TKK Dissertations in Information and Computer Science
Espoo 2008 TKK-ICS-D9

# DISCRIMINATIVE LEARNING WITH APPLICATION TO INTERACTIVE FACIAL IMAGE RETRIEVAL

Zhirong Yang

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium TU2 at Helsinki University of Technology (Espoo, Finland) on the 14th of November, 2008, at 12 noon.

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O.Box 5400
FI-02015 TKK
FINLAND
URL: http://ics.tkk.fi
Tel. +358-9-451 1
Fax +358-9-451 3277
E-mail: series@ics.tkk.fi



ISBN 978-951-22-9625-5 (Print)
ISBN 978-951-22-9626-2 (Online)
ISSN 1797-5050 (Print)
ISSN 1797-5069 (Online)
URL: http://lib.tkk.fi/Diss/2008/isbn9789512296262/

Multiprint Oy
Espoo 2008

Yang, Z. (2008): **Discriminative learning with application to interactive facial image retrieval**. Doctoral thesis, Helsinki University of Technology, Dissertations in Information and Computer Science, TKK-ICS-D9, Espoo, Finland.

**Keywords:** Content-based image retrieval (CBIR), relevance feedback, Self-Organizing Map, discriminant analysis, facial image.

# ABSTRACT

The amount of digital images is growing drastically and advanced tools for searching in large image collections are therefore becoming urgently needed. Content-based image retrieval is advantageous for such a task in terms of automatic feature extraction and indexing without human labor and subjectivity in image annotations. The semantic gap between high-level semantics and low-level visual features can be reduced by the relevance feedback technique. However, most existing interactive content-based image retrieval (ICBIR) systems require a substantial amount of human evaluation labor, which leads to the evaluation fatigue problem that heavily restricts the application of ICBIR.

In this thesis a solution based on discriminative learning is presented. It extends an existing ICBIR system, PicSOM, towards practical applications. The enhanced ICBIR system allows users to input partial relevance which includes not only relevance extent but also relevance reason. A multi-phase retrieval with partial relevance can adapt to the user's searching intention in a from-coarse-to-fine manner.

The retrieval performance can be improved by employing supervised learning as a preprocessing step before unsupervised content-based indexing. In this work, Parzen Discriminant Analysis (PDA) is proposed to extract discriminative components from images. PDA regularizes the Informative Discriminant Analysis (IDA) objective with a greatly accelerated optimization algorithm. Moreover, discriminative Self-Organizing Maps trained with resulting features can easily handle fuzzy categorizations.

The proposed techniques have been applied to interactive facial image retrieval. Both a query example and a benchmark simulation study are presented, which indicate that the first image depicting the target subject can be retrieved in a small number of rounds.

# Preface

The work related to this thesis has been carried out in the Neural Networks Research Centre and the Adaptive Informatics Research Centre of the Laboratory of Computer and Information Science (Department of Information and Computer Science (ICS) since 2008) at the Helsinki University of Technology. The research has been mainly funded by the Helsinki Graduate School in Computer Science and Engineering. Additional funding has come from projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *Finnish Center of Excellence in Adaptive Informatics Research* supported by the Academy of Finland.

First of all, I would like to express my gratitude to my instructor and the manager of our research project, D.Sc. Jorma Laaksonen, for exquisite daily guidance, teaching me a lot about writing practice, and bearing the main responsibility for the PicSOM programming. I am obliged to my supervisor, Prof. Erkki Oja, for lasting comments and support throughout my work, and also for expert commenting on the content of this thesis. Special thanks go to the pre-examiners of my thesis, D.Sc. Sami Brandt and D.Sc. Joni Kämäräinen, who truly made my thesis much better with their thorough examination and offering their knowledge through a multitude of very insightful comments.

I wish to thank all (present and former) members of our research group, especially D.Sc. Markus Koskela, Mats Sjöberg, and Ville Viitaniemi. I am also grateful to many other co-workers in the ICS department for creating a nurturing academic atmosphere to work in.

Last but certainly not least I would like to thank the most important thing in my life, my family. My parents and my sister have been supporting my aspirations all the way through and have done their best to help me in any way they can. Finally, I am deeply indebted to my beloved wife, Yanyi, — without your support this would not have been possible.

Otaniemi, October 2008

Zhirong Yang

# Contents

# List of symbols

| | |
|---|---|
| $\mathbf{x}$, $\mathbf{x}_i$ | samples in the original data space |
| $\mathbf{y}$, $\mathbf{y}_i$ | samples in the projected space |
| $c(\mathbf{x}_i)$ | class label of $\mathbf{x}_i$ |
| $c_i$ | class label of the $i$-th sample |
| $N$ | total number of samples |
| $N_c$ | number of samples in the $c$-th class |
| $M$ | dimensionality of the original data space |
| $R$ | dimensionality of the projected space |
| $Q$ | number of classes |
| $K$ | number of subjects |
| $\tau$ | number of units in an SOM |
| $\mathbf{W}$ | projection matrix |
| $\lVert \cdot \rVert$ | Euclidean norm |
| $\lVert \cdot \rVert_{\mathbf{G}}$ | Riemannian norm with respect to metric $\mathbf{G}$ |
| $\mathbf{G}$ | Riemannian metric matrix |
| $\hat{\mathbf{G}}$ | approximation of $\mathbf{G}$ with only principal components |
| $[\mathbf{A}]_{ij} = A_{ij}$ | the $(i, j)$-th entry of matrix $\mathbf{A}$ |
| $\mathcal{J}$ | main objective function |
| $\mathcal{F}$ | regularization objection function |
| $\mathbf{S}$ | (total) scatter matrix of all data |
| $\mathbf{S}_W$ | within-class scatter matrix |
| $\mathbf{S}_B$ | between-class scatter matrix |
| $\boldsymbol{\mu}$ | empirical mean of all samples |
| $\boldsymbol{\mu}_c$ | empirical mean of samples in the $c$-th class |
| $\mathcal{I}_c$ | the index set of the $c$-th class |
| $\lambda$ | tradeoff parameter in Parzen Discriminant Analysis |
| $\boldsymbol{\Omega}$ | smoothing matrix in Parzen Discriminant Analysis regularization |
| $p(\mathbf{x}; \boldsymbol{\theta})$ | probability density of $\mathbf{x}$ given the parameter $\boldsymbol{\theta}$ |

| | |
|---|---|
| $\ell(\mathbf{x};\boldsymbol{\theta})$ | equal to $-\log p(\mathbf{x};\boldsymbol{\theta})$ |
| $\mathcal{N}(x;\rho)$ | zero-mean normal density function with variance $\rho$ |
| $\delta(a,b)$ | Kronecker delta, equal to 1 if $a=b$ and 0 otherwise |
| $\phi_{ij}$ | shorthand notation for $\delta(c_i,c_j)$ |
| $e_{ij}$ | radial basis kernel between the $i$-th and $j$-th projected samples, $e_{ii}=0$ |
| $\sigma$ | kernel width in radial basis kernel |
| $\mathbf{M}$ | sample-class probability contingency matrix |
| $\mathbf{H}$ | weights for the pairwise differences |
| $\mathbf{B}$ | unsupervised weights for pairwise differences |
| $\mathbf{B}^k$ | supervised weights for the pairwise differences in the $k$-th class |
| $\eta$ | learning rate |
| $\mathrm{diag}(\mathbf{a})$ | creates a diagonal matrix $\mathbf{A}$ from a vector $\mathbf{a}$ with $A_{ii}=a_i$ |
| $\circ$ | elementwise product |
| $j_t$ | position of first hit of the $t$-th subject with random retrieval |
| $\varrho$ | mean position of first subject hit with random retrieval |
| $i_t$ | position of first subject hit with an improved retrieval |
| $\tilde{\varrho}$ | first subject hit advantage |
| $r[j]$ | cumulative number of relevant images retrieved after $j$ rounds |
| $n[j]$ | cumulative number of total images retrieved after $j$ rounds |
| $s$ | number of relevant images or number of images of the target subject |

# List of abbreviations

| | |
|---|---|
| BMU | Best-matching unit |
| CBIR | Content-based image retrieval |
| DSOM | Discriminative self-organizing map |
| FRHA | First relevant hit advantage |
| FSHA | First subject hit advantage |
| ICA | Independent component analysis |
| ICBIR | Interactive content-based image retrieval |
| IDA | Informative discriminant analysis |
| IFIR | Interactive facial image retrieval |
| IR | Information retrieval |
| LBP | Local binary pattern |
| LDA | Linear discriminant analysis |
| LSI | Latent semantic index |
| NCA | Neighborhood component analysis |
| NMF | Non-negative matrix factorization |
| P-NMF | Projective non-negative matrix factorization |
| PCA | Principal component analysis |
| PDA | Parzen discriminant analysis |
| PWG | Principal whitened gradient |
| RF | Relevance feedback |
| SIFT | Scale Invariant Feature Transform |
| SOM | Self-organizing map |
| SVM | Support vector machine |
| TS-SOM | Tree structured self-organizing map |

# Chapter 1

# Introduction

## 1.1 Goals and scope of the thesis

*Content-Based Image Retrieval* (CBIR) addresses the problem of finding images relevant to the users' information needs from image databases. The indexing is principally based on low-level visual features for which automatic extraction methods are available. Due to the semantic gap, i.e. the inherently weak connection between the high-level semantic concepts that humans naturally associate with images and the low-level visual features that the computer is relying upon, the task of developing this kind of systems is very challenging. *Interactive Content-Based Image Retrieval* (ICBIR) improves the retrieval performance by employing *relevance feedback* (RF), i.e. adjusts the subsequent retrieval process by using information gathered from the user's feedback.

Some CBIR systems require the user to supply starting or example images for the retrieval. These images provide extra similarity information in addition to the relevance feedback. The use of example images quite naturally leads to improved retrieval performance. However, the user-supplied initial images are not available in many applications, for example, when searching a photo of a specific criminal only through the recalling of an eyewitness. This work focuses on the cases without starting images.

Many existing ICBIR systems confront the *evaluation fatigue* problem. This is especially evident when the target class is extreme small or rare so that the searching requires too many interaction rounds to retrieve the first satisfactory

image. Consequently the user probably gets tired and ceases searching before the retrieval goal is successfully attained. One of the underlying reasons is that most existing ICBIR systems employ only unsupervised methods in extracting visual features and in building associated indexes. By this means the learning from users' feedback progresses with only mediocre efficiency. This thus motivates incorporation of supervised information for building a predictive subsystem to speed up the retrieval.

Deficient user interfaces are another obstacle for using ICBIR. Humans evaluate image similarity intelligently and usually in various focuses or semantic aspects. However, users fail to input the aspect information through a single relevance criterion which is adopted by most existing ICBIR systems. Consequently, the collected similarity data used by the relevance feedback learning algorithm is often inconsistent with the evaluation focus, and thus the prediction will deviate from the retrieval goal. To overcome the above discrepancy, a successful ICBIR system must allow users to specify certain properties of the target image, for example, the gender or race in facial image retrieval, and perform prediction consistent with such partial relevance assessments.

A simple way to accommodate partial relevance is to select image candidates according to the user-specified properties. The selection can be implemented by using available ground truths or previously trained classifiers. However, such hard filtering methods would be problematic in handling fuzzy cases where there is no common criterion for classifying the image contents, e.g. the hairstyle in facial image retrieval. It has been demonstrated by the PicSOM system [66] that *Self-Organizing Maps* (SOMs) are a powerful tool for learning from relevance feedback in ICBIR. A natural way to reuse and improve the PicSOM framework is to extend Self-Organizing Maps by incorporating supervised information. This thesis concentrates on extracting discriminative components of images as a preprocessing step before the SOM training.

The most influential factor in the complexity of the image retrieval problem is the repertoire of images in the database—the image domain [110]. A *narrow image domain* has only a limited and predictable variable in all aspects of appearance whereas a *broad image domain* has in principle an unlimited and unpredictable variability as well as ambiguous and subjective semantics. For the latter, the performance of automatically extracted visual features remains moderate and additional features (e.g. text) may thus be required for reaching an acceptable retrieval performance level (see e.g. Idée[1]). On the other hand, most real-world applications of CBIR (e.g. [9, 50, 85, 86]) belong to the former case. This work focuses on a certain domain of images, in particular facial images, in line with

[1] http://labs.ideeinc.com/visual/

the goal of pushing ICBIR closer to practical use.

Digital images are usually sampled as pixel matrices or vectors. Supervised learning based on the vectorized representations of images is thus performed in very high-dimensional spaces and inherently prone to overfitting. Complexity control is therefore necessary. This thesis discusses a number of unsupervised dimensionality reduction methods which encode images of a certain domain in a compact format. Moreover, regularization as another method for complexity control is described as well. By attaching a penalization term to the original objective, one actually incorporates a certain prior for the parameters to be learned, which can effectively compensate the data scarcity.

The success of a discriminant analysis method for images requires an efficient optimization algorithm. This work aims at finding a projection matrix that maximizes the likelihood of image labels. A speedup method should make use of the information geometry principle and fulfill the orthonormality constraint. A fast implementation by matrix multiplication will be described as well.

*Interactive Facial Image Retrieval* (IFIR) is a particular case of ICBIR. In IFIR the user searches the images of a target subject in a setting mimicking the recalling of an eyewitness. Most facial image databases contain only a small number of images for each subject, which forms the major challenge in IFIR. Conventional ICBIR systems often fail to return one of the subject images before the user feels tired. This thesis will present and analyze the IFIR query process and examine the improvement by using the proposed techniques with a popularly used database. The performance of various methods is quantified by measuring the advantage in the first relevant hit over random retrieval as well as precisions at various recall levels.

## 1.2 Contributions of the thesis

In this thesis, a novel framework that incorporates supervised information for improving the performance of *Interactive Content-Based Image Retrieval* (ICBIR) is suggested with the following major contributions:

- The appearance of the first relevant image is accelerated by utilizing partial relevance from the user feedback. In addition to filtering by ground truths and classifiers, partial relevances are implemented by a DSOM, a discriminative extension of Self-Organizing Maps, which can easily be integrated into the PicSOM CBIR system.

- Discriminative feature extraction methods are studied and a new discriminant analysis method is proposed. The proposed *Parzen Discriminant Analysis* (PDA) regularizes the *Informative Discriminant Analysis* (IDA) objective with a prior addressing the piecewise smoothness of images. Three strategies are applied to speed up the optimization:
  - The gradient calculation is reformulated by matrix operations without explicitly going through all pairwise differences.
  - The updates that approximate geodesic flow in the Stiefel manifold are adopted to avoid expensive calculation of the Givens rotation.
  - The *Principal Whitened Gradient* (PWG) method for information geometry is proposed to improve the convergence of optimization.
- A new non-negative feature extraction method, *Projective Non-negative Matrix Factorization* (P-NMF), is presented. The underlying reasons that lead to high sparseness by P-NMF are analyzed.
- The *First Relevant Hit Advantage* (FRHA) measurement is proposed to quantify the performance of an ICBIR system in retrieving the first relevant image.
- The above ICBIR improvement techniques have been applied to interactive facial image retrieval. In addition to a query demonstration, quantitative simulation study is presented to compare the performance of six ICBIR approaches.

## 1.3 Outline of the thesis

After this introductory chapter, Chapter 2 reviews the related work in content-based image retrieval and relevance feedback techniques which attack the semantic gap problem, including the essential ingredients of the PicSOM ICBIR system. It is then pointed out that the evaluation fatigue problem is one of the major difficulties in ICBIR. A solution based on partial relevance is introduced, with implementation possibilities discussed.

Chapter 3 describes discriminative feature extraction for images. First, various approaches for encoding images in a specific domain are explored. After reviewing some related work in discriminant analysis, the *Parzen Discriminant Analysis* (PDA) is presented, with extension to aspects with fuzzy classification. The

PicSOM retrieval procedure can then be enhanced by using the discriminatively trained SOMs.

The applications of the proposed ICBIR techniques on a facial image database are illustrated in Chapter 4. The motivation of interactive facial image retrieval is first described. Next, a concrete query example demonstrates the retrieval process. Quantitative simulation results are then provided.

Finally, Chapter 5 gives the conclusions of the work presented in this thesis and discusses the potential future work. This is followed by references and a set of publications detailing the proposed methods and showing the results of experiments.

## 1.4 Included publications

The following three journal articles and five conference papers have been included in this thesis.

1. Z. Yang and J. Laaksonen. Interactive retrieval in facial image database using Self-Organizing Maps. In *Proceedings of IAPR Conference on Machine Vision Applications (MVA 2005)*, pages 112–115, Tsukuba Science City, Japan, May 2005.

2. Z. Yang and J. Laaksonen. Approximated classification in interactive facial image retrieval. In *Proceedings of 14th Scandinavian Conference on Image Analysis (SCIA 2005)*, volume 3540 of *Lecture Notes in Computer Science*, Springer, pages 770–779, Joensuu, Finland, June 2005.

3. Z. Yang and J. Laaksonen. Partial relevance in interactive facial image retrieval. In *Proceedings of 3rd International Conference on Advances in Pattern Recognition (ICAPR 2005)*, volume 3687 of *Lecture Notes in Computer Science*, Springer, pages 216–225, Bath, UK, August 2005.

4. Z. Yang and J. Laaksonen. Regularized neighborhood component analysis. In *Proceedings of 15th Scandinavian Conference on Image Analysis (SCIA 2007)*, volume 4522 of *Lecture Notes in Computer Science*, Springer, pages 253–262, Aalborg, Denmark, June 2007.

5. Z. Yang and J. Laaksonen. Face recognition using Parzenfaces. In *Proceedings of International Conference on Artificial Neural Networks (ICANN*

*2007)*, volume 4669 of *Lecture Notes in Computer Science*, Springer pages 200–209, Porto, Portugal, September 2007.

6. Z. Yang and J. Laaksonen. Multiplicative updates for non-negative projections. *Neurocomputing*, 71(1-3):363–373, 2007.

7. Z. Yang, Z. Yuan, and J. Laaksonen. Projective non-negative matrix factorization with applications to facial image processing. *International Journal on Pattern Recognition and Artificial Intelligence*, 21(8):1353-1362, 2007.

8. Z. Yang and J. Laaksonen. Principal whitened gradients for information geometry. *Neural Networks*, 21(2-3):232-240, 2008.

## 1.5 Contribution of the author in the publications

In Publication 1, the author proposed to solve the IFIR problem with Self-Organizing Maps. The empirical study was cooperatively performed by the author and D.Sc. Jorma Laaksonen. Compared with the existing FACERET system, the proposed method is advantageous because it can handle multiple features simultaneously and improve retrieval performance by long-term learning from other users' relevance feedback.

In Publication 2, the author proposed to implement the partial relevances by approximated classifiers and performed the experiments with assistance from D.Sc. Jorma Laaksonen. The main finding in Publication 2 is that the appearance of the first relevant image can be significantly accelerated if and only if the employed classifiers have very high accuracy.

In Publication 3, the author identified the two phases in IFIR and proposed to improve the retrieval in the first phase by using partial relevance, which was implemented by replacing the unsupervised SOMs in the PicSOM system with DSOMs. The empirical study was accomplished by the author and D.Sc. Jorma Laaksonen. The experimental results indicate that the proposed method can remarkably speed up the appearance of the first relevant hit.

Publication 4 argues that the *Informative Discriminant Analysis* (IDA) and *Neighborhood Component Analysis* (NCA) are prone to overfitting for high-dimensional data. To address this problem, the author proposed a regularization technique which is equivalent to maximizing the posterior of the linear transformation matrix with a Gaussian prior. The experimental results on facial images

show that the regularized NCA can greatly improve the generalization performance. The problem identification, the improved algorithm, and the empirical study were all accomplished by the author.

In Publication 5, the *Parzen Discriminant Analysis* (PDA) is presented and applied to the face recognition problem. PDA regularizes the IDA objective by emphasizing the piecewise smoothness of the projection matrix entries. Another major improvement over IDA is to accelerate the gradient calculation by matrix multiplications. Furthermore, the PDA updates maintain the orthonormality by following the approximated geodesic flow in the Stiefel manifold. The above innovations were proposed by the author, who also performed the face recognition comparison experiments.

Publication 6 describes how to construct multiplicative update rules for non-negative projections based on Oja's iterative learning rule. The derivation provides a sound interpretation of learning non-negative projection matrices based on iterative multiplicative updates—a kind of Hebbian learning with normalization. A convergence analysis is sketched by interpreting the multiplicative updates as a special case of natural gradient learning. Two application examples of the proposed technique are also demonstrated: a non-negative variant of the linear Hebbian networks and a non-negative Fisher discriminant analysis, including its kernel extension. The resulting algorithms demonstrate interesting properties for data analysis tasks in experiments performed on facial images. The author was responsible for both theoretical and empirical contributions of this publication.

The *Projective Non-negative Matrix Factorization* (P-NMF) was previously presented in [134] by Prof. Erkki Oja and Zhijian Yuan. In addition to recapitulating the model and algorithms of P-NMF, Publication 7 discusses the underlying reasons that lead to high orthogonality or sparseness by P-NMF, and presents both qualitative and quantitative comparison with *Non-negative Matrix Factorization* (NMF). Furthermore, the features obtained by P-NMF can be used as an input to classifiers for facial images. Publication 7 also corrects two errors in the P-NMF update rules in [134]. The author was responsible for interpreting P-NMF, comparison experiments against NMF, classification tests on facial images, and the two corrections in the P-NMF update rules.

In Publication 8, the author devised two strategies to improve the optimization in information geometry. First, a local Euclidean embedding is identified by whitening the tangent space, which leads to an additive parameter update sequence that approximates the geodesic flow to the optimal density estimation. Second, removal of the minor components of the gradients enhances the estimation of the Fisher information matrix and reduces the computational cost. It is proven in this

publication that dimensionality reduction is necessary for learning multidimensional linear transformations. The optimization based on the principal whitened gradients demonstrates faster and more robust convergence in simulations on (i) unsupervised learning with synthetic data, and (ii) discriminant analysis of breast cancer data. Publication 8 is the extended version of [132] by the same authors. In the original and extended papers, the author was responsible for both the theoretical and empirical contributions.

# Chapter 2

# Interactive Content-Based Image Retrieval

In this chapter the related work in the interactive content-based image retrieval is reviewed. The major difficulty of the relevance feedback learning is the evaluation fatigue problem. We propose the partial relevance technique to attack this problem. Partial relevances in the PicSOM system can be implemented either by direct filtering or by Discriminative Self-Organizing Maps described in the next chapter.

## 2.1 Content-based image retrieval

An image retrieval system is a computer program for browsing, searching and retrieving images from a large database of digital images. *Content-based image retrieval* (CBIR), also known as *query by image content* (QBIC) and *content-based visual information retrieval* (CBVIR) is the application of computer vision techniques to the image retrieval problem, that is, the problem of searching for digital images in large databases. The term of CBIR first appeared in a paper by T. Kato [56] to describe automatic retrieval of images from a database based on visual features, such as color and shape. Since then, the term has been used to describe the process of retrieving desired images from a large collection on the basis of statistical image features.

### 2.1.1 Search by textual annotations

The research on image retrieval or searching for digital images in large databases can be dated back to 1970s. The traditional approach is based on textual annotations. That is, the images are represented and indexed by such metadata as filenames, captions and keywords, and the retrieval is then carried out with the text information.

Manual insertion of textual annotations is however laborious and thus impractical for large-scale image collections. Moreover, the retrieval results could be unsatisfying because manual annotations are subjective and the annotation focus may differ among human annotators. To avoid these problems, many text-based image retrieval systems, such as Google Image Search[1], extract the annotations from the text surrounding the images. Nevertheless, the text description nearby may be irrelevant or only partially relevant to the image. The language of the query may also be different from that of the text. In addition, it is impossible to query images that do not exist in a text document by such a technique.

### 2.1.2 Search by image content

Content-Based Image Retrieval emerged as a method to overcome the evident problems of text-based image retrieval (see e.g. [26, 114]). "Content-based" means that the search will analyze the actual content of the image. The term "content" in this context refers to the descriptors or information that can be derived from the image itself using automatic image processing techniques. Indexing images in CBIR differs substantially from indexing textual documents since images or visual information in general do not consist of such fundamental building blocks as words in text, which could be directly utilized. Instead, the desired attributes of images for efficient indexing are complex functions of image regions or the whole image. In this sense, image retrieval can be considered as a discipline in the intersection of traditional information retrieval (IR) and image processing.

CBIR has been a subject of very intensive research effort for more than a decade. A great number of research papers have been published and many notable CBIR systems, such as *QBIC* [81, 28], *Photobook* [92], *VisualSEEk* [111], *Tiltomo*[2],

[1] http://images.google.com
[2] http://www.tiltomo.com

*Behold Image Search*[3], *xcavator*[4], have been developed since early 1990s. Some open-source CBIR software systems such as *imgSeek*[5] and *GIFT*[6] have also appeared in recent years. For reviews of CBIR systems, see, for instance, [122] and [19]. In addition, textbooks and surveys on CBIR and multimedia retrieval in general are available, see e.g. [133].

There exists growing interest in CBIR because of the limitations inherent in text-based media search systems, as well as the large range of possible uses for efficient image retrieval. Potential uses of CBIR include photograph archives [100], retail catalogs [123], medical diagnosis [85, 108], geography and remote sensing [112], astronomy [18], crime prevention and investigation [86], education [13, 80], searching device-parts [74], trademark retrieval [9], and paper industry [50], etc.

Multiple usage types may be supported by even a single CBIR system [16]. The most precise search task is *target search*, in which the user is trying to find a specific target image, which may or may not be actually present in the database and which is the only relevant image for this query. *Category search* occurs when the user is looking for images belonging to a certain category or class of images and all images fulfilling the category criteria are considered relevant. In *open-ended search* or *browsing*, the user has a vague or inexact search goal in mind and she browses the database for any interesting images.

A query can start with an example image supplied by the user, but from the user perspective an ideal CBIR system should implement more convenient retrieval with a mental configuration. That is, the search goal only exists in the user's mind and can be expressed by using query sentences of natural languages. Among the above search types and query settings, this thesis will focus on category search without user-provided starting images.

## 2.2 Retrieval with relevance feedback

*Relevance feedback* (RF) is an interaction process, introduced in the mid-1960's for text retrieval techniques, to improve retrieval effectiveness. Original work on RF includes [99, 49, 104]. It has been defined in [102] as the process of automatically adjusting an existing query by using information fed back from

[3] http://photo.beholdsearch.com
[4] http://www.scavator.net
[5] http://www.imageseek.net/
[6] http://www.gnu.org/software/gift/

the user about the relevance of previously retrieved documents. In the context of CBIR, researchers soon felt the need to integrate RF in order to overcome the *semantic gap problem*.

### 2.2.1 Semantic gap

The image content in its original format is a pixel array. On the other hand, humans possess a highly sophisticated visual system and have a lot of *a priori* information in object recognition and semantic combination. This information is based on previous experience, personal preferences and interests, cultural issues, and the context in which the image is represented. Unfortunately, this kind of knowledge is practically impossible to implement in a computer vision application. The discrepancy between low-level visual features and high-level semantics is commonly referred to as the *semantic gap*.

The varying difficulty of the CBIR problem can also be examined from the viewpoint of different users' needs. A categorization of image retrieval was proposed by Eakins [24] who identified three distinct levels of image queries:

- Level 1, retrieval by *primitive (visual)* features.
- Level 2, retrieval by *logical* features or *semantic* attributes.
- Level 3, retrieval by *abstract attributes*.

Level 1 queries concentrate on basic low-level components of visual content such as color, texture and shape. Queries at level 2 may contain specific objects (e.g. "car") and scenes (e.g. "beach"). Level 3 queries may involve abstract concepts (e.g. images depicting "freedom" or "humor"), which requires sophisticated image understanding, knowledge representation, and reasoning about the relations and significance of objects and scenes. Generally, users are interested in searching at level 2 or 3. On the other hand, most current image retrieval systems are constructed using level 1 processing and recent research has been moving toward level 2 retrieval. Indexing and retrieval at level 3 are currently possible only by using textual descriptions.

Given the current state of machine vision technology, this thesis focuses on the narrow image domain of images to make the work closer to practical applications. The search tasks in such systems are more determined and generally at retrieval level 2. Two strategies are adopted to bridge the semantic gap. One is to steer the query toward the desired target by learning from the relevance feedback of the

user. The other is to extract semantically more informative and discriminative features from the images by using supervised learning techniques.

### 2.2.2 Interactive content-based image retrieval

The weakness of the connection between semantic concepts and visual low-level features is a serious limitation and reduces the usefulness of the content-based approach of image retrieval. As a result, many content-based retrieval applications cannot be expected to return the best available images or reach a high precision of relevant items. The systems can nevertheless serve as valuable semi-automatic tools and make retrieving of images manageable even from large-scale wide-domain image collections. More satisfactory results can often be obtained in *Interactive Content-Based Image Retrieval* (ICBIR), where the image query is refined in an iterative process with the relevance feedback technique.

Each round of ICBIR implements the relevance feedback in three steps. First, the system shows the user a set of images and remembers what images have already been shown, not to display them again. Second, the user somehow indicates which images are to some extent relevant to the present query and which are not. Third, the system changes its behavior depending on the relevance feedback provided for the seen images. During the retrieval process more and more images are assessed and the ICBIR system gradually learns the correspondence between the high-level concepts humans use and the low-level features obtained from the images.

In this work, we assume that at each iteration the user selects the relevant images and the non-selected images can be implicitly regarded as non-relevant. Of course, there exist other possibilities of granularity in relevance assessments. In some systems, for example, the non-relevant examples must also be explicitly provided and the non-selected images are considered to be neutral. The relevance scale may also be finer, e.g. containing options like “very relevant”, “relevant”, “somewhat relevant”, and so on.

Relevance feedback can be seen as a form of supervised learning to steer the subsequent query toward relevant images by using the information gathered from the user’s feedback. Three specific characteristics of relevance feedback, however, distinguishes it from many other machine learning problems [142]: (a) *The small number of training samples.* Compared with many supervised learning tasks, the number of samples relative to the dimensionality of the feature spaces is very small in relevance feedback. Only a rather small number of images are usually evaluated in one round of the query, and users are often impatient to provide much feedback.

This makes many traditional inductive learning methods ill-posed since they fail to produce stable results. (b) *The asymmetry of the training data.* The number of evaluated non-relevant images usually cannot represent the distribution of all non-relevant images well. (c) *The real-time processing requirements.* Relevance feedback is used when the user is interacting with the system and thus waiting for the completion of the algorithm. Fast response time is essential as an image query may well take several rounds until the results are satisfactory. With large databases, this usually limits the range of possible methods to ones which do not rely on processing the whole database on each query round. Furthermore, an RF learning algorithm should be accompanied with an efficient indexing technique that speeds up the access to image candidates.

### 2.2.3 Relevance feedback techniques

During a retrieval session, an ICBIR system returns at each iteration a set of unseen images that it deems to be the most relevant ones for the search goal based on the accumulated feedback information. This can be implemented by a ranking algorithm with various relevance score definitions. An excellent survey on the ranking algorithms based on relevance feedback can be found in [63]. This section summarizes the essential elements of some newly appeared ICBIR systems.

A straightforward method is to calculate the relevance score as a membership function measuring the likelihood that a given image belongs to the relevant class. In terms of probability, the supervised learning can thus be interpreted as a density estimation problem. Ves et al. [22] assumed a multivariate Gaussian distribution for user preferences and applied Bayesian learning on the Gaussian mean and covariance. Wu et al. [130] modeled the membership function by Gaussian mixtures and employed a *Discriminant Expectation Maximization* (D-EM) algorithm which estimates the parameters of a generative model and meanwhile finds the associated linear projection. Gaussian mixture models were also employed by Kherfi and Ziou in [58], where the negative examples were also taken into account. The densities of individual features can be estimated by the Parzen Window method [109]. A factorized joint density can then be obtained by assuming mutual independence of the features [4]. The membership functions can be other than densities. In [64], Kushki et al. aggregated the individual Cauchy memberships [51] which were calculated with five MPEG-7 descriptors. Alternatively, the aggregated memberships can be obtained by fuzzy methodology [120].

From the machine learning point of view, the ranking problem is closely related

to classification. Thus the relevance scores may be defined as the output of such classifiers as *Support Vector Machines* (SVMs) [17]. Zhang et al. [137] proposed to directly return the SVM classifier outputs as the relevance scores. Tao et al. modified SVM by two methods, *asymmetric bagging* [118] and *random subspace* [118, 116], which overcome two problems of relevance feedback, i.e. the small training set and the asymmetry between positive and negative samples. Instead of the raw SVM output, Hong et al. [44] proposed to rank the samples by their distances to the classification hyperplane. Wang et al. further modified the distances by SVM active learning embedding Euclidean search [126]. To attack the imbalanced data set problem, Hoi et al. presented the *Biased SVM* which adapts the 1-SVM method [42] by using double sphere hyperplanes. Note that classifiers can easily be related to the membership approach through, for example, the sigmoid function [127].

Another stream of ranking techniques is based on measuring the similarity between the query and the images in the database. In this setting, the query can be represented by the user-provided starting image, but more often by a point moving in the image feature space (e.g. [52, 32]). The Euclidean metric is non-parametric and thus cannot be adapted according to relevance feedback. A straightforward method to overcome this is to use weighted Euclidean distances, where the supervised learning becomes tuning the weights [32]. Besides feature weighting, the adaptive similarities can be calculated in a transformed space through e.g. learning a kernel discriminative projection [11, 117] or building a biased *Marginal Convex Machine* (MCM) [115]. In addition to Minkowski metrics, Gondra and Heisterkamp [30] proposed normalized information distances on raw intensities by treating images as strings.

In relevance feedback learning, labeled images account for only a small portion of a large image collection. Thus an effective way to overcome the problem of a small training set is to utilize unlabeled data. Hertz et al. [41] proposed to learn distance functions by semi-supervised boosting in a product space, where classifiers are updated by a constant factor with unlabeled data. The augmentation of the data set can greatly help clustering and reveal the compact structure before supervised learning [12, 41]. Another use of unlabeled data relies in constructing a regularization term which emphasizes the smoothness of the mapping function. He et al. applied the above manifold ranking theory to image retrieval with relevance feedback [38], and later generalized their method by initializing a pseudo seed vector, based on neighborhood relationships [39]. The similarity information among the images in a collection has also motivated a cluster-based image retrieval technique [14]. This approach actually introduces a subsystem to any CBIR systems and retrieves image clusters instead of a set of ordered images. Similar subsystems can be found also in [74, 128].

A school of thought, generally called *long-term learning*, has emerged in recent years. Humans evaluate the similarity between images intelligently, bringing semantic information in the made relevance assessments. The marking actions by previous users can be seen as hidden annotations of the images. The annotations may subsequently serve as cues for similarity in semantic contents of the images. An idea, presented in [63] and borrowed from text retrieval, is to factorize the query–image matrix by *Latent Semantic Indexing* (LSI) [68] and to construct a statistical *user interaction* feature to be used alongside with the visual features. Chan and King [10] utilized an SOM where the model vectors were modified so that neurons containing similar images in the feedback are moved closer to each other. Han et al. [34] proposed a memory learning framework where they constructed a semantic correlation graph. With this graph their system can calculate such indicators as *authoritative rank* and *hidden semantic correlation* between image and feedback examples and propagate annotations from one image to another. A classifier approach for long-term relevance feedback was proposed by Hoi et al. [43], where the standard SVM algorithm is modified to handle noisy labels.

## 2.3 Relevance feedback with Self-Organizing Maps

Most relevance feedback techniques, for example the ones described in Section 2.2.3, treat the feature space in a global manner. However, a global distance measure or feature extraction technique which is advantageous in the vicinity of a set of images similar to each other, may not produce favorable results for images that are more dissimilar from one another.

In this section, we describe how relevance feedback can be implemented by using multiple Self-Organizing Maps (SOMs). Contrary to most existing methods, the SOM-based relevance feedback technique is local in the sense that it operates only in the local neighborhoods of images marked positive or negative by the user. Therefore, the method respects the nonlinear nature of image similarity better. The method also dynamically produces an implicit weighting of different features. The empirical results [66] indicate that the PicSOM CBIR system is able to effectively utilize a set of parallel SOMs so that the combined retrieval result exceeds the performance of any of the features used separately.

### 2.3.1 Self-Organizing Map

An SOM consists of a (usually two-dimensional) regular lattice or grid of map units. In the PicSOM system, SOM grid type for image retrieval is rectangular [66, 63]. A model vector $\mathbf{m}_i \in \mathbb{R}^M$ is associated with each map unit $i$. During the training phase, a set of feature vectors is presented to the map multiple times and the model vectors are modified to match the distribution and topological ordering of the training data in the feature vector space. The fitting of the model is usually carried out by a sequential regression process, where $t = 0, 1, 2, \ldots, N-1$ is the step index: For each input sample $\mathbf{x}(t)$, first the index $c(\mathbf{x})$ of the *best-matching unit* (BMU) or the *winner model* $\mathbf{m}_{c(\mathbf{x})}(t)$ is identified by the condition

$$\forall i: \quad \|\mathbf{x}(t) - \mathbf{m}_{c(\mathbf{x})}(t)\| \leq \|\mathbf{x}(t) - \mathbf{m}_i(t)\| \ . \tag{2.1}$$

The usual distance metric used here is the Euclidean distance. After finding the BMU, a subset of the model vectors constituting a neighborhood centered around the BMU node $c(\mathbf{x})$ are updated as

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h(t; c(\mathbf{x}), i)(\mathbf{x}(t) - \mathbf{m}_i(t)) \ . \tag{2.2}$$

Here $h(t; c(\mathbf{x}), i)$ is the *neighborhood function*, a decreasing function of the distance between the $c(\mathbf{x})$-th and $i$-th nodes on the map grid. This regression is reiterated over the available samples and the value of $h(t; c(\mathbf{x}), i)$ is allowed to decrease in time to guarantee the convergence of the prototype vectors $\mathbf{m}_i$. Large values of the neighborhood function $h(t; c(\mathbf{x}), i)$ in the beginning of the training initialize the map and small values at later iterations are needed in fine-tuning.

The SOM algorithm has a number of important properties (see e.g. [37]) that make it suitable for indexing image feature data. As a dimensionality reduction tool, SOM summarizes and represents the image collection in a rather compact manner. More importantly, the SOM grid provides a good approximation of the input space in a way that preserves topological ordering, which is a useful property lacking in basic clustering algorithms. This ordering is especially convenient for efficient RF implementation as similar images reside in neighboring map units.

The search for the BMU dominates the computing time of the SOM algorithm and makes training of large SOMs computationally too expensive especially if the dimensionality of the input vectors is high. The basic algorithm uses linear search, in which all map units must be evaluated to find the BMU. This makes the complexity of the search $O(\tau)$, where $\tau$ is the number of map units. To speed up the BMU search, [61] introduced a variant of SOM called the *Tree Structured Self-Organizing Map* (TS-SOM) [61, 60]. TS-SOM is a tree-structured vector quantization algorithm that uses normal SOMs in each of its hierarchical levels.

It is loosely based on the traditional depth-first tree-search algorithm. Due to the tree structure, the number of map units increases when moving downwards the SOM levels of a TS-SOM. The search space for the BMU (2.1) on the underlying SOM level is restricted to a fixed-sized portion just below the BMU of the above SOM. Unlike most tree-structured algorithms, the search space is not limited to the direct children of the upper-level BMU. Instead, the search space can be set to include also neighboring nodes having different parent nodes in the upper level. With the TS-SOM one can follow the hierarchical structure and reduce the complexity of the search to $O(\log \tau)$. Thus the reduced computational requirements obtained by using the TS-SOM algorithm facilitate the creation and use of large SOMs, needed for indexing huge image databases.

### 2.3.2 Relevance feedback in PicSOM

The indexing and relevance feedback techniques with SOMs form the backbone of the PicSOM ICBIR system. In the experiments in Publications 1–3, the used TS-SOM structure had three levels with sizes $4 \times 4$, $16 \times 16$, and $64 \times 64$ map units. The training data consisted of 2409 feature vectors and each vector was presented 100 times for the adaptation of each map level. After training a TS-SOM, all images in the database were mapped in their BMUs at the bottom-most level. This mapping of each TS-SOM thus forms an index of the images.

In each round of the image query, the PicSOM system presents the user a set of images she or he has not seen before. The user is then expected to mark the relevant image as positive, and the system implicitly interprets the unmarked images as negative. With the TS-SOM index, it is straightforward to locate the positive and negative images on each feature map. The map units are awarded a positive impulse for every positive image mapped in them. Likewise, associated negative images result in negative impulses. These positive and negative impulses are scaled so that the total sum of all scores on each map is equal to zero.

SOM units get raw score value in each zero-sum sparse value field. Next, PicSOM applies low-pass filtering on the SOM surface. Fig. 2.1 illustrates how the positive and negative responses, displayed with white and black map units, respectively, are first mapped on s SOM surface and how the responses are expanded in the convolution. The relevance score of each image can then be obtained from the convolved value of the BMU associated with the image. Laaksonen et al. [67] showed that the scores obtained in this way actually form a discrete estimation of the Bayesian decision.

When multiple features are used, the combined score for each image can be

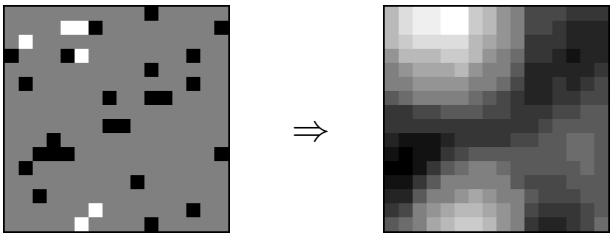

Figure 2.1: An example of how a SOM surface, on which the images selected and rejected by the user are shown with white and black marks, respectively, are convolved with a low-pass filter.

obtained most simply as just the sum of those from the individual feature maps. Unseen images with highest combined scores will be displayed to the user in the next round. This was justified [67] if the used features are independent. Yet, the experiments in [66, 67] also suggest that PicSOM's retrieval accuracy can be increased by adding new features in the system even if the added feature is highly correlated with one or more existing features in the sense of large mutual information.

With the Bayes theorem, it can be shown that an RF learning algorithm should increase the probability of relevance in the vicinity of images marked by the user [67]. The same effect is obtained in PicSOM by the simple method of adding the signed impulses caused by newly marked images to the distribution, convolving, and renormalizing. Exactly the same is done for the distribution of non-relevant images, and thus both the positive and negative sample distributions get tuned in the query process.

One may argue that two cells far in the SOM may present more similar images than some cells closer. Actually this seldom affects the PicSOM ranking. First, two cells far in the SOM map are of little interest because (i) SOM usually preserves local distances well; (ii) in retrieval only images with top ranking, i.e. closest to the target, are returned in each round. Two dissimilar images may be mapped to close SOM cells, which often happens in the cluster boundaries which are of low density. However, the SOM training algorithm allocates only a small number of cells for the low-density areas. Furthermore, the images in dense areas, e.g. the cluster centers, are more likely to be displayed in early rounds. Therefore, the probability of wrong behavior of PicSOM ranking is small.

## 2.4 Evaluation of system performance

As content-based image retrieval is a particular type of IR, its system performance can be evaluated by the metrics used for generic information retrieval. One popularly adopted measurement metric is the plot of *precision versus recall*. Suppose $r[j]$ is the cumulative number of relevant images retrieved after $j$ rounds; $s$ is the total number of relevant images in the database, and $n[j]$ is the cumulative number of total images retrieved after $j$ rounds, i.e. $n[j] = j \cdot m$, where $m$ images are displayed in each round. The performance statistics *recall* and *precision* [98] are then defined as:

$$\text{recall}[j] = r[j]/s, \tag{2.3}$$

$$\text{precision}[j] = r[j]/n[j]. \tag{2.4}$$

In the plot of precision versus recall, larger area under the precision curve indicates better overall ranking.

Yet, interactive retrieval usually involves multiple rounds and the importance of the rounds are asymmetric. It is often desired that all relevant images should be retrieved in early rounds. Such performance difference can be emphasized by *Average-precision* as illustrated in Publication 1.

$$\text{average-precision}[j] = \frac{1}{s}\sum_{i=1}^{j}(cc[i] \cdot \text{precision}[i]) \tag{2.5}$$

where $cc[i]$ is the relevant ratio in $i$-th round, i.e. the number of relevant images shown in $i$-th round divided by $m$. A larger average-precision value indicates the retrieval is better in early rounds.

The above conventional IR performance statistics concern all relevant items. However, as will be discussed in Chapter 4, the importance of the first relevance image is much more significant for ICBIR without initial user-provided images. In Publications 1–3, a new evaluation metric has been proposed to address such performance differences. Suppose $N$ is the number of all images. Let $j$ denote the random variable for the position of the first relevant hit when using random retrieval. It is not difficult to show that the mean of $j$ is

$$\tilde{\varrho} \equiv E\{j\} = (N - s)/(s + 1). \tag{2.6}$$

Thus the improvement compared with the random retrieval can be quantified by the following *First Relevant Hit Advantage* (FRHA) measurement:

$$\text{FRHA} = \varrho = \frac{\tilde{\varrho}}{i} = \frac{N - s}{i \cdot (s + 1)}, \tag{2.7}$$

where $i \in \{1, \ldots, N - s + 1\}$ is the position of the first relevant hit using the retrieval method being evaluated. FRHA equals one when the retrieval is done in a random manner and increases when the method is able to return the first relevant image earlier. For example, it equals two when the first relevant hit occurs in the position whose index is half of the expected index in the random retrieval.

## 2.5 Retrieval by partial relevance

In most ICBIR systems, the relevance feedback is assumed to be consistent with the underlying image ranking model. On the other hand, the user can freely use any relevance criterion in mind, but the evaluation criterion is by no means input to the RF learning algorithm. Consequently, the ranking at each iteration is only suboptimal and more iterations are needed to accomplish successful retrieval.

### 2.5.1 Evaluation fatigue problem

The relevance feedback techniques try to reduce the semantic gap through user interaction. Involving a human user in the procedure may however lead to the *evaluation fatigue* problem. Systems that require a substantial amount of evaluation labor may cause the user getting tired and ceasing the search before the retrieval goal is successfully attained.

Evaluation fatigue problem is the major challenge in interactive content-based image retrieval. It becomes especially significant for precise searching tasks where the relevant class is much smaller than the database size. In the interactive facial image retrieval application, which will be discussed in Chapter 4, the user usually seeks a target person with only two or three images while there are thousands of candidates in the database. In this extreme, there may be few or no relevant images displayed in the early rounds of interactive retrieval if the relevance criterion is fixed to the exact membership in the target class. Consequently, the retrieval would get stuck due to the *no-positive-feedback* problem. That is, only negative or neutral responses are input to the RF learning, but they provide only little semantic information about the target class. The iteration may therefore progress in a random manner, which probably makes the user feel tired and cease searching. To relieve the evaluation fatigue, an ICBIR system should allow the user to input relevance information in various focuses. In turn, the RF learning can better adapt towards the search goal.

### 2.5.2 Refining relevance criterion

The evaluation fatigue problem discussed in Section 2.5.1 is caused by small target classes and is mainly induced by the rigorous relevance criterion. The user is unable to input knowledge about the retrieval goal at different focus levels. To overcome the no-positive-feedback problem, one can divide the whole retrieval process into multiple *phases*. A good ICBIR system should allow the user to use relaxed relevance criteria in the early phases and later switch to more precise ones.

However, most existing ICBIR systems do not accommodate such broad-to-narrow search goal changes. Their RF algorithms inherently assume the relevance criterion employed by the user remains unchanged. This strategy seems to work well for simple broad category search like "buildings" or "birds", but is often problematic in precise target searches with narrow categories such as facial images. Although the user can freely use any criterion in mind, the resulting relevance feedback is not necessarily consistent with the underlying index and learning algorithm.

### 2.5.3 Partial relevance

To support the above shift in relevance criterion, we implemented in the PicSOM ICBIR system both *exact relevance* for membership in the precise target class and *partial relevance* for membership in some wider semantic class. The former criterion represents the ultimate query goal while the latter corresponds to a certain limited property of the target.

An *aspect* of partial relevance is a descriptive domain that consists of a number of semantic classes. For example, in the *ethnic origin* or *race* aspect there can be three major classes: *Asian*, *Black*, and *White*. The following query procedure illustrates how a user might look for images of a specific woman of Asian origin by using partial relevance assessments in the first retrieval phase:

1. The system displays a random set of facial images.

2. If one of the subject images appears, the first retrieval phase terminates; goto the optional second retrieval phase.

3. The user marks all images that depict Asian women as relevant, while leaving the others unmarked.

4. The system applies the retrieval algorithm, and displays the images with the highest scores; goto step 2.

The system should also be able to learn partial relevance of multiple aspects in parallel. A more concrete query example of interactive facial image retrieval of this kind will be provided in Section 4.5.

It is worth to notice that partial relevance does not mean the metadata associated with images. Instead, it is a means by which the user supplies the reason or intention that is used in assessing relevance. A conventional CBIR system does not accommodate partial relevance. The user specifies relevance of the images, but the system does not know why. In the above example, the user assesses a facial image as relevant because both the person in the image and the searched person are Asian. However, without such reason information, the relevance feedback can be meaningless or even harmful for the retrieval. For instance, if the underlying system metric is based on eigenface features (see Section 3.1.1), which are known to probably encode hairstyles (Publication 1), the relevance feedback algorithm will consequently return answers irrelevant to the Asian race. Partial relevance is proposed to address such a mismatch.

### 2.5.4 Filtering

For some partial relevance aspects, the semantic class labels take categorical values such as *male* and *female* in the aspect *gender*. It is thus intuitively appealing to implement the partial relevance by filtering. Once the user has specified the semantic classes of the target, the images which do not satisfy the partial relevance criteria can be filtered out and never displayed to the user. Filtering can be implemented in two different ways, either by using existing deterministic *ground truth* information or by constructing feature-based *classifiers*.

There often exist metadata associated with images in a database. This metadata may contain useful ground truth information of the images. For example, in interactive facial image retrieval, most face databases contain gender information for each image. As the gender can be reliably remembered in most facial image retrieval scenarios, we can filter out images of the wrong gender. This will effectively reduce the processing time and evaluation labor.

For some aspects, e.g. the possible existence of *mustache* or *eyeglasses*, even though there are no ground truths available in databases, most people have a common understanding about the categorization. In such cases, it is feasible to implement stochastic filtering by training a classifier. If a state-of-the-art clas-

sifier works well for these aspects of partial relevance, one can use classification techniques to construct the filters, which will further reduce the number of candidate images. Such filtering might prevent the appearance of the correct images in some rare cases, but this can still be acceptable considering the significant reduction of evaluation fatigue.

Filtering is a simple method that implements partial relevance. It is however not applicable for some aspects where people do not have a common understanding about the categorization. For example, the decision of *racial class* by appearance may differ among people. Similar variability also holds for the categorization of the *hairstyle* aspect. It is therefore impossible to implement filtering for such fuzzy aspects. In the next chapter, a solution for handling such aspects will be presented, where the supervised information is incorporated in constructing the discriminative extension of the SOM-based index.

# Chapter 3

# Discriminative Learning for Images

Partial relevance works like a bridge enforcing the consistence between the search goal and the relevance feedback learning algorithm. A user interface equipped with partial relevance assessment allows the user to specify particular properties of the target image. In the other end, the ranking according to partial relevances requires discriminative machine learning on such properties of images. Such learning makes use of some supervised training data to enhance inference on high-level semantics of images. In this chapter, we address the learning as a feature extraction procedure. The resulting image representations can enhance the discriminative power of the PicSOM ranking technique described in Section 2.3. The inference can thus be more pertinent to the search goal using partial relevance feedbacks.

## 3.1 Domain image representations

Digital images are originally stored as pixel arrays. However, directly using image pixels is very inefficient for visual recognition. For example, a $32 \times 32$-sized gray-level image, which looks only as small as an icon, leads to a feature vector in a 1024-dimensional space. Finding patterns among a large number of such vectors is clearly infeasible due to massive storage and computation requirements. Moreover, pixel values as a visual feature are very sensitive to noise and varying

imaging conditions such as shifts, rotations and scale changes.

Some visual features of images in a broad domain can be computed by using certain image processing techniques. The conventional feature types, including *color*, *texture*, and *shape*, have been extensively studied in many review articles as well as CBIR textbooks (see e.g. [71, 8]). MPEG-7, a standardization initiative for describing multimedia content also follows this categorization. In recent years, wavelet-based feature extraction, which analyzes image content in multi-resolutions (see e.g. [20, 21]), has been introduced to obtain compressed representation of images.

This thesis focuses on ICBIR systems for images in a narrow domain, e.g. facial images. The above feature extraction methods which serve for general purposes cannot capture the underlying statistical properties of such an image collection. In the following we will address a number of dimensionality reduction approaches which not only compute the compact image representations, but also encode a certain characteristic of the image collection as a whole in the original space. In this work we only consider grayscale images, while the discussion can be easily extended to involve color information.

A vectorized image $\mathbf{x}$ can be viewed as a random variable in a high-dimensional space $\mathbb{R}^M$. We may further assure that it has been normalized to zero mean by subtracting its expected value. In projective feature extraction, the representation of the image $\mathbf{x}$ in a low-dimensional space is given by

$$\mathbf{y} = \mathbf{W}^T\mathbf{x}, \tag{3.1}$$

where $\mathbf{W}$ is an $M \times R$ orthonormal matrix with $R \leq M$. The orthonormal subspace projection method is widely used in image representations because: (1) the images can be encoded in a compact manner; (2) the decoding function can be easily obtained by simply transposing the encoding projection matrix; and (3) both encoding and decoding can be efficiently implemented with standard mathematical software that supports matrix multiplication.

### 3.1.1 Eigenfeatures

Suppose that the zero-mean input data is given in the form of an $M \times N$ matrix $\mathbf{X}$, each column $\mathbf{x}_i$ for a vectorized sample. *Principal Component Analysis* (PCA) finds an orthonormal projection matrix $\mathbf{W}_{\mathrm{PCA}}$ that minimizes the sample square

construction error:

$$\mathbf{W}_{\mathrm{PCA}} = \arg \max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{W}\mathbf{W}^T\mathbf{x}_i\|^2. \tag{3.2}$$

The above optimization problem can be solved by eigenvalue decomposition:

$$\mathbf{S}\mathbf{W}_{\mathrm{PCA}} = \mathbf{W}_{\mathrm{PCA}}\mathbf{D}_{\mathrm{PCA}}, \tag{3.3}$$

where $\mathbf{D}_{\mathrm{PCA}}$ is a diagonal matrix with $D_{ii}$ the $i$-th eigenvalue of the sample data covariance matrix $\mathbf{S} = \sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T = \mathbf{X}\mathbf{X}^T$, with $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$.

As a widely used method in machine learning and signal processing, PCA has a number of merits. The unique principal subspace can be identified in a single step. The compression and decompression with a projection matrix can easily be implemented. PCA is known for its denoising capability and performs robustly in presence of outliers. Furthermore, it has been shown that the PCA objective is a lower bound of K-means clustering [23] and PCA thus forms an approximation in finding data clusters.

The features extracted through a PCA projection are called *eigenfeatures*. Since 1990s, this technique has been extensively used in computer vision problems such as object detection (see e.g. [79]), face recognition [121] and fingerprint preselection [53]. In particular, eigenfeatures of facial images are called *eigenfaces* [121]. Note that the *Discrete Cosine Transformation* [31], the core of JPEG image compression standard [91, 125], is an approximation of PCA if the images are not constrained to a specific domain.

### 3.1.2 Non-negative projections

It was recently pointed out that the positivity or non-negativity of a linear expansion is a very powerful constraint that seems to lead to sparse representations of images [69, 70]. This kind of representations have some biological significance, as the sparse features seem to correspond to the receptive fields of simple cells in the area V1 of the mammalian visual cortex [69].

Incorporating the non-negativity constraint to the PCA objective, one can obtain the *Projective Non-negative Matrix Factorization* (P-NMF) method ([134] and Publication 7). Suppose that all elements of the input data matrix are non-negative. P-NMF equipped with the Frobenius matrix norm solves the following

optimization problem:

$$\underset{\mathbf{W}\geq 0}{\text{minimize}} \quad \sum_{m=1}^{M}\sum_{i=1}^{N}\left[\mathbf{X}-\mathbf{W}\mathbf{W}^T\mathbf{X}\right]_{mi}^2 = \sum_{i=1}^{N}\|\mathbf{x}_i-\mathbf{W}\mathbf{W}^T\mathbf{x}_i\|^2. \tag{3.4}$$

A locally optimal $\mathbf{W}$ can be obtained by iteratively applying the update rule

$$W_{mr} \leftarrow W_{mr}\frac{2\left[\mathbf{X}\mathbf{X}^T\mathbf{W}\right]_{mr}}{\left[\mathbf{W}\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}\right]_{mr}+\left[\mathbf{X}\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{W}\right]_{mr}}, \tag{3.5}$$

where $m = 1, \ldots, M$ and $r = 1, \ldots, R$. Compared with the PCA objective 3.2, the P-NMF optimization does not involve explicit orthonormalization. However, empirical results show that P-NMF implicitly enforces orthonormality on the transformation matrix $\mathbf{W}$ ([134] and Publications 6 & 7), which is insensitive to random initialization of $\mathbf{W}$ (Publications 6 & 7).

P-NMF can be viewed as an improved variant of *Non-negative Matrix Factorization* (NMF) proposed by Lee and Seung [69, 70]. The new method differs from NMF in that it replaces the weight matrix in NMF with the inner product of the base vectors and the input images. Unlike other variants of NMF, P-NMF does not involve any regularization terms or trade-off parameters, but is still able to learn more spatially localized, part-based representations of visual patterns.

The objective for learning a non-negative projection is not necessarily restricted to (3.4). In Publication 6 we showed that by using Oja's learning rule [83] a great variety of additive learning rules can easily be converted to corresponding multiplicative versions that maintain the non-negativity after each iteration. With this technique, almost identical results to P-NMF can be obtained by imposing the non-negativity constraint on linear Hebbian networks. The derivation of our approach provides a sound interpretation of learning non-negative projection matrices based on iterative multiplicative updates. This can be interpreted as a kind of Hebbian learning with normalization.

### 3.1.3 Other representation methods

The PCA method only utilizes the first and second order statistics, which might not be sufficient for many non-Gaussian distributed data sets. *Independent Component Analysis* (ICA) [48] takes into account also the higher-order statistics by e.g. maximizing the non-Gaussianity of the projected data. Trained with natural image blocks, the projection basis identified by ICA looks like Gabor filters and coincides with wavelet results [48]. It is also reported that ICA is applicable to

computer vision problems such as face recognition [5, 72]. However, ICA requires a large amount of training data to accurately estimate the high-order statistics and is thus sensitive to the curse of dimensionality. In practice, prewhitening with dimensionality reduction is required before ICA learning. Furthermore, the ICA rotation does not alter the relationships among data points in the prewhitened space and thus produces only little discriminative information for subsequent pattern recognition.

Images in a narrow domain can be seen as samples taken from a manifold according to certain probabilities. In recent years, manifold learning methods such as *Isomap* [119], *Laplacian Eigenmap* [6], and *Local Tangent Space Alignment* [140] have been introduced to find a low-dimensional embedding to represent the data samples. The above nonlinear methods, however, are only applicable to the training set and cannot extract features for newly coming data. Linearization has later been incorporated to address this problem [40, 138], where the transformation is restricted to centering and projections. Although manifold learning algorithms have been reported to produce good performance for some data sets, they assume the data manifold is sampled densely enough. This might be problematic if the training data is scarce. For high-dimensional data like images, most manifold learning methods require PCA as a preprocessing step (see. e.g. [40, 138]).

Many linear projective methods are readily extended to their nonlinear version by using the *kernel trick* [106]. The resulting algorithms may possess better expressive capability if an appropriate kernel is used. The kernel trick, nevertheless, requires extra effort to select among kernel types and to adjust the parameters associated with the kernel, which might lead to an even more complicated problem.

In recent years segmentation- or region-based based image representations have become an active research field. Contrasting to approaches that use only global features of images, the region-based methods, e.g. [59, 65, 75, 94, 95, 96, 97, 124], attempt to partition an image into a number of segments in which the pixels share homogeneous color or texture. The main objective of using region features is to enhance the ability of capturing user's perception of image content [59]. However, fully automated segmentation of images into objects itself is an unsolved problem. Even in fairly specialized domains, fully automated segmentation causes many problems and is usually not easy to be realized [80]. Especially, the shape feature of each segment is highly unreliable and thus often discarded in the representation (e.g. [59, 75]). In addition, it is more difficult to encode the spatial relationship among segments than pixels. Therefore, many existing region-based methods (e.g. [59, 65, 75, 94, 95, 124]) only treat an image as a bag of segments. Matching between two such unordered sets requires more computationally demanding

techniques such as the *earth mover's distance* [101].

Local descriptors are another popularly used approach to obtain certain invariant features of images. The *Scale Invariant Feature Transform* (SIFT) descriptor [76] is a 3D histogram of gradient locations and orientations where the contribution to location and orientation bins are weighted by the gradient magnitude and a Gaussian window overlaid over the region. Many variants of SIFT such as the GLOH descriptor [78] and PCA-SIFT [57] have later been proposed. Besides so-called interest point descriptors for general images, local descriptors for certain image types exist, for instance, *oriented chamfer matching* [107] for contour fragments, *shape context* [7] for shape images, and Gabor filters (e.g. [33]), *cosine-modulated wavelets* [62], and *Local Binary Patterns* (LBP) [84] for textures. LBP has later been extended to other vision problems such as face recognition [1] and facial expression recognition [141].

Local descriptors provide certain robustness in object detection and classification, of which the research are albeit still ongoing and facing one or more difficulties. First, interest point descriptors confront the costly matching problem as there are variable numbers of unordered interest points. Second, many local image descriptor algorithms incorporate some form of histogramming [129], which only works for image patches of certain sizes. An histogram is inaccurate for too small patches while losing spatial information for too large patches, which prevents the use of local descriptors in multi-resolutive scenarios. Moreover, dimensionality reduction is usually not of concern in the local descriptor approaches. The length of image coding using such methods is often comparable to the raw representation (e.g. [1]) or even much larger than the latter (e.g. [73, 139]). This would make the subsequent supervised learning ill-posed and computationally expensive. Additional feature selection or subsampling might be needed to obtain a compact representation (e.g.[73]).

## 3.2 Discriminant analysis

Discriminant Analysis (DA) in general aims at extracting discriminative components of data. Consider a supervised data set which consists of pairs $(\mathbf{x}_i, c_i)$, $i = 1, \ldots, N$, where $\mathbf{x}_i \in \mathbb{R}^M$ is the primary data of zero mean, and the auxiliary data $c_i$ takes categorical values that are numbered from 1 to $Q$. We seek for an $M \times R$ matrix $\mathbf{W}$ by which the primary data $\mathbf{x}_i$ are projected into a lower-dimensional space $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$. The objective is to maximize a certain discriminative criterion in the projected space.

### 3.2.1 Fisher Discriminant Analysis

In Publication 3, the Fisher's *Linear Discriminant Analysis* (LDA) [27] was applied for discriminative feature extraction. Let

$$\mathbf{S}_B = \frac{1}{Q}\sum_{c=1}^{Q} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T \tag{3.6}$$

denote the between-class scatter matrix and

$$\mathbf{S}_W = \frac{1}{N}\sum_{c=1}^{Q}\sum_{i\in\mathcal{I}_c}(\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T \tag{3.7}$$

the within-class scatter matrix, where $N_c$ and $\mathcal{I}_c$ are the number and indices of the samples in class $c$, respectively, with $\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{i\in\mathcal{I}_c}\mathbf{x}_i$ and $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \mathbf{0}$.

Then the projection matrix $\mathbf{W}_{\mathrm{LDA}}$ can be obtained by solving the eigenproblem

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{W}_{\mathrm{LDA}} = \mathbf{W}_{\mathrm{LDA}}\mathbf{D}_{\mathrm{LDA}} \tag{3.8}$$

with the largest $r$ eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$ in the diagonal of $\mathbf{D}_{\mathrm{LDA}}$ and the associated eigenvectors in the columns of $\mathbf{W}_{\mathrm{LDA}}$.

To reduce noise and computational cost, the LDA method is often accompanied with an unsupervised dimensionality reduction by PCA. Given the covariance matrix $\mathbf{S}$ of the primary data, PCA produces a matrix $\mathbf{W}_{\mathrm{PCA}}$ whose columns are the eigenvectors of $\mathbf{S}$ corresponding to the largest eigenvalues. One can then obtain the joint PCA+LDA projection

$$\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i = \mathbf{W}_{\mathrm{LDA}}^T\mathbf{W}_{\mathrm{PCA}}^T\mathbf{x}_i. \tag{3.9}$$

When applied to the face recognition problem, the projection method (3.9) is named *Fisherface*, where the extracted features are called *Fisherfaces*.

### 3.2.2 Parzen Discriminant Analysis

The PCA+LDA method is attractive for its simplicity, but its performance is restricted by the implicit assumption that each class has a single Gaussian distribution with a common variance. Moreover, the discriminative information may be lost during the unsupervised PCA preprocessing. Although many variants of LDA (e.g. [46, 47, 15]) have been proposed, they use only the first- and

second-order statistics of the class distributions while discarding the higher-order statistics.

In Publication 5 we have proposed a novel DA method that maximizes the *Parzen Discriminant Analysis* (PDA) objective:

$$\mathcal{J}_{\mathrm{PDA}}(\mathbf{W}) = \mathcal{J}_{\mathrm{IDA}}(\mathbf{W}) - \lambda \mathcal{F}(\mathbf{W}), \tag{3.10}$$

where

$$\mathcal{J}_{\mathrm{IDA}}(\mathbf{W}) = \frac{1}{2} \sum_{i=1}^{N} \log p(c_i | \mathbf{W}^T \mathbf{x}_i) \tag{3.11}$$

is the *Informative Discriminant Analysis* (IDA) objective [87] that measures the discriminative information in the projected space.

The regularization term $\mathcal{F}(\mathbf{W}) = \frac{1}{2}\mathrm{trace}\left(\mathbf{W}^T \boldsymbol{\Omega} \mathbf{W}\right)$ emphasizes the smoothness prior of images [35], where the constant matrix $\boldsymbol{\Omega}$ is constructed by $\Omega_{st} = \mathcal{N}(d(s,t);\rho)$. Here $d(s,t)$ is the 2-D Euclidean distance of the pixel locations $s$ and $t$, and $\mathcal{N}$ the zero-mean normal distribution. The variance parameter $\rho$ controls the neighborhood size in smoothing and its value depends on the resolution of the images used. We have found that $\rho \in (0.3, 0.8)$ works fine with $32 \times 32$- and $46 \times 56$-sized facial images in our experiments to be detailed in Chapter 4. It is not difficult to see that $\mathrm{trace}(\mathbf{W}^T \boldsymbol{\Omega} \mathbf{W})$ is an approximated version of the Laplacian smoothness constraint used in [35]. The feature extraction method is called *Parzenface* when PDA is applied to the face recognition problem as in Publication 5 and Chapter 4.

Denoting again $\mathbf{y}_i = \mathbf{W}^T \mathbf{x}_i$, the IDA objective can be rewritten by the Bayes theorem as:

$$\mathcal{J}_{\mathrm{IDA}}(\mathbf{W}) = \frac{1}{2} \left[ \sum_{i=1}^{N} \log \frac{p(\mathbf{y}_i | c_i)}{p(\mathbf{y}_i)} + \sum_{i=1}^{N} \log p(c_i) \right]. \tag{3.12}$$

If we estimate $p(\mathbf{y}_i|c_i)$ and $p(\mathbf{y}_i)$ by the Parzen window method, the objective becomes

$$\mathcal{J}_{\mathrm{IDA}}(\mathbf{W}) = \frac{1}{2} \sum_{i=1}^{N} \log \sum_{j=1}^{N} \phi_{ij} e_{ij} - \frac{1}{2} \sum_{i=1}^{N} \log \sum_{j=1}^{N} e_{ij} + \mathrm{const}, \tag{3.13}$$

$$= \mathcal{J}^c - \mathcal{J} + \mathrm{const}. \tag{3.14}$$

where $\phi_{ij} = \delta(c_i, c_j)$, and $e_{ij} = (1 - \delta(i,j)) \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\sigma^2}\right)$, with $\sigma$ a positive parameter which controls the Parzen window width.

### 3.2.3 Fuzzy discriminant analysis

Conventional DA can only handle hard categorization problems where the probability that a sample belongs to a class is either 1 or 0. However, the partial relevance of some aspects often requires intermediate membership probabilities. For example, in facial image retrieval *hairstyle* plays an important role for recognizing a person. Although there are people of typically long hair or short hair, one can always find many different in-between cases. The decision whether such an intermediate sample belongs to the *long* or to the *short* class varies between different assignments. Therefore it is not enough to implement the categorization by a binary scale since it might induce an even more severe semantic gap.

The problem can be solved by introducing a *fuzzy set* [136] whose membership function quantifies the grade of membership of a facial image in a semantic class. This can be implemented by an $N \times Q$ matrix $\mathbf{M}$ with the $ic$-th element.

$$M_{ic} = \text{membership of } \mathbf{x}_i \text{ to class } c. \tag{3.15}$$

In practice, annotating the samples with continuous labels requires too much labor and a discrete approximation is therefore usually used instead.

With this modification we can slightly extend the LDA algorithm for fuzzy cases by replacing the within-class scatter matrix and class means by

$$\tilde{\mathbf{S}}_W = \frac{1}{N} \sum_{c=1}^{Q} \sum_{i=1}^{N} M_{ic} (\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_c)(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_c)^T, \tag{3.16}$$

where $\tilde{\boldsymbol{\mu}}_c = \left(\sum_{i=1}^{N} M_{ic}\mathbf{x}_i\right) / \left(\sum_{i=1}^{N} M_{ic}\right)$. For PDA, the fuzzy extension just replaces $\phi_{ij}$ in (3.13) by $\tilde{\phi}_{ij} = 1 - \xi_{ij}/\Xi$, where $\xi_{ij} = \sum_{c=1}^{Q} |M_{ic} - M_{jc}|$ is the total labeling difference between the $i$-th and $j$-th samples, and $\Xi = \max\{\xi_{ij}\}$ is the maximum difference over all $(i, j)$ for normalization.

## 3.3 Optimization of PDA

The optimization of Parzen Discriminant Analysis requires more computation than LDA. In [87], Peltonen and Kaski employed the stochastic gradient ascent method to optimize the IDA objective, but their algorithm runs rather slowly. Furthermore, to obtain an orthonormal transformation matrix, they employed a reparameterization based on Givens rotation, which even aggravates the computation and prevents its application to high-dimensional data. Peltonen et al. [89]

later proposed a modified version to speed up the computation by using a small number of Gaussian mixtures instead of the Parzen method. This nevertheless loses the advantage of nonparametric estimation. One has to insert additional EM iterations before computing the gradient, and how to select an appropriate number of Gaussians is unclear.

In Publications 5 and 8, we have proposed a number of advances in the IDA optimization while preserving the advantages of Parzen estimations.

### 3.3.1 Computing batch gradients

The gradient of $\mathcal{J}_{\mathrm{IDA}}(\mathbf{W})$ as defined in ([87] and used in Publication 5) is:

$$\nabla \mathcal{J}_{\mathrm{IDA}}(\mathbf{W}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} H_{ij} \left(\mathbf{x}_i - \mathbf{x}_j\right) \left(\mathbf{x}_i - \mathbf{x}_j\right)^T \mathbf{W}, \tag{3.17}$$

where

$$H_{ij} = -\frac{1}{\sigma^2} \left( \frac{\phi_{ij} e_{ij}}{\sum_{s=1}^{N} \phi_{is} e_{is}} - \frac{e_{ij}}{\sum_{s=1}^{N} e_{is}} \right). \tag{3.18}$$

Apparently the computation is too expensive if it goes through all the sample pairs. However, in Publication 5 we have shown that the gradient can be computed by matrix operations after the formula rearranged:

$$\nabla \mathcal{J}_{\mathrm{IDA}}(\mathbf{W}) = \mathbf{X}(\mathbf{D} - \mathbf{H})\mathbf{X}^T \mathbf{W}, \tag{3.19}$$

where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^{N} H_{ij}$. As there exist fast algorithms that implement matrix multiplication [29], the gradient computation (3.19) runs much faster than the original IDA optimization algorithm.

### 3.3.2 Geodesic flows in the Stiefel manifold

Orthonormality of the transformation matrix in (3.1) is preferred in feature extraction because it enforces the matrix to encode the intrinsic subspace in the most economic way. The orthonormality constraint also prevents the learning algorithm from falling into some trivial local minima.

The set of $M \times R$-sized real orthonormal matrices forms a Stiefel manifold $\mathrm{St}(M, R)$ [36]. Given the gradient $\nabla$ of $\mathcal{J}$ at $\mathbf{W}$, it has been shown [82] that

the natural gradient in such a manifold is given by

$$\mathrm{grad}_{\mathbf{W}}^{\mathrm{St}(M,R)}\mathcal{J} = \nabla - \mathbf{W}\nabla^T\mathbf{W}. \tag{3.20}$$

An approximated geodesic learning flow with the starting point $\mathbf{W}$ is then

$$\mathbf{W}_{\mathrm{new}} = \mathrm{expm}\left(\eta\left(\nabla\mathbf{W}^T - \mathbf{W}\nabla^T\right)\right)\mathbf{W}, \tag{3.21}$$

where expm represents the matrix exponential [45] and $\eta$ a usually small positive learning rate. In Publication 5 we have applied the geodesic flow technique to the conventional gradients (3.17). Yet, the update rule (3.21) can be applied to learning directions other than gradients, including the *Principal Whitened Gradients* proposed in Publication 8 and reviewed next.

### 3.3.3 Principal Whitened Gradient

In Publication 5 we directly applied the steepest ascent method [113] with gradients calculated by (3.19). The resulting optimization speed is only mediocre because the parameter space is not Euclidean. It has been pointed out by Amari that the geometry of the Riemannian space must be taken into account when calculating the learning directions [2]. In Publication 8 we derived a more efficient optimization method that approximates a geodesic flow and maximizes local information change at each iteration.

Many statistical inference problems can be reduced to probability density estimation. *Information geometry* proposed by Amari [3] studies a manifold of parametric probability densities $p(\mathbf{x};\boldsymbol{\theta})$, where the Riemannian metric is defined by the Fisher information matrix [105]

$$[\mathbf{G}(\boldsymbol{\theta})]_{ij} = E\left\{\frac{\partial\ell(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial\ell(\mathbf{x};\boldsymbol{\theta})}{\partial\theta_j}\right\}. \tag{3.22}$$

Here $\ell(\mathbf{x};\boldsymbol{\theta}) \equiv -\log p(\mathbf{x};\boldsymbol{\theta})$. Amari also applied the natural gradient update rule

$$\boldsymbol{\theta}^{\mathrm{new}} = \boldsymbol{\theta} - \eta\mathbf{G}(\boldsymbol{\theta})^{-1}\nabla\mathcal{J}(\boldsymbol{\theta}) \tag{3.23}$$

for the optimization in the information geometry by using $\mathcal{J}(\boldsymbol{\theta}) = \ell(\mathbf{x};\boldsymbol{\theta})$ as the online objective function which is equivalent to the maximum likelihood approach [2]. Similarly, the batch objective function can be defined to be the empirical mean of $\ell(\mathbf{x}_i;\boldsymbol{\theta})$ over $i$.

It is worth to notice that the Fisher information matrix is the covariance of ordinary online gradients. In Publication 8, it has been shown that this property is

invariant of linear transformations in the sense that the Fisher information metric measures the local Kullback-Leibler divergence between the current estimate and updated estimates. This allows us to look for a proper linear transformation matrix $\mathbf{F}$ for facilitating the optimization. By setting $\mathbf{F}$ as the whitening matrix $\mathbf{G}^{-\frac{1}{2}}$, the transformed Fisher information metric becomes identity. That is, the whitening matrix locally transforms the Riemannian space into its Euclidean embedding, which leads to an ordinary steepest descent update rule in the whitened tangent space:

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta} - \eta \tilde{\nabla} \mathcal{J}(\boldsymbol{\theta}), \tag{3.24}$$

where $\tilde{\nabla} \mathcal{J}(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta})^{-\frac{1}{2}} \nabla \mathcal{J}(\boldsymbol{\theta})$ is the *whitened gradient*. The above update rule has a form similar to the natural gradient one (3.23) except the square root operation on the eigenvalues of the Fisher information matrix at each iteration. Such a difference brings two distinguished advantages. First, the additive updates using (3.24) better approximate the geodesic flow to the optimum because additive updates are equivalent to exponential maps when the Riemannian metric becomes Euclidean. Second, selecting the learning rate for the whitened gradient updates is as easy as for the ordinary steepest descent approach because $\eta$ in (3.24) is independent of the local Riemannian metric.

Another strategy to improve the natural gradient is to remove the minor components during the whitening procedure. In Publication 8, it has been shown that the principal direction that maximizes the variance of the projected gradient coincides with the one that maximizes local information change. By this motivation, we preserve only the principal components and discard the minor ones that are probably irrelevant for the true learning direction. The whitening procedure through e.g. singular value decomposition thus runs much faster than inverting the whole matrix $\mathbf{G}$ when the number of principal components is far less than the dimensionality of $\boldsymbol{\theta}$ [29]. Furthermore, we proved in Publication 8 that the Fisher information matrix is always singular when optimizing an objective of the form like (3.12) if $\mathbf{W}$ as the parameter consists of multiple columns. In this case, dimensionality reduction is therefore necessary during the whitening stage. We thus obtained the *Principal Whitened Gradient* (PWG) update rule:

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta} - \eta \hat{\mathbf{G}}(\boldsymbol{\theta})^{-\frac{1}{2}} \nabla \mathcal{J}(\boldsymbol{\theta}), \tag{3.25}$$

where $\hat{\mathbf{G}}(\boldsymbol{\theta})^{-\frac{1}{2}}$ is the whitened matrix with only principal components.

The Principal Whitened Gradient method can be applied to the IDA optimiza-

tion. Let us first separate the class-dependent and -independent parts in (3.17):

$$\begin{aligned}\nabla \mathcal{J}_{\mathrm{IDA}}(\mathbf{W}) =& \frac{1}{2}\sum_{k=1}^{Q}\sum_{i=1}^{N}\sum_{j=1}^{N} B_{ij}^{k}\left(\mathbf{x}_i-\mathbf{x}_j\right)\left(\mathbf{x}_i-\mathbf{x}_j\right)^T\mathbf{W} \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} B_{ij}\left(\mathbf{x}_i-\mathbf{x}_j\right)\left(\mathbf{x}_i-\mathbf{x}_j\right)^T\mathbf{W}, \end{aligned} \tag{3.26}$$

where

$$B_{ij}^{k} = -\frac{\delta(c_i,k)}{\sigma^2}\frac{\phi_{ij}e_{ij}}{\sum_{s=1}^{N}\phi_{is}e_{is}}, \tag{3.27}$$

$$B_{ij} = -\frac{1}{\sigma^2}\frac{e_{ij}}{\sum_{s=1}^{N}e_{is}}. \tag{3.28}$$

Denote $\mathbf{w}$ a column of $\mathbf{W}$ for notational simplicity. After some mathematical manipulations, the online gradients of $\mathcal{J}$ with respect to $\mathbf{w}$ can be obtained from the columns of

$$\mathbf{X}\left[\mathrm{diag}(\mathbf{X}^T\mathbf{w}) + \mathbf{X}^T\mathbf{w}\mathbf{1}^T \circ \mathbf{B}^T - \mathrm{diag}(\mathbf{B}\mathbf{X}^T\mathbf{w}) - \mathbf{B}^T\mathrm{diag}(\mathbf{X}^T\mathbf{w})\right], \tag{3.29}$$

where $\mathrm{diag}(\mathbf{a})$ creates a diagonal matrix $\mathbf{A}$ from a vector $\mathbf{a}$ with $A_{ii} = a_i$, $\mathbf{1}$ is the $M$-dimensional vector with all elements equal to one, and $\circ$ stands for the element-wise product. The online gradients for the $k$-th class can be obtained by similar calculation with $\mathbf{B}^k$. The above calculation is repeated for each column of $\mathbf{W}$. With the online gradients, one can compute the Fisher matrices $\mathbf{G}$ and $\mathbf{G}_k$, $k = 1,\ldots,Q$. Then the learning direction is the vectorized representation of $\nabla\mathcal{J}_{\mathrm{IDA}}(\mathbf{W})$ left-multiplied by the principal whitened components of $\mathbf{G} - \sum_k p(k)\mathbf{G}_k$.

## 3.4 Discriminative Self-Organizing Maps

As discussed in Section 2.5.4, filtering by ground truths or classifiers is only suitable for the aspects with rigorous classification. However, for some aspects like *race* or *hairstyles* the categorization by different users may greatly vary. In these cases, the classification boundaries are not rigorous but vague. Filtering by hard classifiers would therefore prevent the appearance of many relevant images.

In Publications 3, a novel method which adaptively learns partial relevance during the interactive retrieval has been proposed. The solution reuses the PicSOM

framework by just replacing the normal SOMs by aspect-oriented feature maps called *Discriminative Self-Organizing Maps* (DSOMs), which are of enhanced discriminative power by incorporating the supervised information. In other words, a DSOM for a specific aspect is a Self-Organizing Map on which the semantic classes are densely spread in nearly separate areas. Figure 4.3 (b-d) in Section 4.4 visualizes the concept of DSOMs on two semantic classes, *race:Asian* and *race:non-Asian*.

In a DSOM, the images of a semantic class are very likely connected to map units that situate nearby. If some images in the semantic class are displayed and marked as relevant, PicSOM's relevance feedback learning algorithm on the DSOM will most likely award higher scores to the other images of the same class, and then show these images to the user in the imminent rounds.

Different ways for obtaining a DSOM exist. One can replace the Euclidean metric in the normal SOM training by a Riemannian one that incorporates the supervised information [55, 90, 88]. The resulting *learning metric* approach however requires modifying the SOM learning algorithm. In addition, its training and inference are computationally more expensive. A more straightforward method is to apply supervised learning as a preprocessing step before the SOM training. This work employs the latter approach because it requires no change in the SOM training algorithm and in PicSOM's online query implementation. Given the projected feature vectors $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$. The normal SOM training described in Section 2.3.1 can then be performed in the $\mathbb{R}^R$ subspace with the $\mathbf{y}_i$ vectors to generate a Discriminative SOM.

# Chapter 4

# Interactive Facial Image Retrieval

Biometric identification by visual appearance is helpful for searching, for example, a criminal or a lost person. The task is to recover a pictorial presentation of the target subject. The procedure is usually done in an interactive manner if only an eyewitness's recalling is available. A satisfying facial image can then hopefully be obtained after multiple rounds of interaction.

This chapter demonstrates the application of the discriminative learning techniques described in the previous chapters which adapt the PicSOM ICBIR system towards a satisfactory interactive facial image retrieval system. Simulations on the FERET database are provided, including a concrete query example and benchmark test results. The empirical study indicates that the occurrence of the first subject hit can be significantly speeded up in the improved PicSOM system.

## 4.1 Related work

The history of interactive facial appearance acquisition can be traced back to decades ago, far before computers were available. A human portraitist drew a picture according to the appearance description by the eyewitness. Due to descriptive gaps, a face-to-face discussion between the portraitist and the eyewitness is necessary to ensure the quality of the drawing. The portraitist must

be professionally trained because the drawing work as well as the discussion is highly skillful. Unfortunately, such a professional drawer might not be available in urgent cases. Furthermore, the interactive drawing procedure usually takes too much time and the eyewitness may feel tired or find the procedure unpleasant.

The face drawing work can be approximated by using *composition methods*, where the face is divided into several parts such as the mouth, nose, eyes, jaw and hair. In this scenario, the eyewitness is asked to select among several transparencies of each facial part. A face collage is then obtained by piling up the selected face parts. With the development of computer imaging, the composition quality can be improved by using a face fusing software such as E-FIT[1], FACES[2], and FACETTE[3]. No discussion between the eyewitness and a face expert is needed. The witness specifies the types for all facial parts, and then the system fuses these parts and returns a face-like composite. If the returned figure is not satisfactory, the eyewitness has to change one or more facial parts and to repeat the fusing. Although composite methods are popularly used, they have several significant drawbacks. The primary difficulty of using such a system is the *evaluation fatigue* problem which has been discussed in Section 2.5.1. The composite approach requires the user to choose among hundreds or even thousands of facial parts, but in many cases the eyewitness' recalling is not strong enough to identify the exact type of each part. Thus the user must resort to a trial-and-error procedure and probably gets tired in the time-consuming process. Another shortcoming of the composition systems is that as the created facial images are virtual they cannot be coupled with the identity information of the depicted person.

An alternative to the above methods, *Interactive Facial Image Retrieval* (IFIR) seeks images that depict the target subject from a large collection of genuine face images based on relevance feedback from the user. Such example-based systems output facial images of real persons instead of merely face-like composites. That is, it is straightforward for IFIR systems to return the identity associated with the output image, by which one can obtain more relevant images or other personal information about the person. Moreover, unlike the *ad hoc* fusing approach that requires a great deal of prior knowledge about the facial parts, an IFIR system can be built as an agnostic solution of content-based image retrieval. Many present CBIR techniques can thus be applied to improve the performance of IFIR.

Some IFIR approaches have emerged in recent years. The FACERET system proposed by Ruiz-del-Solar and Navarrete [103] calculates a ranking for images on a single SOM trained with eigenface features [121]. In [135], Yuen and Man

[1] http://www.efitforwindows.com/
[2] http://www.iqbiometrix.com/
[3] http://www.facette.com

combined the face sketch recognition and IFIR techniques, where the relevance feedback learning is performed by the Fisher's Linear Discriminant Analysis [27]. An interactive mental face retrieval system by Fang et al. maximizes the mutual information between the target and the responses to the displayed images [25]. However, the above methods only provide heuristic solutions under an ideal response model in which the user's choice is perfectly consistent with the system metric. In practice, the user's selection criteria can be arbitrary and such consistence may not hold. Thus, most existing IFIR systems may fail to return an image of the target subject in an acceptable number of rounds.

## 4.2 Two-phase interactive facial image retrieval

Interactive facial image retrieval aims at finding all images of a specific subject in a facial image database. The whole retrieval procedure can be divided into two independent phases as already discussed in Section 2.5.2: (1) looking for one of the images that depict the subject, and (2) retrieving other images of that subject.

The major challenge of the IFIR problem lies in the first phase. Unlike CBIR systems on general images, the query precision on facial images suffers from the inherent problem of extremely small class sizes, as discussed in Publication 2. In a popular facial image collection, FERET [93], most subjects possess only two frontal images. As there are 2409 images of 867 subjects in the FERET database, the average *a priori* probability of each subject is close to 0.001. If only images that depict the correct person are regarded as relevant, many zero pages (i.e. all images in these rounds are non-relevant) will be displayed before the first relevant image emerges. This is because the negative responses from the user in the early retrieval rounds provide only little semantic information and—as a result—the iteration progresses in a nearly random manner. The first retrieval phase finishes when the *first subject hit* appears. Therefore, making the emergence of the hit as fast as possible is critical for the success of interactive facial image retrieval.

The retrieval in the second phase is of minor research interest if the identity of each portrayed subject is also stored in the database. By this identity one can easily locate the other images of the same subject if they really are needed at all. If such information is not available, the second phase can optionally be treated as a face recognition problem—looking for the most similar faces of the first-found relevant image. Only the interactive approach to obtain the rest of the images depicting the target subject will be in the topic of this thesis. The interactive retrieval allows inclusion of some images of other subjects, but close

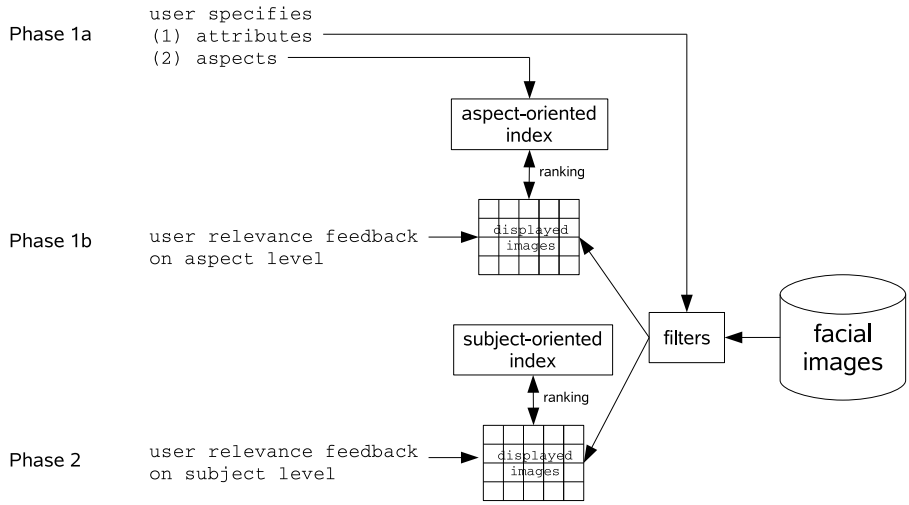


Figure 4.1: System diagram for the two-phase interactive facial image retrieval.

to the eyewitness' recalling, as well as exclusion of faces belonging to the target subject, but unfamiliar to the eyewitness. In many cases, the interactive approach in the second phase can serve as a supplemental verification method and enhance the robustness of the retrieval.

The system diagram shown in Figure 4.1 illustrates the process flow of two-phase interactive facial image retrieval. First in Phase 1a the user supplies the partial relevance specification by choosing the aspects the eyewitness remembers, such as gender or race. For the selected aspects with rigorous categorization, the user also specifies the attributes that indicate the semantic class membership. The system then employs the corresponding deterministic or stochastic filters to preselect the facial image candidates that can be displayed to the user. For the aspects with vague categorizations, the system loads the corresponding aspect-oriented image indices. Then in Phase 1b the iterative query by pictorial examples starts, in which the user provides partial relevance information at each round. The system calculates the ranking by using these feedbacks and then displays the unseen images with the highest scores. Such interaction proceeds until one of the correct images is found. The retrieval in Phase 2 is carried out in a similar manner except the relevance feedback is given at subject level and the image indices become subject-oriented.

In other words, the extended PicSOM system implements from-coarse-to-fine

| aspect | semantic classes (images/subjects) | | method | raw feature |
|---|---|---|---|---|
| gender | male (1495/501) | female (914/366) | ground truth | — |
| eyeglasses | yes (262/126) | no (2147/741) | SVM | innerface |
| mustache | yes (256/81) | no (2153/786) | SVM | innerface |
| beard | yes (144/51) | no (2265/816) | SVM | mpegface |
| race | white (1541/558) | Asian (388/131) | DSOM | innerface |
| | black (199/72) | misc (371/106) | | |
| hairstyle | (see Table 4.2) | | DSOM | mpegface |
| all | 2409/867 | | | |

Table 4.1: Statistics of the used aspects and respective semantic classes.

IFIR: 1) Images are first filtered using their associated textual metadata; 2) The qualified image candidates are then filtered by highly accurate classifiers; 3) The qualified images after the filterings 1) and 2) are interactively retrieved with DSOMs. In Section 4.6 we shall see that the evaluation labor in the first retrieval phase can be significantly reduced using such a cascade-like framework.

## 4.3 Data

We have used the FERET database [93] of facial images to evaluate the discriminative learning techniques described in Chapter 3 within the PicSOM system. After face segmentation, 2409 frontal facial images (poses "fa" and "fb") of 867 subjects were stored in the database for the experiments. The average number of images belonging to a subject is 2.78, and most subjects have only two images.

We used the ground truth data of five deterministic aspects from the FERET collection: *gender*, *eyeglasses*, *mustache*, *beard* and *race*. The statistics of the respective semantic classes are shown in Table 4.1. In addition, we manually labeled four fuzzy *hairstyles* of the FERET images. The images were categorized into three groups[4]: *Yes*, *In-between* or *No*, according to the ternary membership scale of each hairstyle *long*, *short*, *bald* or *capped*, of which the statistics are shown in Table 4.2. For example, the typically bald faces were labeled as *bald:Yes*, the typically non-bald ones as *bald:No* group, and the unsure ones as *bald:In-between* group. We then calculated the class matrix (3.15) by such grouping information: for class $c$, $M_{ic}$ equals 1, 1/2 or 0 if the $i$-th image is labeled *Yes*, *In-between* or *No*, respectively.

[4]available in http://www.cis.hut.fi/projects/cbir/hairstyles.zip

| hairstyle | labeling statistics (images) | | |
|---|---|---|---|
| long | Yes (469) | In-between (317) | No (1623) |
| short | Yes (1240) | In-between (515) | No (654) |
| bald | Yes (153) | In-between (170) | No (2086) |
| capped | Yes (32) | In-between (0) | No (2377) |

Table 4.2: Statistics of the hairstyle semantic classes.

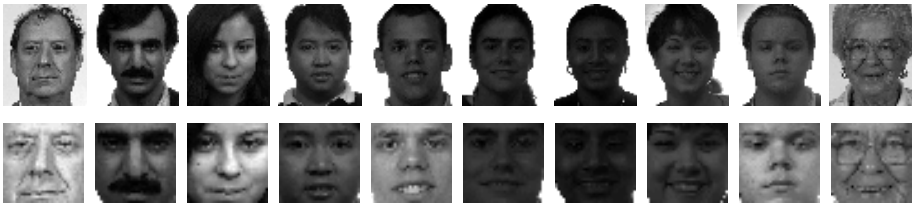

Figure 4.2: Preprocessed example images from FERET. Top: *mpegfaces* cropped according to the MPEG-7 standard. Bottom: *innerfaces* of the same images.

In our experiments the coordinates of the facial parts (eyes, nose and mouth) were acquired from the ground truth data of the FERET collection, with which we calibrated the head rotation so that all faces are upright. Afterwards, all face boxes were normalized to the same size $46 \times 56$, i.e. $M = 2576$, with fixed locations for the left eye (31,24) and the right eye (16,24) in accordance to the MPEG-7 standard [77]. To reduce irrelevant variability, we also used the inner part of faces that are normalized to size $32 \times 32$, i.e. $M = 1024$, with fixed locations for the left eye (26,9) and the right eye (7,9). Later we will call these two kinds of raw features as *mpegfaces* and *innerfaces*. Example FERET images after these normalizations are shown in Figure 4.2.

In our experiments, we supposed the gender information is available for all images. Thus the images of wrong gender could accurately be filtered out. For the other aspects, half of the data was used for training and the rest for testing. The division of the subsets was based on subjects. That is, the images of each subject appear either in the training set or testing set, but not in both. Such division can better reveal the generalization performance of a supervised learning method. In the training and testing sets we had images of 434 and 433 subjects, respectively.

We implemented the filtering by two-class SVM classifiers for *eyeglasses*, *mustache* and *beard*, and applied the DSOM technique to the partial relevance of *race* and *hairstyle*. The parameters in the supervised learning methods were obtained by 5-fold cross-validations with the training set. The raw features used for the aspects are shown in Table 4.1.

## 4.4 Training of DSOMs

For the DSOMs of *race* and *hairstyle*, we first obtained the projection matrix $\mathbf{W}$ of (3.1) by applying discriminant analysis to the training data of zero-mean. All images, both training and testing, were then projected to an $R$-dimensional space, in which we performed a normal SOM training (2.2) with the Euclidean metric. We have compared two discriminant analysis methods, PCA+LDA (3.9) and PDA (3.10), for each aspect. For PCA+LDA we set the number of principal components to 200, which captures 95% of the variance of the training images. LDA outputs at most two discriminants for *race* and three for *hairstyle*. Nevertheless, we found that two discriminants for *hairstyle* perform better than three by the trained SOM, probably because the *capped* class is very small. Therefore we set the number of discriminants of both aspects to two. The PDA learning started from the PCA+LDA results. We have used TS-SOMs of three levels of sizes $4 \times 4$, $16 \times 16$, and $64 \times 64$ units.

The images of the testing set are visualized on the trained DSOMs in Figure 4.3. The SOM trained with eigenfaces (Figure 4.3a) has poor discrimination in separating the races (white and black dots) as the white dots corresponding to *Asian* faces scatter all over the feature map. For the DSOM trained by using PCA+LDA (Figure 4.3b), one can perceive that the *Asian* class mainly resides in the upper part, especially in the upper right corner of the map, but there are still many outliers that drift away and are mixed with the *non-Asian* faces. The discrimination is further improved by the Parzen Discriminant Analysis described in Section 3.2.2. There are much fewer outliers in Figure 4.3c and the Asian class distributes more densely in the upper right corner. Figure 4.3d illustrates the faces in the DSOM using PDA features. It can be seen that the typical images of *Asian*, *Black*, and *White* distribute separately in the upper right corner, upper left corner and the lower part, respectively. The in-between images can be found around the class boundaries.

## 4.5 A query example

In this section we provide a query example to illustrate the IFIR procedure using the extended PicSOM system. Suppose one is looking for a lost woman shown in Figure 4.4a.[5] An eyewitness who has seen a likely person is asked to help the identification task, in which the PicSOM system plays a role in retrieving the

[5]This example is provided only for the illustrative purpose. All the assumptions here are virtual and the simulation has nothing to do with the real persons in the displayed images.

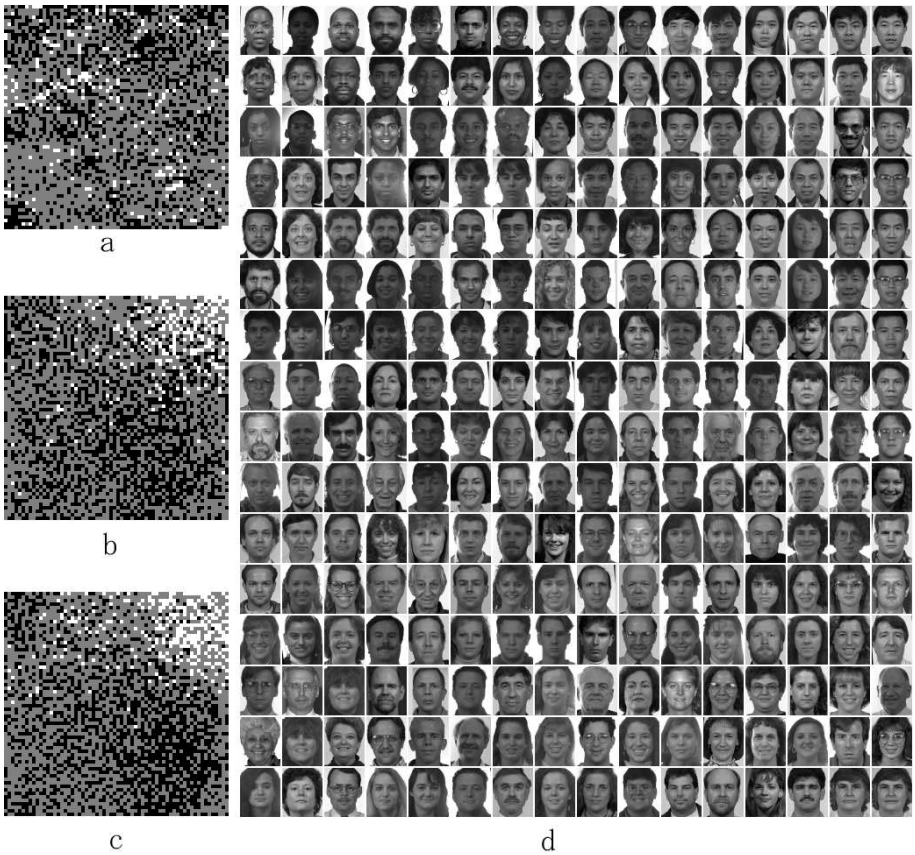


Figure 4.3: The distributions of the *Asian* class on the bottom level ($64 \times 64$) of the trained TS-SOMs: (a) eigenface, (b) PCA+LDA, and (c) PDA. The white dots stand for the Asian faces and the black for the non-Asian. The gray dots represent unoccupied units. (d): Representative faces on the intermediate level ($16 \times 16$) of the TS-SOM trained by the PDA-extracted features.

images of the seen person by the eyewitness recalling.

First a questionnaire-like user interface of Figure 4.4b, corresponding to Phase 1a of Figure 4.1 is presented, where the eyewitness checks the aspects she or he remembers. Here we assume the eyewitness recalls all the other aspects but the eyeglasses in this example. The PicSOM system then applies filtering by the deterministic *gender:Female* ground truth. The *mustache* and *beard* classifiers are redundant for searching a woman. PicSOM thus loads the DSOMs for the *race* and *hairstyle* aspects.

Among the images displayed in the first interactive query round, the eyewitness ticks the relevant ones for the specified aspects: one for *race* and four for *hairstyle* (Figure 4.4c). The relevance feedback is fed to the PicSOM ranking algorithm which then outputs the most probable unseen images in the second round (Figure 4.4d). From the retrieved images, the eyewitness recognizes that the second rightmost one in the last row depicts the correct woman and the first phase finishes.

If the user wants to continue the search for more images of the same woman, the retrieval proceeds to Phase 2 of Figure 4.1, where the relevance criterion is set to the subject aspect. One can see that all the other three images of the same woman can then be found in the first round of the second phase.

## 4.6 Quantitative simulation study

A substantial number of queries from human users would be required to evaluate the performance of an IFIR system in a realistic retrieval setting. Therefore, it is desirable to quantify the performance in offline simulations in a standardized and non-subjective benchmark.

The evaluation of an ICBIR system should involve human users. However, a small number of samples may induce great bias due to subjective difference in assessing similarity among images. Automated evaluation has therefore been employed in this work to avoid such subjectivity. In each retrieval round, 20 images were "displayed"— supplied to the virtual user, a computer program, which assesses the image similarity by using associated ground truth metadata. We repeated the retrieval for each of the 433 testing subjects ten times with different random selections of initial images in the first round.

We compared the following six approaches of IFIR based on the PicSOM framework:

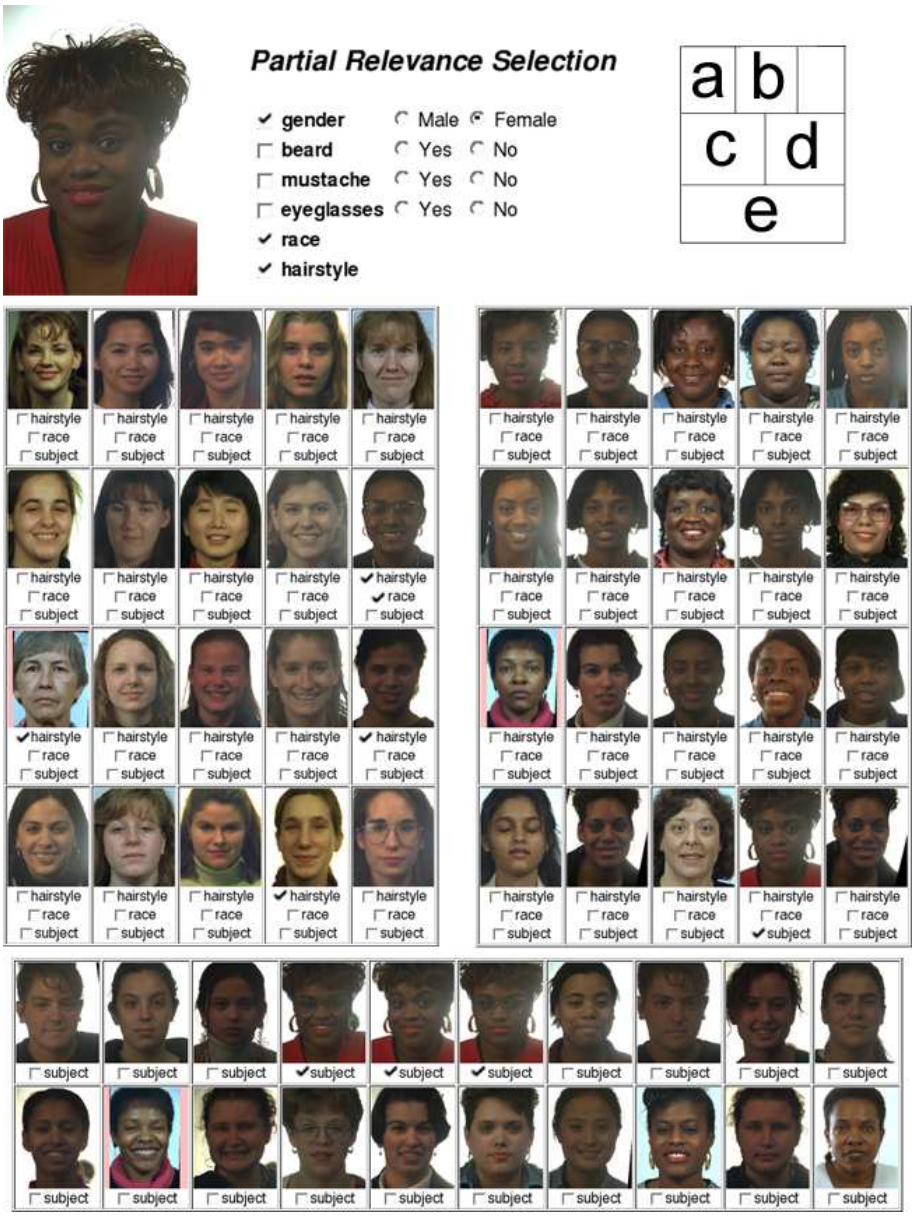


Figure 4.4: A query example using the PicSOM IFIR system: (a) the target person in search; (b) the user interface for specifying the aspects of partial relevance; the displayed images in the first IFIR phase, in (c) the first round and (d) the second round; (e) the images displayed in the first round of the second IFIR phase, where the other images depicting the same subject appears as the fourth to sixth in the first row.

(1) Simulated *FACERET* [103] that uses eigenfaces for normal SOM training;

(2) *Fisherface* that uses Fisherfaces for normal SOM training;

(3) *filter+predict* is the same as (2) except that it filters images by the *gender* ground truth and by SVM-predictions of *mustache*, *beard*, *eyeglasses*, and *race* aspects before iterative retrieval;

(4) *filter+predict+race* that applies the same filtering as (3) for *mustache*, *beard*, *eyeglasses*, but implements partial relevance of *race* through a DSOM created by PCA+LDA.

(5) *filter+predict+race+hairstyle* that is same as (4) except the inclusion of the partial relevance of *hairstyle* through a DSOM created by PCA+LDA.

(6) *PDA* which is same as (5) except that the two DSOMs have been trained with Parzen Discriminant Analysis.

The first three approaches are unable to make use of the available aspect-wise partial relevance feedbacks while the fourth to the sixth genuinely employ the two-phase iterative retrieval principle. The DSOMs created by using Fisherfaces are loaded in the second phase for the approaches (2)–(5), while the *PDA* method (6) employs the Parzenface method which replaces PCA+LDA with PDA.

Two performance measures described in Section 2.4 have been used: the FRHA statistic addresses the efficiency in the first retrieval phase, while the second compares the conventional precision/recall ratios in both phases.

### 4.6.1 First Subject Hit Advantage

The automated evaluation loops over all subjects, each of which serves as the ultimate relevant class. Thus we rename the FRHA measurement as *First Subject Hit Advantage* (FSHA) in this chapter for clarity. Accordingly, denote $\varrho_t$ the FSHA of of the $t$-th subject class.

We found that the density of FSHA is asymmetric and heavy-tailed. The arithmetic mean of FSHAs is dominated by large $\varrho$'s in the tail and thus has high variability. Alternatively, we have found the distribution of log-FSHAs is roughly symmetric around its mean, which motivates us to compare the retrieval results through the means of log-FSHAs instead of the original FSHAs. Furthermore, the arithmetic mean of log-FSHAs can be connected to FSHAs' geometric mean

by an exponential function:

$$\left(\prod_{t=1}^{K} \varrho_t\right)^{1/K} = \exp\left(\frac{1}{K}\sum_{t=1}^{K} \log \varrho_t\right), \tag{4.1}$$

where $K$ is the number of subjects, i.e. $K = 433$, in our testing data set. Figure 4.5 shows the error bars of the resulting log-FSHAs and the corresponding geometric means with the six studied ICBIR methods. The middle point of each error bar represents the mean of $10 \cdot 433 = 4330$ experiments and the extremes indicate the range of one standard deviation.

The method (2) that replaces eigenfaces by Fisherfaces cannot accelerate, but even hinders the early occurrence of the first subject hit. An expected rise is achieved by the method (3) which incorporates the *gender* ground truth. However, the overall speedup is still mediocre due to the SVM classification errors of the other involved aspects. From the method (4) on, we implement the *race* partial relevance by the DSOM technique, which leads to another significant increase of performance. A still higher mean of FSHAs can be achieved by including the *hairstyle* partial relevance in the method (5). Finally, a further improvement is obtained by replacing all PCA+LDA with PDA features in method (6).

The best mean FSHA we obtained is 5.46, the meaning of which can be interpreted by a simple calculation. We used $1,210$ FERET images for testing, and recall that there are 2.78 images per subject. Inserting these numbers into the FSHA definition (2.7), we can deduce that the corresponding first subject hit is expected to take place at index

$$i = \frac{1210 - 2.78}{(2.78 + 1) \cdot 5.46} = 58.5. \tag{4.2}$$

That is, one can on average retrieve the first image of the target subject within three rounds if 20 images per round are displayed.

The above analysis shows that our approach outperforms two other IFIR approaches [135, 25] in the average case. The compared two systems were also experimented with the FERET face database. In [25], the smallest number of displayed images until the first subject image appears is 104 if four images are presented to the user at each iteration. The experimented image set used in [135] contains only 432 images of 72 subjects, which is only about 1/3 of our test set. However, the average rank of the first target image is no less than one fifth of the size of the database, i.e. about 90. It should be noticed that both [135] and [25] employ supervised feature extraction of facial images, but how the training set and testing sets have been divided in their experiments remains unclear.

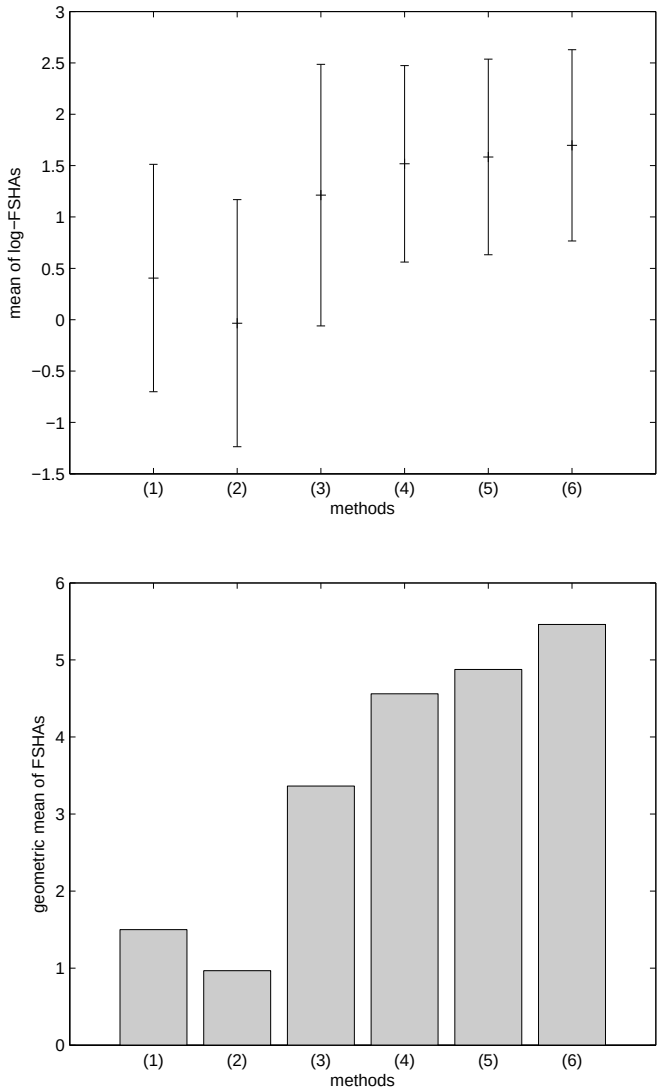


Figure 4.5: Top: Error bars of log-FSHA. Bottom: geometric means of FSHAs. The methods used are (1) *FACERET*, (2) *Fisherface*, (3) *filter+predict*, (4) *filter+predict+race*, (5) *filter+predict+race+hairstyle*, (6) *PDA*.

### 4.6.2 Precision versus recall

We have also recorded the recall and precision performance statistics after each round of 20 "displayed" images. The overall performance statistics were obtained by averaging those from the 433 individual testing subjects. We repeated the experiments ten times with different random seeds and recorded the average results.

The plots of precision vs. recall are shown in Figure 4.6. For clarity the curves have been divided in two groups. The first group includes the results of the two methods, *FACERET* and *Fisherface*, without partial relevance. The use of Fisherfaces instead of eigenfaces results in a slight improvement in the middle recall levels, but the overall precisions remain very low with the maximum less than 0.01.

By contrast, we can see in the second group that the precisions obtained by using partial relevance are much higher than the baseline FACERET method at all recall levels. Starting from a low precision level due to the random initial set, the curves rise clearly and the precision maxima occur between the recall values 0.2 and 0.4, which reveals that the learning from the relevance feedback is taking effect. The mutual ordering of the curves also verifies the results in Figure 4.5 with different kinds of partial relevance.

Because hard classification may prevent the appearance of some true positive images, the performance by using the *filter+predict* method (3) is mediocre, especially in the high recall levels. This disadvantage can be overcome by replacing the SVM classifiers of *race* and *hairstyle* with DSOM-based fuzzy retrieval. Further improvement can be obtained by employing PDA, a more advanced discriminant analysis. PDA significantly outperforms PCA+LDA especially in the high recall levels which correspond to the second retrieval phase.

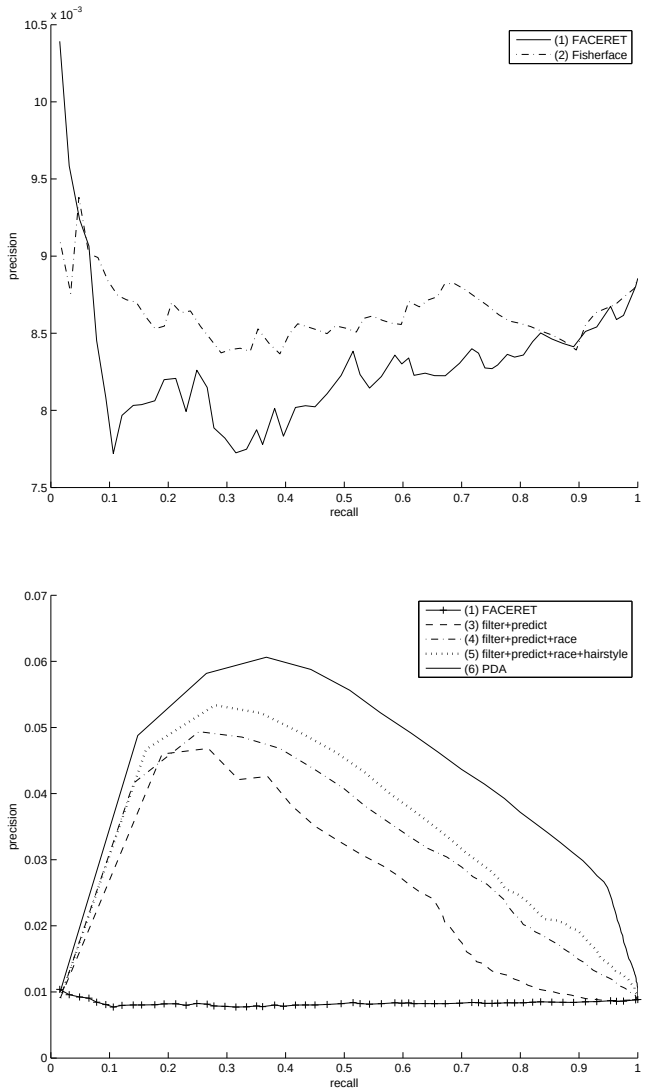


Figure 4.6: Precision vs. recall curves by using the six compared methods.

# Chapter 5

# Conclusions

Content-based image retrieval is a method to overcome the labor and subjectivity in manual image annotation and to implement image searching by using automatically extracted visual features. Relevance feedback has later been introduced to implement interactive content-based image retrieval systems and to reduce the semantic gap between high-level semantic concepts and the low-level visual features. With Tree-Structured Self-Organizing Maps, the PicSOM ICBIR system is able to index large-scale image databases. In addition, relevance feedback can be accomplished efficiently in a local manner.

Conventional relevance feedback approaches require a vast amount of user interaction data, especially when the size of the target class is extremely small. This often leads to the evaluation fatigue problem. To solve this, an ICBIR system should allow the user to supply partial knowledge of aspects, such as gender or hairstyle of the searched person, which probably divide the whole collection into several semantic classes. The search target can then be characterized by only one or an intersection of some such semantic classes. This provides multiple levels of relevance, which is beneficial as aspect-wise partial relevance in early retrieval phases is helpful for gradually reducing the number of displayed images. In this way, partial relevance implements from-coarse-to-fine retrieval and it is probably consistent with human searching behavior.

Partial relevance can be implemented by using the state-of-the-art computer vision techniques and by utilizing supervised information available in the image collection. In this work, the PicSOM retrieval system has been extended in early phases by replacing the membership of a target class with that of a se-

mantic class as the new relevance criterion until the first relevant hit appears. Supervised learning has been applied so that the resulting SOMs have stronger discriminative power. Such an implementation guarantees that the retrieval is theoretically sound and facilitates interpretation.

A straightforward method to obtain discriminative Self-Organizing Maps is to employ discriminant analysis as a preprocessing step before normal SOM training. However, direct application of discriminant analysis on high-dimensional image data yields poor results. Unsupervised dimensionality reduction is therefore required to obtain compact representations of images in a particular domain. The PCA+LDA method, which is widely used in pattern recognition, has been applied as the baseline in this work. Beyond the first and second order statistics, the Informative Discriminant Analysis is adopted to maximize the discrimination for more complicated distributions. Furthermore, the presented Parzen Discriminant Analysis regularizes the IDA objective by emphasizing the prior of piecewise smoothness in images. Both LDA and PDA have been extended in this work for handling fuzzy cases.

The original IDA optimization algorithm is computationally expensive. Three acceleration strategies have been presented: First, the computation cost of batch gradients is reduced by using matrix multiplication. Second, the updates follow a geodesic flow in the Stiefel manifold without Givens reparameterization. Third, a more efficient learning direction is calculated by preserving only the principal whitened components of the batch gradient at each iteration.

In this thesis, an application example, interactive facial image retrieval, has demonstrated the use of the proposed techniques. Simulations were performed on a popularly used image database, FERET. In addition to a concrete query example, the work has included a quantitative study on the obtained advantage in terms of the occurrence of the first subject hit as well as retrieval precisions at various recall levels.

The research on ICBIR is still going on. The elements of the presented solution and analysis can be further improved. People tend to perceive visual patterns from coarse overview to fine details. On the other hand, discriminative patterns generally are meaningful only in a specific domain. Therefore, a more advanced image representation method should utilize both information in multi-resolutions and statistics of the image domain in analysis.

The image variability is limited at this stage due to the state-of-the-art of current computer vision techniques. When more advanced methods are developed, more difficult data sets, such as the Face Recognition Grand Challenge (FRGC)

database [1], could be studied. Potential research directions on images without pre-registration should be investigated. In particular, more representative methods might be used to generate more invariant features.

For the discriminant analysis, some nonlinear feature extraction techniques may further improve the performance of resulting DSOMs. Moreover, the ideal supervised learning must overcome the problem of small-sized training sets, where some discriminative patterns, for instance, scars and moles in facial images, are difficult to be encoded using low-order statistics. In the optimization with Principal Whitened Gradient, the whitening matrix and the batch gradient are computed separately. Actually one may achieve a direct and faster method for computing their product, for example, by adopting online Principal Component Analysis [131]. Another potential extension of the PWG update rule is to make it accommodate additional constraints such as orthonormality or sparseness. Furthermore, many conventional optimization techniques, such as the conjugate gradient, can be applied in the Euclidean embedding to further improve the convergence speed.

In addition to discriminative power, the topological ordering of the within-class clusters also plays an important role for the Discriminative Self-Organizing Map. Some potential methods, like Discriminative Clustering [54], could discover this kind of information, by which one would be able to devise a more user-friendly interface for partial relevance specifications.

This thesis has focused on content-based image retrieval, which, however, does not prevent the use of other information sources such as text annotations. Instead, incorporating such information in a multi-modal retrieval system would be conducive to designing better user interfaces and to developing more efficient retrieval.

Interactive content-based image retrieval can be further accelerated with a considerately initialized image set. For example, the IFIR system could be used together with facial image composition systems. In the presented empirical study, PicSOM randomly selected the starting images. Actually the retrieval performance could be improved by using a considerate initial set of images. A potential way to achieve this is to use the fused result from an image composition system. A rough output from the composition system would suffice for the subsequent retrieval. In this way, the initial set that is comprised of images similar to the composite could lead to further speedup of the first subject hit.

The automated benchmark study aimed at providing an objective comparison

[1] http://www.bee-biometrics.org/

among various IFIR implementations. However, no retrieval procedure that involves the human factor can by any means be deterministic. Rigorous test settings might in some cases cause an evaluation bias from the true performance of an IFIR system. For example, DSOMs have been proposed for the fuzzy aspects like *hairstyle*, but the current version of PicSOM analysis supports only boolean-valued relevance for a set of enumerated choices. Consequently the IFIR performance by using the *hairstyle* partial relevance could be underestimated. More studies are therefore needed to investigate the actual stochastic user behavior in IFIR systems.

# Bibliography


[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[3] S. Amari and H. Nagaoka. *Methods of information geometry.* Oxford University Press, 2000.

[4] M. Arevalillo-Herráez, J. Domingo, and F. J. Ferri. Combining similarity measures in content-based image retrieval. *Pattern Recognition Letters*, 2008. In press.

[5] M. Bartlett, J. Movellan, T.J., and Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, 13(6):1450–1464, 2002.

[6] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14:585–591, 2002.

[7] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[8] V. Castelli and L. D. Bergman, editors. *Image Databases - Search and Retrieval of Digital Imagery.* John Wiley & Sons, Inc., 2002.

[9] A. Cerri, M. Ferri, and D. Giorgi. Retrieval of trademark images by means of size functions. *Graphical Models*, 68(5-6):451–471, 2006.

[10] C.-H. Chan and I. King. Using biased support vector machine to improve retrieval result in image retrieval with self-organizing map. In *Proceedings of The 11th International Conference on Neural Information Processing*, pages 714–719, Calcutta, India, November 2004.

[11] H. Chang and D.-Y. Yeung. Kernel-based distance metric learning for content-based image retrieval. *Image and Vision Computing*, 25(5):695–703, 2007.

[12] H. Chang, D.-Y. Yeung, and W. K. Cheung. Relaxational metric adaptation and its application to semi-supervised clustering and content-based image retrieval. *Pattern Recognition*, 39(10):1905–1917, 2006.

[13] S.-F. Chang, A. Eleftheriadis, and R. McClintock. Next-generation content representation, creation, and searching for new-media applications in education. *Proceedings of the IEEE*, 86(5):884–904, 1998.

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

[15] D. U. Cho, U. D. Chang, B. H. Kim, S. H. Lee, Y. L. J.Bae, and S. C. Ha. 2D direct LDA algorithm for face recognition. In *Proceedings of the Fourth International Conference on Software Engineering Research, Management and Applications*, pages 245–248, Seattle, Washington, USA, August 2006.

[16] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos. The Bayesian image retrieval system, PicHunter: Theory, implementation and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, 2000.

[17] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[18] A. Csillaghy, H. Hinterberger, and A. O. Benz. Content-based image retrieval in astronomy. *Information Retrieval*, 3(3):229–241, 2000.

[19] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. Technical Report CSE 06-009, Penn State University, 2006.

[20] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Information Theory*, 36(9):961–1005, 1990.

[21] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.

[22] E. d. Ves, J. Domingo, G. Ayala, and P. Zuccarello. A novel bayesian framework for relevance feedback in image content-based retrieval systems. *Pattern Recognition*, 39(9):1622–1632, 2006.

[23] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of International Conference on Machine Learning*, pages 225–232, Banff, Alberta, Canada, July 2004.

[24] J. P. Eakins. Towards intelligent image retrieval. *Pattern Recognition*, 35(1):3–14, 2002.

[25] Y. Fang, D. Geman, and N. Boujemaa. An interactive system for mental face retrieval. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 193–200, Singapore, November 2005.

[26] R. Fidel, T. B. Hahn, E. M. Rasmussen, and P. J. Smith, editors. *Challenges in Indexing Electronic Text and Images*. ASIS Monograph Series. Learned Information, Inc., 1994.

[27] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1963.

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

[29] G. H. Golub and C. F. v. Loan. *Matrix Computations*. The Johns Hopkins University Press, 2nd edition, 1989.

[30] I. Gondra and D. R. Heisterkamp. Content-based image retrieval with the normalized information distance. *Computer Vision and Image Understanding*, 111(2):219–228, 2008.

[31] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice-Hall, 3rd edition, 2008.

[32] A. Grigorova, F. G. B. D. Natale, C. Dagli, and T. S. Huang. Content-based image retrieval by feature adaptation and relevance feedback. *IEEE Transactions on Multimedia*, 9(6):1183–1192, 2007.

[33] J. Han and K.-K. Ma. Rotation-invariant and scale-invariant gabor features for texture image retrieval. *Image and Vision Computing*, 25(9):1474–1481, 2007.

[34] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang. A memory learning framework for effective image retrieval. *IEEE Transactions on Image Processing*, 14(4):511–524, 2005.

[35] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

[36] A. Hatcher. *Algebraic Topology.* Cambridge University Press, 2002.

[37] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, Inc., 2nd edition, 1999.

[38] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang. Manifold-ranking based image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 9–16, New York, NY, USA, October 2004.

[39] J. He, H.-J. Zhang, H. Tong, and C. Zhang. Generalized manifold-ranking-based image retrieval. *IEEE Transactions on Image Processing*, 15(10):3170–3177, 2006.

[40] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 27(3):328–340, 2005.

[41] T. Hertz, A. Bar-Hillel, and D. Weinshall. Learning distance functions for image retrieval. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 570–577, Washington, DC, USA, June 2004.

[42] C.-H. Hoi, C.-H. Chan, K. Huang, M. R. Lyu, and I. King. Biased support vector machine for relevance feedback in image retrieval. In *Proceedings of IEEE International Joint Conference on Neural Networks*, volume 4, pages 3189–3194, Budapest, Hungary, July 2004.

[43] S. C. Hoi, M. R. Lyu, and R. Jin. A unified log-based relevance feedback scheme for image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):509–524, 2006.

[44] P. Hong, Q. Tian, and T. S. Huang. Incorporate Support Vector Machines to content-based image retrieval with relevance feedback. In *Proceedings of IEEE International Conference on Image Processing*, volume 3, pages 750–753, Vancouver, Canada, September 2000.

[45] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.

[46] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):995–1006, 2005.

[47] P. Howland, J. Wang, and H. Park. Solving the small sample size problem in face recognition using generalized discriminant analysis. *Pattern Recognition*, 39(2):277–287, 2006.

[48] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[49] E. Ide. New experiments in relevance feedback. *The Smart System Experiments in Automatic Document Processing*, pages 337–354, 1971.

[50] J. Iivarinen and J. Pakkanen. Content-based retrieval of defect images. In *Proceedings of Advanced Concepts for Intelligent Vision Systems*, pages 62–67, Ghent, Belgium, September 2002.

[51] J.-S. Jang, C.-T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing*. Prentice-Hall, 1996.

[52] F. Jing, M. Li, H.-J. Zhang, and B. Zhang;. Relevance feedback in region-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5):672–681, 2004.

[53] T. Kamei and M. Mizoguchi. Fingerprint preselection using eigenfeatures. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 918–923, Santa Barbara, California, USA, June 1998.

[54] S. Kaski, J. Sinkkonen, and A. Klami. Discriminative clustering. *Neurocomputing*, 69:18–41, 2005.

[55] S. Kaski, J. Sinkkonen, and J. Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

[56] T. Kato. Cognitive view mechanism for content-based multimedia information retrieval. *Interfaces to Database Systems*, pages 244–262, 1992.

[57] Y. Ke and R. Sukthankar. PCA-SIFT: A more dinstinctive representation ofr local image descriptors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513, June 2004.

[58] M. L. Kherfi and D. Ziou. Relevance feedback for CBIR: A new approach based on probabilistic feature weighting with positive and negative examples. *IEEE Transactions on Image Processing*, 15(4):1017–1030, 2006.

[59] D.-H. Kim, C.-W. Chung, and K. Barnard. Relevance feedback using adaptive clustering for image similarity retrieval. *Journal of Systems and Software*, 78(1):9–23, 2005.

[60] P. Koikkalainen. Progress with the tree-structured self-organizing map. In A. G. Cohn, editor, *11th European Conference on Artificial Intelligence*, pages 211–215, Amsterdam, The Netherlands, August 1994. European Committee for Artificial Intelligence, John Wiley & Sons, Ltd.

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

[62] M. Kokare, B. N. Chatterji, and P. K. Biswas. Cosine-modulated wavelet based texture features for content-based image retrieval. *Pattern Recognition Letters*, 25(4):391–398, 2004.

[63] M. Koskela. *Interactive Image Retrieval using Self-Organizing Maps*. PhD thesis, Laboratory of Computer and Information Science, Helsinki University of Technology, November 2003. Available online at: http://lib.hut.fi/Diss/2003/isbn9512267659/.

[64] A. Kushki, P. Androutsos, K. N. Plataniotis, and A. N. Venetsanopoulos. Query feedback for interactive image retrieval. *IEEE Transaction on Circuits and Systems for Video Technology*, 14(5):644–655, 2004.

[65] S. Kwok and J. L. Zhao. Content-based object organization for efficient image retrieval in image databases. *Decision Support Systems*, 42(3):1901–1916, 2006.

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

[67] J. Laaksonen, M. Koskela, and E. Oja. Class distributions on SOM surfaces for feature extraction and object retrieval. *Neural Networks*, 17(8-9):1121–1133, 2004.

[68] T. Landauer, P. W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[69] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[70] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Neural Information Processing Systems*, pages 556–562, Denver, CO, USA, 2000.

[71] M. S. Lew, editor. *Principles of Visual Information Retrieval*. Springer-Verlag, 2001.

[72] C. Liu. Enhanced independent component analysis and its application to content based face image retrieval. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 34(2):1117–1127, 2004.

[73] C. Liu. Gabor-based kernel pca with fractional power polynomial models for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):572–581, 2004.

[74] R. Liu, Y. Wang, T. Baba, D. Masumoto, and S. Nagata. Svm-based active feedback in image retrieval using clustering and unlabeled data. *Pattern Recognition*, 41(8):2645–2655, 2008.

[75] Y. Liu, D. Zhang, and G. Lu. Region-based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition*, 41(8):2554–2570, 2008.

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

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

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

[79] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[80] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, 2004.

[81] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. H. G. D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin. The QBIC project: Querying images by content using color, texture and shape. In *Storage and Retrieval for Image and Video Databases (SPIE)*, volume 1908 of *SPIE Proceedings Series*, pages 173–181, San Jose, CA, USA, February 1993.

[82] Y. Nishimori and S. Akaho. Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. *Neurocomputing*, 67:106–135, 2005.

[83] E. Oja. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273, 1982.

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

[85] M. C. Oliveira, W. Cirne, and P. M. d. Azevedo Marques. Towards applying content-based image retrieval in the clinical routine. *Future Generation Computer Systems*, 23(3):466–474, 2007.

[86] K. Pastra, H. Saggion, and Y. Wilks. Intelligent indexing of crime scene photographs. *IEEE Intelligent Systems*, 18(1):55–61, 2003.

[87] J. Peltonen and S. Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 16(1):68–83, 2005.

[88] J. Peltonen, A. Klami, and S. Kaski. Improved learning of Riemannian metrics for exploratory analysis. *Neural Networks*, 17:1087–1100, 2004.

[89] J. Peltonen, J. Goldberger, and S. Kaski. Fast discriminative component analysis for comparing examples. In *Neural Information Processing Systems Workshop on Learning to Compare Examples*, 2006.

[90] J. Peltonen, A. Klami, and S. Kaski. Learning more accurate metrics for self-organizing maps. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 999–1004, London, UK, 2002.

[91] W. B. Pennebaker and J. L. Mitchell. *JPEG: Still Image Data Compression Standard (Digital Multimedia Standards)*. Springer, 1992.

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

[93] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[94] B. G. Prasad, K. K. Biswas, and S. K. Gupta. Region-based image retrieval using integrated color, shape, and location index. *Computer Vision and Image Understanding*, 94(1-3):193–233, 2004.

[95] X. Qi and Y. Han. A novel fusion approach to content-based image retrieval. *Pattern Recognition*, 38(12):2449–2465, 2005.

[96] G. Qiu. Embedded colour image coding for content-based retrieval. *Journal of Visual Communication and Image Representation*, 15(4):507–521, 2004.

[97] X. Ren and J. Malik. Learning a classification model for segmentation. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 10–17, October 2003.

[98] C. V. v. Rijsbergen. *Information Retrieval*. Butterworth, 2nd edition, 1979.

[99] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART retrieval system: Experiments in automatic document processing*, pages 313–323. Prentice-Hall, 1971.

[100] K. Rodden and K. Wood. How do people manage their digital photographs? In *Proceedings of ACM Conference on Human Factors in Computing Systems (ACM CHI 2003)*, pages 409–416, Fort Lauderdale, FL, USA, April 2003.

[101] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a metric for image retrieval. Technical Report CS-TN-98-86, Stanford University, 1998.

[102] Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in MARS. In *Proceedings of IEEE International Conference on Image Processing*, pages 815–818, Santa Barbara, CA, USA, October 1997.

[103] J. Ruiz-del-Solar and P. Navarrete. FACERET: An interactive face retrieval system based on Self-Organizing Maps. In *Proceedings of Image and Video Retrieval : International Conference, CIVR 2002*, pages 157–164, London, UK, July 2002.

[104] G. Salton. *The SMART retrieval system: Experiments in automatic document processing*, chapter Relevance feedback and the optimization of retrieval effectiveness. Prentice-Hall, 1971.

[105] M. J. Schervish. *Theory of Statistics*. Springer, 1995.

[106] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[107] J. Shotton, A. Blake, and R. Cipolla. Multiscale categorical object recognition using contour fragments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1270–1281, 2008.

[108] C.-R. Shyu, C. Brodley, A. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick. ASSERT: A physician-in-the-loop content-based retrieval system for HRCT image databases. *Computer Vision and Image Understanding*, 75(1-2):111–132, 1999.

[109] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[110] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[111] J. R. Smith and S.-F. Chang. Searching for images and videos on the world-wide web. Technical Report #459-96-25, Columbia University, 1996.

[112] T. R. Smith. A digital library for geographically referenced materials. *IEEE Computer*, 29(5):54–60, 1996.

[113] J. A. Snyman. *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms (Applied Optimization)*. Springer, 2005.

[114] H. Tamura and N. Yokoya. Image database systems: A survey. *Pattern Recognition*, 17(1):29–43, 1984.

[115] D. Tao, X. Li, and S. J. Maybank. Negative samples analysis in relevance feedback. *IEEE Transactions on Knowledge and Data Engineering*, 19(4):568–580, 2007.

[116] D. Tao and X. Tang. Random sampling based SVM for relevance feedback image retrieval. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume II, pages 647–652, Washington, DC, USA, June 2004.

[117] D. Tao, X. Tang, X. Li, and Y. Rui. Direct kernel biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm. *IEEE Transactions on Multimedia*, 8(4):716–727, 2006.

[118] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1088–1099, 2006.

[119] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[120] C.-F. Tsai, K. McGarry, and J. Tait. Using neuro-fuzzy techniques based on a two-stage mapping model for concept-based image database indexing. In *Proceedings of the IEEE Fifth International Symposium on Multimedia Software Engineering*, pages 6–12, Taichung, Taiwan, December 2003.

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

[122] R. C. Veltkamp and M. Tanase. Content-based image retrieval systems: A survey. Technical Report 2000-34, Utrecht University, Information and Computing Sciences, Utrecht, The Netherlands, October 2000. Available at: http://www.cs.uu.nl/research/techreps/UU-CS-2000-34.html.

[123] V. Viitaniemi and J. Laaksonen. Browsing an electronic mail-order catalogue with PicSOM content-based image retrieval system. In *Proceedings of 10th Finnish Artificial Intelligence Conference*, pages 170–181, Oulu, Finland, December 2002.

[124] J. Vogel and B. Schiele. Performance evaluation and optimization for content-based image retrieval. *Pattern Recognition*, 39(5):897–909, 2006.

[125] G. K. Wallace. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, 1992.

[126] L. Wang, K. L. Chan, and Y. P. Tan. Image retrieval with SVM active learning embedding euclidean search. In *Proceedings of International Conference on Image Processing*, volume I, pages 725–728, Barcelona, Spain, September 2003.

[127] L. Wang and B. Manjunath. A semantic representation for image retrieval. In *Proceedings of International Conference on Image Processing*, volume II, pages 523–526, Sept. 2003.

[128] M. Wang, Z.-L. Ye, Y. Wang, and S.-X. Wang. Dominant sets clustering for image retrieval. *Signal Processing*, 88(11):2843–2849, 2008.

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

[130] Y. Wu, Q. Tian, and T. S. Huang. Discriminant-EM algorithm with application to image retrieval. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2000)*, volume 1, pages 222–227, Hilton Head Island, SC, USA, June 2000.

[131] B. Yang. Projection approximation subspace tracking. *IEEE Transactions on Signal Processing*, 43(1):95–107, 1995.

[132] Z. Yang and J. Laaksonen. Approximated geodesic updates with principal natural gradients. In *Proceedings of The 2007 International Joint Conference on Neural Networks*, pages 1320–1325, Orlando, USA, August 2007.

[133] A. Yoshitaka and T. Ichikawa. A survey of content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):81–93, 1999.

[134] Z. Yuan and E. Oja. Projective nonnegative matrix factorization for image compression and feature extraction. In *Proc. of 14th Scandinavian Conference on Image Analysis*, pages 333–342, Joensuu, Finland, June 2005.

[135] P. C. Yuen and C. H. Man. Enhanced human face image searching system using relevance feedback. In *Proceeding of The Sixth International Conference on Automatic Face and Gesture Recognition*, pages 177–182, Seoul, Korea, May 2004.

[136] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[137] L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. In *Proceedings of IEEE International Conference on Image Processing (ICIP 2001)*, volume 2, pages 721–724, Thessaloniki, Greece, October 2001.

[138] T. Zhang, J. Yang, D. Zhao, and X. Ge. Linear local tangent space alignment and application to face recognition. *Neurocomputing*, 70(7-9):1547–1553, 2007.

[139] W. Zhang, S. Shan, X. Chen, and W. Gao. Local gabor binary patterns based on kullback-leibler divergence for partially occluded face recognition. *IEEE Signal Processing Letters*, 14(11):875–878, 2007.

[140] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):318–338, 2005.

[141] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[142] X. S. Zhou and T. S. Huang. Relevance feedback for image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.