval. In: *Proceedings of the 9th ACM international conference on Multimedia*, ACM Press, pp. 137–146.

Zhou, X.S., and Huang, T.S. (2001b). Small sample learning during multimedia retrieval using BiasMap. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1, pp. 11-17.