TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY
Julkaisu 734 • Publication 734

Zhong Daidi

Image Database Retrieval Methods Based on Feature Histograms

Tampere 2008

Tampereen teknillinen yliopisto. Julkaisu 734
Tampere University of Technology. Publication 734

Daidi Zhong

Image Database Retrieval Methods Based on Feature Histograms

Thesis for the degree of Doctor of Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB222, at Tampere University of Technology, on the 23th of May 2008, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2008

ISBN 978-952-15-1970-3 (printed)
ISBN 978-952-15-1974-1 (PDF)
ISSN 1459-2045

IMAGE DATABASE RETRIEVAL METHODS BASED ON FEATURE HISTOGRAMS

Thesis for the degree of Doctor of Technology

By

Zhong Daidi

June 2007

Abstract

Proliferation of digital capture, storage and networking systems makes creation of image collections very easy. A difficult and unsolved problem is the formation of computerized image databases which would enable precise retrieval using key images and specified search criteria like expressed in task "find me the most similar picture". The task is easy for humans but it is hard to implement in computers due to the complexity of visual information and lack of highly efficient algorithms. In this thesis a class of algorithms for image database retrieval is proposed and investigated. These algorithms are based on two conceptual principles. First principle is that image information used for retrieval should be effectively reduced in order to preserve key information and prevent the growth in computational complexity. Second principle is proper combination of image statistical and structural information in which the latter one is used as little as possible since the former one is much easier to describe and compute.

Several methods are developed in the thesis to fulfill those principles, acting on three levels of processing hierarchy from local to global. At a bottom level, local features are constructed from the coefficients of quantized block transforms. Block transforms are widely used in image and video compression and are well known for their excellent ability of preserving perceptual information under heavy quantization. Quantization acts for the concentration of block-wise information in a more condense way, which is highly desirable for the retrieval tasks. In the thesis several new types of local features are introduced and their properties are described. At an intermediate level, histograms of local image features are used as descriptors of global statistical information. Histogram similarity measure is introduced and methods for combining feature histograms are investigated. Finally, at the top level, in the thesis the combination of histograms from image sub-areas is defined as a way to incorporate structural information. The three information processing levels are composed into an overall image database retrieval system. The system parameters, like quantization level, histogram length and image subareas, are optimized iteratively using training datasets. The optimized system performance is evaluated on the example of available face databases using standardized evaluation procedures. The results show that the performance approaches best other methods proposed and sometimes exceeds them. This indicates that proposed methods for the description and combination of statistical and structural information are very effective for the image database retrieval.

Preface

The work presented in this thesis has been carried out at the Institute of Signal Processing, Tampere University of Technology, Finland. Part of the work is financially supported by the Centre for International Mobility (CIMO) Scholarship, Nokia Foundation Award and the Tampere Graduate School in Information Science and Engineering (TISE).

In the first place, I am deeply grateful to my supervisor Irek Defée for his guidance, encouragement and kindly support throughout my studies. I would like to thank the thesis reviewers Professor Spiros Fotopoulos and Professor Dietrich Paulus for their feedback and constructive comments on the manuscripts, as well as towards Dr. Adrian Bors for his effort during my defense.

I would like to thank all the colleagues at the Institute of Signal Processing and the Nokia Research Center for the pleasant working atmosphere. Special thank goes to all those who made our institute scientifically as well as culturally very rich.

I would also like to expend my thanks to my friends from the local Chinese community. They helped me a lot with my daily life and academic career. I have enjoyed a wonderful time together with them in this small city.

Portions of the research in this thesis uses the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office. I would like to thank National Institute of Standards and Technology (NIST) for providing the FERET data. For the similar reason, I also wish to thank the AT&T Laboratories Cambridge, who is the provider of ORL database. Furthermore, the financial support to me by Tampere Graduate School in Information Science and Engineering (TISE) scholarship, Nokia scholarship and Chinese Government Outstanding Oversea Student scholarship are also gratefully acknowledged.

Finally, I am indebted to my wife, my parents and my baby. Their encouragement, support and endless love are the ultimate power resource which allows me to make this possible. For this reason, I would like to dedicate this thesis for them.

List of Publications

I. DaiDi Zhong, Irek Defée, "Pattern Recognition in Compressed DCT Domain", in Proceedings of IEEE International Conference on Image Processing (ICIP 2004), pp. 2031-2034, October 2004.

II. DaiDi Zhong, Irek Defée, "Global Pattern Selection For Compression Histogram Database Retrieval", in Proceedings of International Workshop on Systems, Signals and Image Processing (IWSSIP 2004), pp. 239-242, September 2004.

III. DaiDi Zhong, Irek Defée, "DCT Histogram Optimization for Image Database Retrieval", Pattern Recognition Letter, Vol. 26, Iss. 14, pp. 2272-2281, 2005

IV. DaiDi Zhong, Irek Defée, "Pattern Retrieval Using Optimized Compression Transform", in Proceedings of SPIE Visual Communications and Image Processing (VCIP 2005), pp. 1571-1578, July 2005.

V. DaiDi Zhong, Irek Defée, "Study of image retrieval based on feature vectors in compressed domain", Proceedings of 7th Nordic Signal Processing Symposium (NORSIG 2006), pp. 202-205, June 2006

VI. DaiDi Zhong, Irek Defée, "Performance of Similarity Measures Based on Histograms of Local Image Feature Vectors", Pattern Recognition Letter, Volume 28, Issue 15, pp. 2003-2010, 2007

VII. DaiDi Zhong, Irek Defée, "Face Recognition In Compressed Domain Using Ternary Feature Vector", in Proceedings of European Signal Processing Conference (EUSIPCO 2007), pp. 1580-1584, Sep. 2007.

VIII. DaiDi Zhong, Irek Defée, "Face Image Retrieval System Using TFV and Combination of Subimages", in Proceedings of International Conference series on Visual Information Systems (VISUAL 2007), June 2007.

IX. DaiDi Zhong, Irek Defée, "A Framework for Combining Statistical and Structural Pattern Retrieval Based on Feature Histograms", in Proceedings of IEEE International Workshop on Machine Learning For Signal Processing (MLSP 2007), August 2007.

X. DaiDi Zhong, Irek Defée, "Fast Searching For The Optimal Area Of TFV Representation", in Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP 2007), October 2007.

XI. DaiDi Zhong, Irek Defée, "Location Detection of Face Features by DCT Coefficients", in Proceedings of Visualization, Imaging, and Image Processing 2005 (VIIP 2005), pp. 99-103, September 2005.

XII. DaiDi Zhong, Irek Defée, "Pattern Recognition by Grouping Areas in DCT Compressed Images", in Proceedings of 6th Nordic Signal Processing Symposium (NORSIG 2004), pp. 312-315, June 2004.

Supplemental Publications:

XIII. DaiDi Zhong, Irek Defée, "A Three-Layer System for Image Retrieval", in Proceedings of International Conference on Signal Processing and Multimedia Applications (SIGMAP 2007), pp. 208-212, July 2007.

XIV. DaiDi Zhong, Irek Defée, "Facial Features Detection by Coefficient Distribution Map", in Proceedings of The 11th International Conference on Computer Analysis of Images and Patterns (CAIP 2005), pp. 822-828, September 2005.

XV. DaiDi Zhong, Irek Defée, "Face Retrieval Based on Robust Local Features and Statistical-Structural Learning Approach", EURASIP Journal on Advances in Signal Processing Volume 2008, Article ID 631297, 12 pages, doi:10.1155/2008/631297, 2008.

List of Figures

List of Tables

List of Abbreviations

ACBP	AC Block Pattern
ADM	ACBP Distribution Map
ANN	Artificial Neural Networks
AR	Accuracy Rate
BDM	Block Density Matching
BFV	Binary Feature Vector
CDM	Coefficient Distribution Map
CMS	Cumulative Match Score
Cor	Correlation distance
Cos	Cosine distance
DCN	DC Normalization
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
FAR	False Acceptance Rate
EER	Equal Error Rate
EWC	Eigenvalue-Weighted Cosine
FERET	The Facial Recognition Technology Database
FID	Full Image Decomposition
FRR	False Rejection Rate
FV	Feature Vector
HE	Histogram Equalization
ICA	Independent Component Analysis
ISOMAP	Isometric Mapping
KPCA	Kernel Principal components analysis
LBP	Local Binary Pattern
LLE	Locally Linear Embedding
NIST	National Institute of Standard and Technology
ORL	Olivetti Research Laboratory Database
PCA	Principle Component Analysis
PID	Partial Image Decomposition
QBIC	Query by Image Content
QP	Quantization Parameter
SE	Standard Euclidean distance
TFV	Ternary Feature Vector

Table of Contents

Chapter 1

Introduction

Digital image capture and processing revolutionized in a short period of time the way images are handled. Instead of keeping film and album collections, images are nowadays stored in digital format, transmitted over networks, and manipulated by sophisticated software. The ease of producing and distribution leads to the huge increase of volume of digital images available from different sources. This creates a problem of how to manage this information, which is normally solved by organizing it into databases. Database is a combination of storage and processing system which facilitates handling of data by structuring them in a way which allows querying by the data content as required by specific applications. Databases are used at present universally for the information described in text format. Text is a kind of description based on relatively simple primitives and rules, which makes it feasible to produce algorithms for managing text databases. An ultimate illustration of this is the web search engines like Google which operate globally over the Internet. Despite of dealing with gigantic amounts of text data, these search engines can provide almost immediate answers to sophisticated queries. They accomplish this by organizing the data into huge databases with very efficient query engines.

Managing of images is a much more complicated problem than text. Image information has no simple primitives and rules for describing the data. In consequence the problem of creating databases for images with sophisticating query possibilities is not solved yet in general. The difficulty of this problem is easily underestimated due to the fact that in biological information processing systems, and especially in the human visual processing, images are stored, compared and queried easily and apparently with no effort as it is well known from daily experience. However this is only a deception achieved by perfectly hiding complexity of underlying information processing. Intensive research over many years on image databases has not produced yet systems which could be considered close to the performance of their biological counterparts. This concerns especially the general domain-independent systems dealing with unrestricted variety of images. On the other hand, for particular classes of images and for specific applications, there exist database retrieval systems producing results at performance level which is not perfect but sufficient for practical needs, e.g. in the fingerprint identification.

Due to the fact that the general image database retrieval problem is still not solved, there is continuing intensive research with many new ideas and solutions being proposed. This line of research is also followed in this thesis, in which several new detailed methods and an overall novel approach are proposed and compared to existing results using standardized methodology.

1.1 Statement of Research Problem

This thesis is formulated within a framework of general image database retrieval problem which can be stated as follows. Given a set of images, how to organize them into a database which would enable querying by image? Querying by image means finding solution to the questions like: Is an image used as a query contained in the database or not? Which are the images in the database most similar to the query image? To provide answers to such questions there have to be developed algorithms which will process the images and produce correct responses. The set of images with the processing algorithms will form image retrieval database system. There are several fundamental difficulties with the development of image retrieval algorithms. They originate from the complexity of visual information which is not easy to describe. The complexity can be seen as originating from three reasons. The first reason is the rich set of primitives - basic local image features composed of pixels. The second reason is complex structural information which depends on locations of features in images. The third reason is the statistical information when details about feature locations are not important but their statistics is. These three aspects of complexity are often coexisting which makes the description of images very hard. In contrast, extremely capable image retrieval is successfully implemented in biological visual information processing systems. The computational principles of how biological systems achieve their capabilities are not known yet, but based on current knowledge, they are realized by combining specialized hardware (e.g. for handling face images) and/or by long training (e.g. recognizing Chinese script). This illustrates that dealing with the complexity of visual information requires approach which should include some domain-dependent aspects achievable by training and adaptation. It also indicates that efficient image data processing including learning phase has to include data reduction to eliminate processing overheads. The question is how to do it. Therefore, above considerations drive the research focus of this thesis to the design of efficient image database retrieval system.

1.2 Objectives and Scope of the Thesis

Main objective of this thesis is to develop a class of image retrieval database algorithms and a processing system within the framework of information processing efficiency, as well as to evaluate its performance. The following technical objectives serve to fulfill the main objective; they are presented with descriptions of the scope of research:

To develop efficient local feature description. The main idea is to use approach based on block transforms which are widely applied in the lossy image compression area. In lossy compression, images are divided into small blocks which are subsequently transformed by a transform. Currently, the most popular one of such kind of transform is the DCT (Discrete Cosine Transform) [1] operating on 8x8 image blocks. After the block transform its coefficients are quantized. This leads to the reduction of information and the reason for using the DCT is its excellent preservation of perceptually important local information even under heavy quantization. Quantization serves for the reduction of information, eliminating the part of it which is perceptually not relevant. This is also very desirable from the view of image retrieval system since it should only deal with perceptual information. To fully utilize this property, in this thesis the 4x4 AC transform, known from the H.264 video compression standard, is used. Although the 4x4 quantized transforms can be used directly as local feature descriptors, in the thesis the so called Feature Vectors (FV) are further developed based on the quantized transform coefficients. The FV is describing local features more effectively as it is shown in thesis.

To develop efficient description of statistical information. Statistical information describes global distribution of features, and in general, does not depend on their locations. In this thesis, the histograms of quantized block transform and feature vectors are developed to describe the statistical information. Histograms are widely used for representing global statistical image information using different attributes. In the thesis histograms of feature vectors are used with suitable norm to measure similarity of images. Parameterized combinations of histograms are introduced which allow to include several types of feature vectors with controlled impact. Feature histograms obtained in this way are simple and efficient tools for the description and comparison of statistical information of images. Size of the feature histograms, measured by the number of bins corresponding to feature vectors, reflects the amount of statistical information used for the description. The size is

a parameter which reflects the amount of statistical information but not all this information is useful and some of it may even be harmful for the retrieval performance. From this reason the size of feature histograms is a parameter which is adjusted during the training to optimize the statistical information used for best performance.

To develop a framework for the incorporation of structural information. Structural information is the one which takes into account locations of local features within images. This information is in general very complex and difficult to handle since it may require dealing with precise locations of thousands of features represented, for example, by the feature vectors. The question is also how to relate the statistical and structural information in order to get clear connection between them. In the thesis this question is resolved by introducing the concept of feature histograms for subareas covering images. The image area is decomposed into subareas and feature histograms are calculated for it. Combined feature histograms of subareas represent approximate description of structural information. This description can be made more precise by making subareas smaller, in the extreme case the subareas are covering single features and this represents full structural information. In the opposite case, full image is just a single subarea and its feature histogram carries no structural information. Combining feature histograms from subareas makes flexible framework for describing structural information linking it naturally with statistical information. In the thesis this framework is used for the reduction of structural information needed to accomplish a retrieval task. Reduction of structural information is equivalent to finding decomposition with lowest number of subareas whose combined histograms can be used to achieve specific retrieval performance.

To develop of an adaptive system with training for optimal retrieval parameters and performance evaluation. In this thesis, a part of the image database retrieval problem is training towards specific domain and task, due to the fact that we do not have a priori knowledge about what type of images will be considered and what type of task has to be performed. The training step is very critical because of the information processing efficiency objective. To achieve this, a retrieval system is defined and its parameters optimization problem is formulated. Iterative procedure for finding parameter values is derived for operating with the training datasets. The parameters concern different system levels. They include block quantization value and thresholds for local features, length of feature histograms, coefficients for the combination of histogram, and the size, location and number of subareas. The optimization criterion is based on the maximization of Cumulative Match Score (CMS).

This selection is related to the domain used for training and performance evaluation. While the methods presented in the thesis are general and not limited to specific domain, face databases are used for the training and evaluation. Such choice is motivated by the existence of standardized databases and evaluation procedures. Moreover, face retrieval has been widely researched which makes possible to compare results obtained with those from other methods.

In the thesis it is described how the objectives are achieved and how the overall system design accomplished the main objective. Performance evaluation allows drawing conclusions about the validity of the approach relative to the results achieved by others.

1.3 Organization of the Thesis

Information within images is composed of data which can be processed hierarchically from local to global level. The thesis is organized following such order of processing in summarizing the content of Publications I-XIV on which it is based.

Chapter 2 deals with methods for the extraction of relevant local information based on block transforms and quantization. Block transform and quantization are briefly described. Several new feature vectors are proposed to express the local information in an efficient and robust manner suitable for usage in the later parts of the thesis.

In Chapter 3 histograms of feature vectors are introduced. It is shown how these histograms describe statistical information of images and how they can be used to measure image similarities. Next, the combination of histograms of different feature vectors is described. Histograms can also be defined for image subareas, it is shown that this introduces structural information and enables its flexible combination with statistical information.

In Chapter 4 the image database retrieval framework is described. Feature histograms are used within this framework by performing optimization of system parameters using training datasets. Standard performance evaluation protocols are explained.

In Chapter 5 performance of the proposed approach is evaluated and compared with other methods using face image database retrieval. First, the face databases used in performance evaluations are described and the experimental setup is presented. Then,

the results obtained for different retrieval system configurations are listed. They are compared with the results obtained by other researchers. The discussion of results is made in the last part of this chapter.

In the Conclusion the contents of the thesis are overviewed and final conclusions are presented.

Chapter 2

Local Features Based on Block Transforms

Image database retrieval is a complex multi-dimensional problem which can be seen easily when considering images as matrices in an $N \times N$ dimensional space. It is almost impossible to use such matrices directly because they are too large. However, meaningful images form only a subset of such matrices and this fact enables development of certain techniques matched to the image information. One approach for achieving this is by assuming that images can be described as composed of some set of basic building characters called local features. From the point of image description, the challenge is to produce appropriate set of features so that they can be used to build representation of specific image attributes, with good discriminative properties for image retrieval. In this Chapter new methods of flexible generation of local features are presented.

2.1 Local Image Features

An image can be seen as a combination of a set of small image patches, each of them representing a certain amount of local visual information. Description of this kind of local information is called local feature. The requirement of a good feature set is that it is maximally informative, that means the descriptions are concise, the feature set is compact, but at the same time it is sufficiently large to describe meaningful variations between local image patches. The proper derivation and collection of local features is essential for the description of images.

Edges are widely used examples of local features [2-4]. Edges are defined as sharp light intensity changes. However, there can be very many types of edges depending on the profile and directions of light intensity change. Precise detection and classification of edges is rather difficult to formalize and describe [5-7].

Local features proposed in this thesis are more comprehensive than edges; they are computationally efficient and adaptable to variations of images. They are proposed, utilized and tested, and they have been proven to be efficient in describing and

discriminating of large class of images. The features are constructed via block transforms which are explained below.

2.2 Block Transforms

The transform techniques are used to convert image pixels into another type of data, which may represent visual information more efficiently for specific processing tasks. Many different types of transforms were proposed for application in the image retrieval and recognition, e.g. the Discrete Wavelet Transform (DWT) [8-13], Gabor transform [14-18], and the Discrete Cosine Transform (DCT) [19-28].

For the development of local features in this thesis we use ideas originating from the image compression area. The image and video compression standards such as JPEG and MPEG are based on block transforms operating on small image patches [29,30]. Specific block transforms are used jointly with quantization because they can robustly eliminate perceptually irrelevant information and preserve only the part which is important for perceptual quality. Example is given by the widely used Discrete Cosine Transform (DCT) which has excellent preservation of perceptual information even under heavy quantization. The two-dimensional N x N DCT transform matrix $\boldsymbol{C}$ is described by the Eq. (1), where $0 \leq k, l \leq N-1$.

$$c_{k,\,l} = \begin{cases} = \dfrac{1}{\sqrt{N}} & l = 0 \\ = \sqrt{\dfrac{2}{N}} \cos\left[\dfrac{(2k+1)l\pi}{2N}\right] & otherwise \end{cases} \tag{1}$$

In matrix notation, the N x N pixel block $\boldsymbol{P}$ is forward transformed to block $\boldsymbol{H}$ using Eq. (2), and block $\boldsymbol{R}$ is subsequently reconstructed from $\boldsymbol{H}$ using Eq. (3)

$$H = C \times P \times C^T \tag{2}$$

$$R = C^T \times H \times C \tag{3}$$

where X denotes the matrix multiplication and T denotes transpose.

The DCT transform can be seen as a kind of frequency decomposition and the result of the transformation $\boldsymbol{H}$ are coefficients of cosines with specific frequencies. The typical size of DCT used in the compression is $N = 8$. Element $H(0,0)$ is called the DC coefficient as it corresponds to zero frequency and represents the average value

of the pixel block. Other coefficients of $\boldsymbol{H}$ are called AC coefficients. Typical distribution of values of the DCT coefficients for a block taken from natural image is shown in Figure 1.

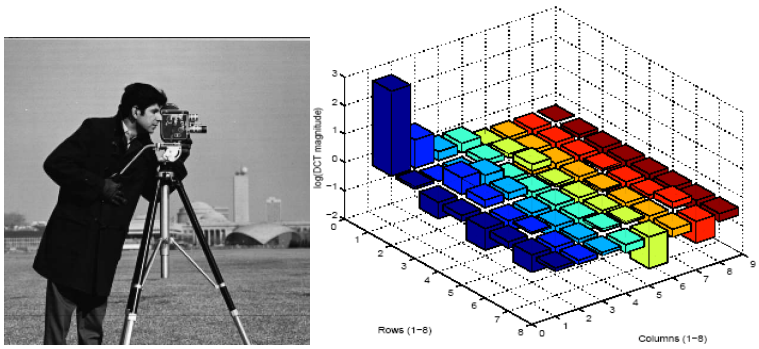


Figure 1: Distribution of the DCT coefficients for typical 8x8 DCT block of natural image. 8x8 DCT transform is taken over the left-hand image, and all the transformed blocks are averaged to produce the right-hand plot.

In Figure 1 one can see example of typical DCT coefficient distribution: higher frequency coefficients (close to the right corner) have generally smaller values than lower frequency coefficients (close to the left corner). It is also known that from the perceptual point of view the higher frequencies are not critical from perceptual point and can be largely reduced or removed by the quantization operation described later. In this respect, the DCT is very robust in the preservation of perceptual information.

In the image and video compression standards traditionally the DCT of size 8x8 has been used. This limits the precision of perceptual details which can be processed and from this reason in the newer H.264 video compression standard a 4x4 transform derived from the DCT is introduced [31]. This transform approximates the DCT well but it is also optimized to be computationally very simple by having only integer coefficients. The forward 4x4 transform matrix is defined as

$$B_f = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \tag{4}$$

while its inverse transform matrix is defined as

$$B_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix} \quad (5)$$

A 4x4 pixel block of image pixels $\boldsymbol{P}$ is forward transformed to block $\boldsymbol{H}$ using Eq. (6), and block $\boldsymbol{R}$ is subsequently reconstructed from $\boldsymbol{H}$ using Eq. (7).

$$H = B_f \times P \times B_f^T \quad (6)$$

$$R = B_i^T \times H \times B_i \quad (7)$$

In Figure 2 we show the results after applying the 4x4 DCT transform and the H.264 4x4 transform (6) to a pixel block are shown. As can be seen the 4x4 transform coefficients are integers while the DCT coefficients require floating-point representation.

164	161	159	162
162	162	163	160
162	162	164	160
160	161	160	161

(a)

645.75	1.63	-0.25	0.67
1.17	1.5303	2.09	-0.07
-1.75	-0.16	3.25	-2.36
0.86	0.42	0.48	0.47

(b)

2583	10	-1	5
7	15	13	0
-7	0	13	-15
6	5	4	5

(c)

Figure 2: (a) A sample 4x4 pixel block, (b) Result of 4X4 DCT transform, (c) Result of 4X4 H.264 AC transform.

As illustrated in Figure 3, given a pixel block transformed by the 4x4 transform:

- The coefficients in the first row correspond to local light intensity variations in vertical direction
- The coefficients in first column correspond to local light intensity variations in horizontal direction
- The coefficients along left diagonal correspond to local light intensity variations along right diagonal

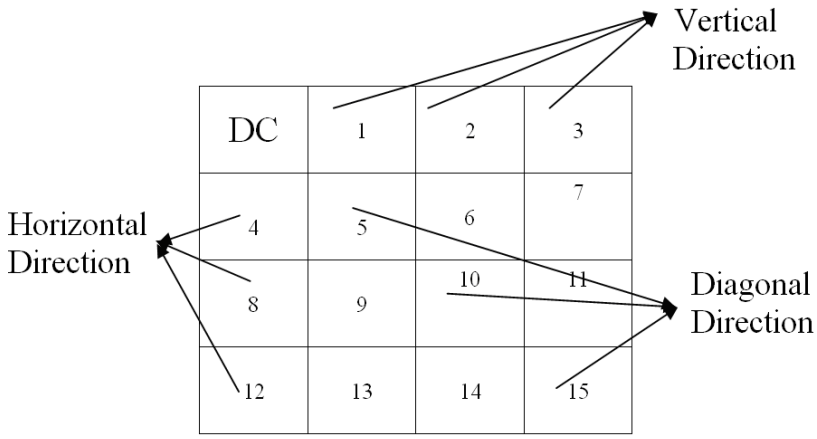


Figure 3: Directional information represented by different coefficient.

The transform coefficients represent thus description of oriented light intensity variations within the 4x4 pixel block.

2.3 Quantization and Local Features

Quantization is the process of reducing the range of variation of the set of data. The data range reduction can be used for the approximation of data and together with concise data description, this process is called lossy data compression. Quantization can be modeled by a mapping a vector of N original data samples into another vector of M quantized samples, with $0 < M < N$. There are two types of quantization:

Scalar quantization: single data samples are mapped by a function to quantized data
Vector quantization: vector of data samples is mapped into a reduced vector of samples

Vector quantization is much more general than scalar quantization, but in turn scalar quantization is very simple. In the image and video compression area, quantization is routinely employed to remove perceptually irrelevant visual information. When applied to the result of block transforms, the quantization is especially effective [30]. The results of the 8x8 DCT block transform in such standards like JPEG and MPEG-2 are quantized by vector quantization using special so called quantization matrices which have been derived to match precisely perceptual sensitivities of human visual system. For the 4x4 transform used in the H.264 standard, much simpler scalar quantization is used in which the transform

coefficients $\boldsymbol{H}$ in Eq. (6) are divided by a number called Quantization Parameter (QP) and rounded to the nearest integer. The bigger the QP value is, the more approximate the quantized transform values will be, and in particular more values will be zero. As mentioned above the block transforms are very special in the sense they approximate perceptually very well the original pixel block after the inverse transforms in Eq. (3) and (7), even if the transform results are strongly quantized with high QP values. The process of inverse quantization and inverse transform can be formulated as in Eq. (8), where $\boldsymbol{Q}[\,]$ stands for the 4x4 scalar quantization by QP and $\boldsymbol{Q^{-1}}[\,]$ means 4x4 inverse scalar quantization, i.e. multiplication of quantized values by QP.

$$R = B_i^T \times Q^{-1}\left[Q(H)\right] \times B_i \tag{8}$$

With suitable selection of QP many of the transform values will be zero and perceptually relevant information of the original pixel block is represented just by a few transform values which are small integer numbers.

For instance, if the blocks (b) and (c) shown in Figure 2 are quantized with QP = 8 and QP = 2 respectively, the result will be as (a) and (b) in Figure 4 respectively. The corresponding reconstructed blocks are shown as (c) and (d) in Figure 4 and it can be seen that they approximate quite well the original block shown in Figure 2 (a).

323	1	0	1
1	2	2	0
-1	0	2	-2
1	1	1	1

(a)

323	1	0	0
1	1	1	0
-1	0	2	-1
0	0	0	0

(b)

166	161	158	162
162	162	164	160
163	162	165	160
159	161	160	161

(c)

165	161	159	162
163	163	163	160
161	162	164	160
160	161	159	161

(d)

Figure 4: (a) Result of quantization over Figure 2-b. (b) Result of quantization over Figure 2-c. (c) Reconstructed result from Figure 2-b. (d) Reconstructed result from Figure 2-c.

The (strongly) quantized result of the 4x4 block transform operating on 4x4 pixel blocks thus represents a very efficient description of perceptually relevant variations of pixel values. In Publication I application of this block transform to the description of local image features is described. This quantized blocks description has many advantages:

- it is very concise, using only few integers
- preserves perceptually important features well
- can be easily adapted to specific task by changing the QP
- it is computationally very effective

It is important to emphasize that while the description of local features by quantized block transform is comprehensive, the number of different features can be controlled. By increasing the QP, the description becomes more approximate and less detailed. In result the overall number of different local features is reduced. This point is illustrated in Figure 5 where this relation is shown for typical image: when the QP increases, the number of local features decreases.

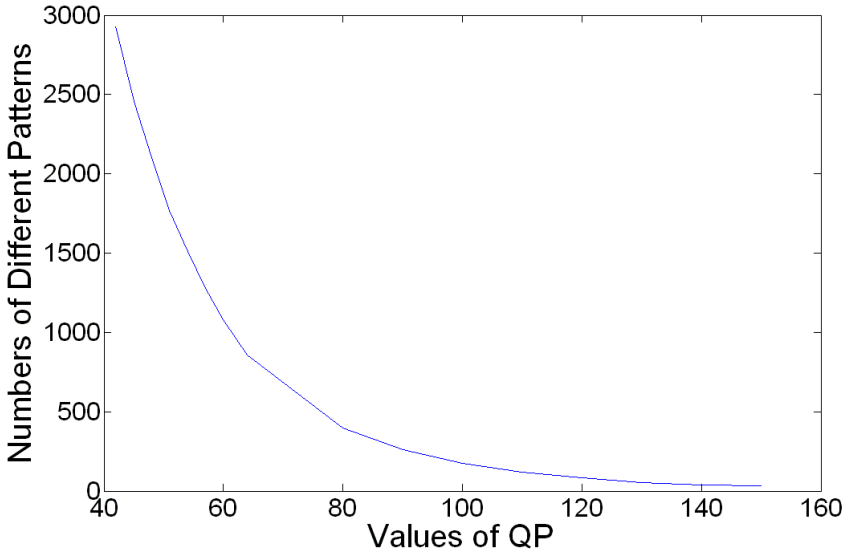


Figure 5: Quantization reduces the number of different blocks when only AC coefficients are considered. This figure is obtained by applying different QP values to the blocks of the same image.

In Figure 6 the effect of increasing the QP on the preservation of perceptual information is shown. It can be noticed that while images are distorted when using high QP, they are still recognizable and may be useful for retrieval purposes.

This phenomenon is helpful from the perspective of finding effective description of image information. By the quantization of block transform, perceptually irrelevant local information can be effectively removed, limited set of local features is left and they are effectively described by the transform coefficients. The question of how the proper value of QP can be selected is studied later in the thesis.

Figure 6: An example image reconstructed from applying quantized block transforms with different QP. From left to right, the QP is increased. When QP is not too large, one may still be able to identify the face, although many details have been lost.

As marked in Figure 3, AC coefficients at different positions in the block represent specific directional information. This is illustrated in Figure 7 for the human face images and their corresponding distributions of fifteen AC quantized transform coefficients whose values are thresholded and binarized. As can be seen the coefficients indicate different aspects of local image patches. They thus can be further used for deriving descriptions of local image features.

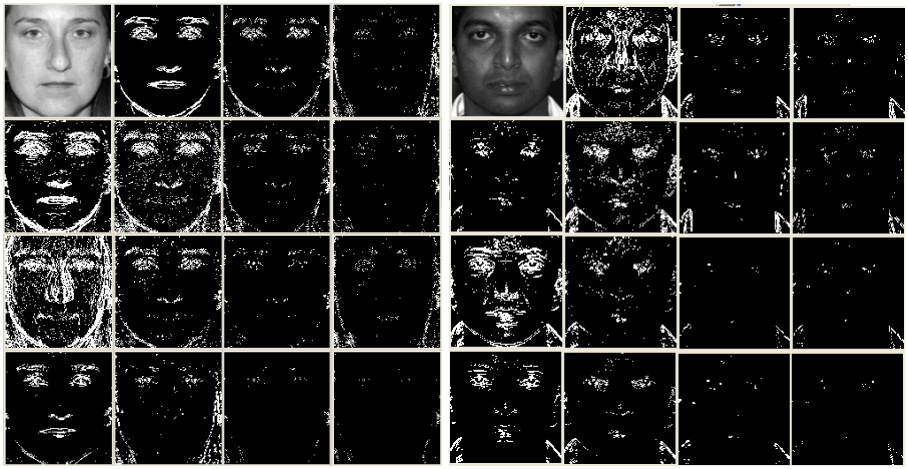


Figure 7: Coefficient Distribution Map (QP=100).

The Coefficient Distribution Map (CDM) was introduced and used in Publication XI for the description of facial image features. While the single AC coefficients of the transform can be used for the extraction of specific information of image features, more advanced local feature descriptors can be constructed from the coefficients. Such descriptors based on quantized block transform coefficients will be presented next.

2.4 Feature Vectors From Transform Coefficients

According to the terminology introduced earlier, the pixel block transformed by the 4x4 transform, Eq. (2) or (6), contains one DC coefficient and fifteen AC coefficients. The DC coefficients carry information about average pixel intensity within the 4x4 blocks. If light intensity is slowly changing over larger image area, single block transform values can not reflect this, which may result in a loss of important perceptual information. From this reason, in Publications III and IV feature vectors constructed separately from the DC and AC coefficients of quantized transform blocks were introduced. The feature vectors are described next starting from the simplest ones which are obtained from reordering the block transform coefficients.

2.4.1 AC Block Patterns

The quantized transform block without the DC coefficient is called the **AC Block Pattern (ACBP)**. The ACBP was introduced in Publications I, II and III, where its performance was investigated. For the purpose of application the ACBP is represented by a vector of block values concatenated in a row-by-row manner as illustrated in Figure 8. The zero values at the end of the vector which result from quantization are skipped which can reduce the size of vector.

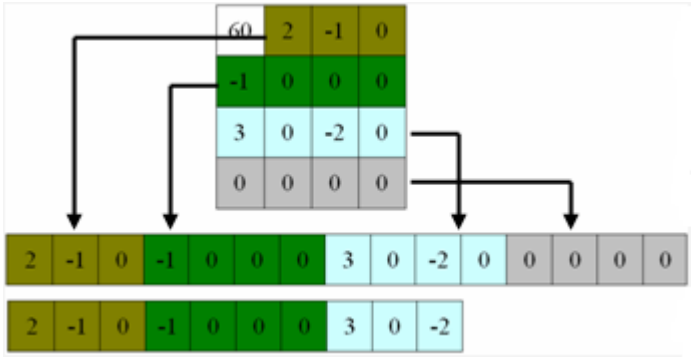


Figure 8: Forming an ACBP pattern from a 4x4 block.

The distribution of ACBP patterns in images depend on their content. As an example, for the face image in Figure 9, certain ACBP vectors are marked as white dots and it can be seen that their location distribution is characteristic for certain facial features. This kind of location distribution was investigated in Publication XII

where it was called ACBP Distribution Map (ADM) and its application for the detection of facial features was described.

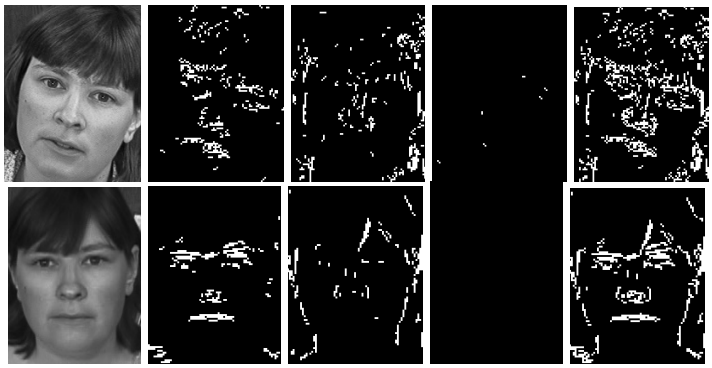


Figure 9: Certain ACBP patterns depict the key facial features. (a) Original image. (b) Location distribution of 4^{th} coefficient (in Figure 3). (c) Location distribution of 1^{st} coefficient (in Figure 3). (d) Location distribution of 5^{th} coefficient (in Figure 3). (e) Combination of (b)-(d).

2.4.2 DC Direction Vectors

The concept of the **DC Direction-Vector (DCDV)** was introduced in Publications III and IV. The DCDV operates on the DC transform values taken from transform blocks of an image and arranged in a matrix.

The process of forming direction vectors is illustrated in Figure 10. For a given DC value, the differences between its value and the eight neighboring DC values are calculated and arranged in a 3x3 matrix. These differences are called direction-values and represents orientation information. The center value of the matrix is calculated as the difference between the current DC value and the mean of the all the nine DC values of the matrix. Next the differences are ordered according to their absolute values, and the corresponding index of the direction is listed to be an 8-bin vector. Subsequently, the first γ direction-values ($1 \leq \gamma \leq 9$) with largest differences form the DCDV; γ is a parameter which can be adjusted to optimize the amount of orientation information carried by the vector. In this case the first $\gamma = 4$ bins are taken to form the DCDV. There is no specific reason why these nine directions are assigned values in such way. Since the DCDV only concentrates on relative differences between different directions, therefore, another way to assign values to different directions doesn't affect the final retrieval performance.

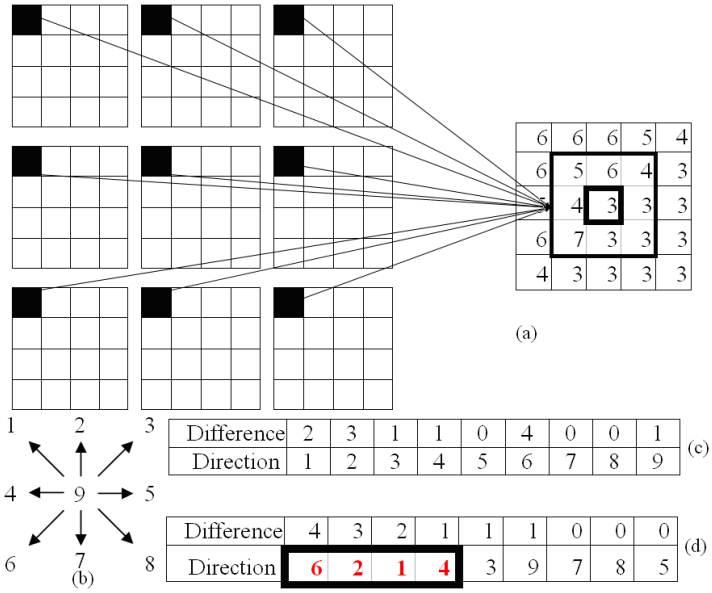


Figure 10: Forming of a Direction-Vector. (a) DC coefficients are extracted from neighboring blocks. (b) Indexes of directions. (c) Difference for each direction. (d) DCDV for γ=4. The indexes of the first form directions with largest differences form the DCDV.

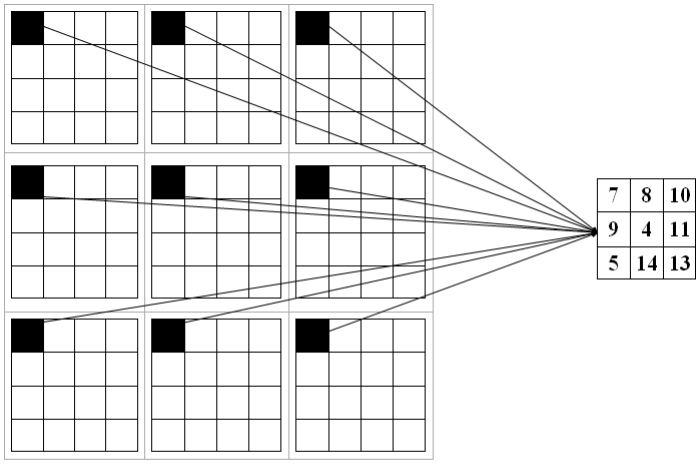
2.4.3 Binary Feature Vectors

The **DC and AC Binary Feature Vectors (DC-BFV and AC-BFV)** were introduced in Publication V. The reason behind their introduction was that it might be worth to combine the DC and AC coefficients from neighboring transform blocks in a similar way like in the DCDV above but in order to avoid increasing complexity their values are binarized by thresholding, producing Binary Feature Vectors (BFV). A BFV is generated from a 3x3 matrix containing the same AC or DC coefficients from nine neighboring transform blocks. The BFV generated from DC coefficients is called the DC-BFV. The BFV generated from AC coefficients is called AC-BFV. Taking the DC-BFV as an example, it is constructed by taking the 3x3 matrix containing nine neighboring DC coefficients. The eight DC coefficients surrounding the center one are thresholded to form a binary vector with length eight:

$$\begin{aligned} &\text{If the DC value} < T \text{ then put } 0 \\ &\text{If the DC value} \geq T \text{ then put } 1 \end{aligned} \qquad (9)$$

where T is the threshold. The value of the threshold T can be defined in many ways, e.g. as the value of the central coefficient in the 3x3 matrix or as the mean value of all its nine coefficients.

The thresholding approach to forming feature vectors is actually the essence of the Linear Binary Pattern (LBP) method [32]. The method operates directly on the image pixels and uses the central coefficient value which may make it sensitive to noise. In our case the thresholding operates on quantized transform coefficients. The process of forming the DC-BFV is illustrated in Figure 11. Eight surrounding coefficients are thresholded by the mean of the nine neighboring coefficients. The result of thresholding is an 8-bin vector which can be further converted to a decimal value. The local feature information is thus minimized to be a single decimal value. In the same way the AC-BFV vectors can be formed for each of the AC coefficients. We have found that for the retrieval framework where the quantized transform coefficients are used, the threshold value given by the mean of the coefficients is better than the center value.



$$Threshold = (7+8+10+9+4+11+5+14+13)/9 = 9$$
$$Thresholding([7\ \ 8\ \ 10\ \ 11\ \ 13\ \ 14\ \ 5\ \ 9]) = [0\ \ 0\ \ 1\ \ 1\ \ 1\ \ 1\ \ 0\ \ 1]$$
$$BFV = 2^5 + 2^4 + 2^3 + 3^2 + 3^0 = 61$$

Figure 11: Illustration of forming the DC-BFV.

2.4.4 Ternary Feature Vector

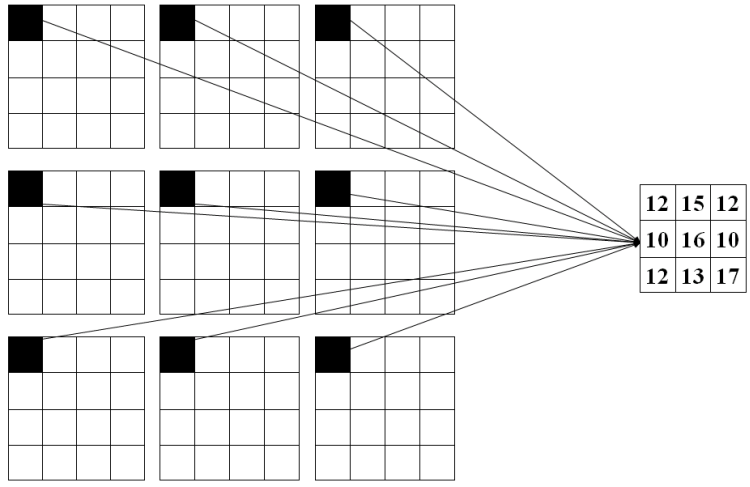
DC and AC Ternary Feature Vector (TFV) was proposed in Publication V as an extension of the BFV by using two thresholds. This is done in the following way:

$$T_{-}^{+} = M \pm (X - N) \times f \tag{10}$$

where T^+ and T_- are two threshold values ($T^+ > T_-$), f is a real number from the interval (0,0.5), X and N are maximum and minimum values in the coefficient matrix, and M is the mean value of the coefficients. The parameter f actually defines a flexible range around the mean value M, which allows all the TFV extracted from image to (approximately) equally distribute into three intervals separated by the two thresholds. The thresholded values are either 0, 1 or 2.

$$\begin{aligned} &\text{If the coefficient value} \le T_- \quad \text{put } 0 \\ &\text{If the coefficient value} \ge T^+ \quad \text{put } 2 \\ &\text{otherwise} \quad \text{put } 1 \end{aligned} \tag{11}$$

The resulting thresholded vectors of length eight are subsequently converted to decimal numbers in the range of [0, 6560], where $6560=3^8-1$. An example of forming the DC-TFV is shown in Figure 12.



$$Mean = (12+15+12+10+16+10+12+13+17)/9 = 13$$
$$Max = \mathbf{17},\ Min = 10$$
$$T^+ = Mean + f \times (Max - Min) = 13 + 0.3 \times (17-10) = 16.1$$
$$T_- = Mean - f \times (Max - Min) = 13 - 0.3 \times (17-10) = 11.9$$
$$Thresholding([12\ \ 15\ \ 12\ \ 10\ \ 17\ \ 13\ \ 12\ \ 10]) = [1\ \ 1\ \ 1\ \ 0\ \ 2\ \ 1\ \ 1\ \ 0]$$

Figure 12: Example of forming a DC-TFV.

The feature vectors described above represent the information about contents of small image areas in a very concise way. This is because they are based on quantized transform coefficients and thresholding, producing binary or ternary values. The number of feature vectors is limited and can be controlled by the quantization and threshold level while the description is still rich enough to pick up important

local information. This is illustrated next on the example of detection of facial landmarks.

2.5 Local Features and Landmark Detection

The ADM (ACBP Distribution Map) and CDM (Coefficient Distribution Map) distribution maps can be applied to the detection of specific information content in images. Based on certain a priori information about the targeted images, detection algorithms based on the local features derived above can be formulated. In Publications XI and XII they have been used for the detection of facial landmarks.

Facial landmark detection is defined as the process of locating specific areas or contours in a given facial image. Human face and its feature landmarks are very important in various applications such as human face identification, virtual human face synthesis, and human face model coding [33]. Figure 13 illustrates the detection of face landmarks from Publication XI by using the 12^{th} AC transform coefficient from Figure 3. First, the image is block transformed and its corresponding CDM of 12^{th} coefficient is shown in Figure 13(b). The number of non-zero coefficients (after quantization) are summed, first horizontally, then vertically, as shown in (a) and (b). In this way the areas of eye location can be detected.

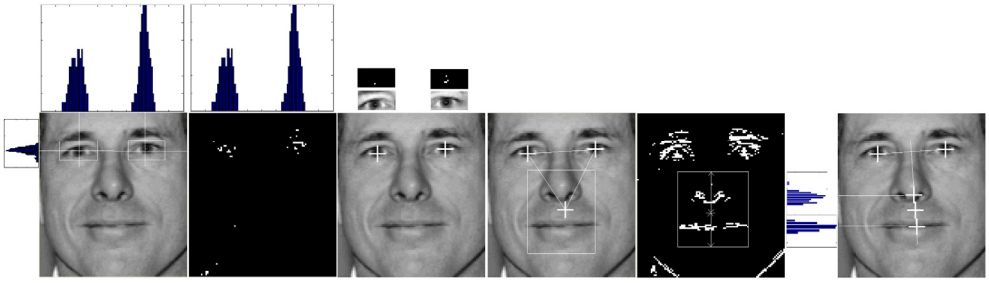


Figure 13: Eye detection process using the CDM.

The eye areas are now processed and the blocks with lowest DC values are identified using a selected threshold. The black color within the eye image of Figure 13(c) shows the locations of eyeballs. Finally, the eye position coordinates are obtained from these black points. This process is illustrated in Figure 13(c).

Approximate location of nose and mouth can now be obtained from the eye position found since they are roughly located on an equilateral triangle base formed by the eyes. The area surrounding the vertex of this triangle is searched and the width of the searching window is the horizontal distance between the eyes. This area is shown in Figure 13 (d),(e). Presuming that the position of nose and mouth is in the middle of eyes, we can calculate the horizontal positions of them.

Another method for landmark detection is proposed in Publication XII using the ACBP. Before locating the nose and mouth areas, the eye area is located first using a pre-defined binary matching template. Such template detects the density of certain ACBP patterns and we call this method Block Density Matching (BDM). The template for the eye is shown in Figure 14.

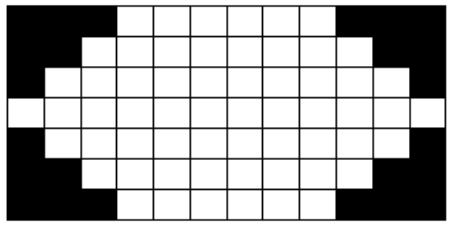


Figure 14: Template used for matching the eye area in Publication XII.

The pixels in the four corners are set to '0' (black points), while the rest are set to '1' (white point), just to approximate the shape of the eye. Moving this template as a sliding window over certain area and matching the area with simple AND operation, enables to locate the eyes. The template can be proportionally adjusted based on the size of targeted image. The number of black points at the four corners can be correspondingly increased or decreased. In the same way position of nose and mouth can be detected using the density of the BDM and the search is done along equilateral triangle as before.

Chapter 3

Feature Histograms

Histograms are widely used for representing global statistical information using some attributes. In image processing, histograms are used to provide visualization of statistical distributions of features like color [2,3,34]. Comparison of the distributions for different images can be used to judge their similarity. In this thesis histograms of feature vectors based on the coefficients of quantized block transforms described in the previous chapter are used for the description of global statistical information. Such histograms are called feature histograms and in this chapter their construction is presented in detail. In particular it is shown how the feature histograms are formed from feature vectors, how different feature histograms are combined, and how the histograms are used for measuring similarity of images.

3.1 Feature Histograms

In Chapter 2 several types of local features and feature vectors based on the 4x4 block transform were described. These features preserve perceptual information even under strong quantization while at the same time quantization allows for flexible control of size of the feature set. Histograms based on the feature sets calculated for natural images will typically have several hundred bins. Description of images via feature histograms only provides information about the distribution of features, without any information about locations of features. In the thesis this is called statistical information, as opposed to structural information which takes locations of features into account. Since histograms are normalized, they can also be regarded as describing probability of features, and also as vectors with unit length.

The feature histogram can be formed in the following way: Let $\boldsymbol{F}=\{F_i,\ 0<i<N\}$ be a set of features. Let $\{B_j,\ 0<j<N\}$ be a partition of $\boldsymbol{F}$ into equivalence classes. If $F_1 \in B_j$, then the "$F_k \in B_j$" happens if and only if $F_k = F_1$. The block histogram is $H_p = \{|B_j|,\ 0<j<N\}$ where $|B_j|$ denotes cardinality of set B_j.

Figure 15 shows histograms of different feature vectors for the same image, the ACBP (AC Block Pattern) histogram in (a), the DCDV (DC Directional Vector) histogram in

(b), the DC-BFV (DC Binary Feature Vector) histogram in (c), and the DC-TFV (DC Ternary Feature Vector) histogram in (d). In these four histograms, the features which constitute the feature histograms are ACBP, DCDV, DC-BFV and DC-TFV respectively. The differences in the shapes of the histograms stem from the fact that each of them describes statistical information of its underlying feature set. One can also observe large variations in the distributions of features within a histogram which indicates that their impacts on statistical information are quite different. This can be seen even more clearly when the histogram bins are sorted in descending order. The histogram distribution has then a long tail since many bins have then very small values, suggesting that they are not crucial for the description of relevant statistical information in images, or they may be even detrimental to it. This leads to the idea that the size of a feature histogram can be reduced by trimming the number of bins. The problem of "what is the optimal number of bins for a given feature histogram" is studied within the context of optimization of histogram retrieval performance in the next chapter.

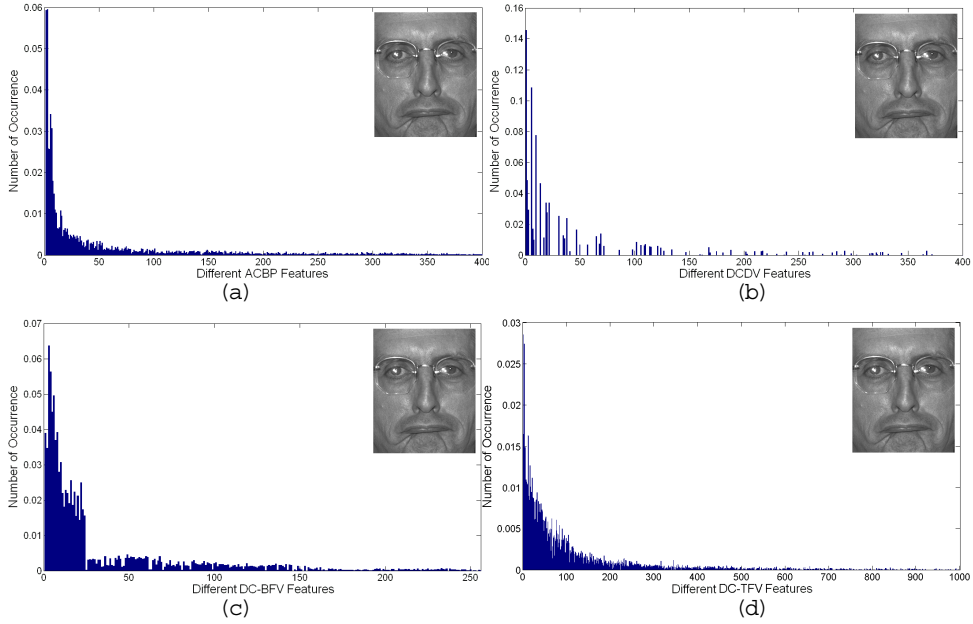


Figure 15: (a) ACBP Histogram. (b) DCDV Histogram. (c) DC-BFV Histogram. (d) DC-TFV Histogram. The X-axis shows different possible variations of certain type of feature. The Y-axis shows their corresponding probability distribution.

3.2 Similarity Measures for Histograms

As mentioned before feature histograms can be seen as normalized vectors. This allows comparing statistical information of images in quantitative way by

calculating the distance between the histograms under suitable norm. Widely used distance metrics include the L_1-norm distance, the L_2-norm distance, Cosine distance (Cos), Correlation distance (Cor) and Standard Euclidean distance (SE). The L_1-norm is also known as the Manhattan (or city-block) distance. It is invariant to the translation or reflection with respect to a coordinate axis, but not to rotation. It is defined as:

$$L1(i,j) = \sum_{b=1}^{L} \left| H_i(b) - H_j(b) \right| \tag{12}$$

The L_2-norm is also known as the Euclidean distance. It is rotation and translation invariant, but highly dependent on the scale of each feature. It can be formulated as:

$$L2(i,j) = \sum_{b=1}^{L} (H_i(b) - H_j(b))^2 \tag{13}$$

The Standard Euclidean distance is proposed based on the L_2-norm, and takes the variances into consideration:

$$SE(i,j) = \sum_{b=1}^{L} \left(\frac{H_i(b) - H_j(b)}{\sigma(b)} \right)^2 \tag{14}$$

with $\sigma(b)$ denoting the variance of all histograms at b^{th} bin.

The Cosine distance measure is actually calculating the angle between two feature vectors, it is defined as

$$Cos(i,j) = \sum_{b=1}^{L} \frac{H_i(b) H_j(b)}{\sqrt{\left\| H_i \right\| \left\| H_j \right\|}} \tag{15}$$

where H_i and H_j are histogram length in the L_2-norm.

The Correlation distance is defined similarly to Cosine distance, but the mean of H_i and H_j are removed respectively.

$$Cor(i,j) = \sum_{b=1}^{L} \frac{(H_i(b) - \overline{H_i})(H_j(b) - \overline{H_j})}{\sqrt{\left\|H_i - \overline{H_i}\right\| \left\|H_j - \overline{H_j}\right\|}} \quad (16)$$

In the later part of the thesis we show the how different similarity measurement affect the performance of proposed retrieval system. According to the results presented in Publication VI, the L_1-norm is identified to be the most suitable measure for our retrieval system giving best results. In addition, one may easily figure out that the L_1-norm is the simplest one to be calculated.

3.3 Combination of Feature Histograms

Feature histograms can be defined for different kinds of features and feature vectors based on the DC and AC coefficients which represent complementary types of local information. These histograms can be combined by the concatenation of histograms defined and applied in Publications VIII – XI as shown in Eq. (17):

[Concatenated_2Histogram] = [Histogram_1 Histogram_2] (17)

In the above concatenation both component histograms have the same impact. More general concatenation is defined by applying weights to the component histograms. Two histograms can be then combined using a single weight α (Eq. (18)) which was defined and applied in Publications I – VIII:

[Combined_2Histogram] = [Histogram_1 α x Histogram_2] (18)

Combination of three histograms was defined with two weights α and β (Eq. (19)) in Publications VI and VII:

[Combined_3Histogram] = [Histogram_1 α x Histogram_2 β x Histogram_3] (19)

With such combinations, different feature histograms can be flexibly integrated and the weights can be optimized to provide most informative description of statistical information as shown later.

3.4 Subarea Histograms and Structural Information

In practice, images contain areas with critical information and areas which are of less importance. An image can be thus seen as a composition of several important subareas. This leads to the idea that feature histogram can be produced for each subarea instead of a single histogram for the whole image. Next, the histograms for subareas can be combined as in Eq. (17). Since each of the subarea histograms is normalized, the combination histograms are in general not equivalent to the full image histogram from the point of similarity measures based on distance norms.

For a formal explanation, assume now that a pattern **P** defined over some area **C** is described by a histogram **H** over a feature set **F** (Figure 16). We shall now define covering of the image area C by a set of subareas C_1 , … , C_n, which do not have to be disjoint. For each of the subareas C_s (s = 1, …, n), its corresponding histogram H_s will be calculated. The description of pattern P is now done using the set of subarea histograms $\{H_1, …, H_n\}$ by forming concatenation of the histogram $H_c = [H_1, …, H_n]$ and patterns can be compared using e.g. city-block metrics.

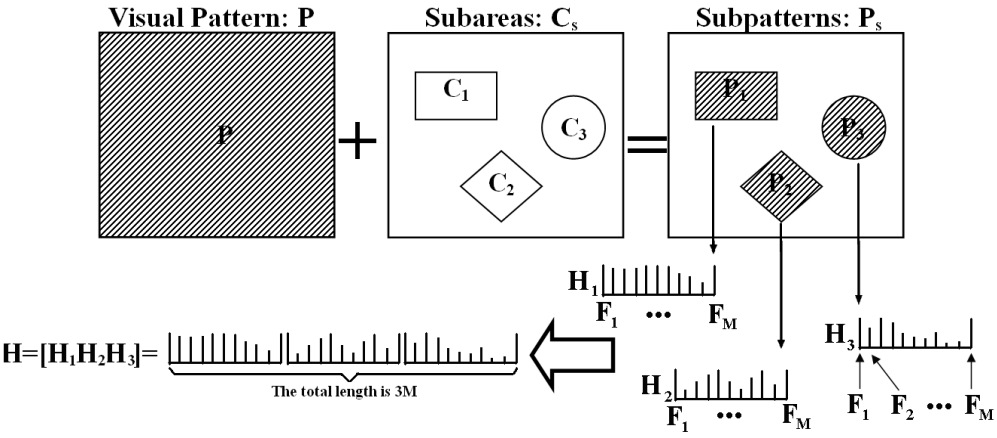


Figure 16: The pattern **P** is covered by the area **C**. The **C** is composed of three subareas: C_1, C_2 and C_3. Single histogram is calculated from each subarea. Each histogram contains M bins, which is corresponding to M features from the feature set **F**. Finally, the three histograms are concatenated in a form of $[H_1\ H_2\ H_3]$, which is description of pattern **P**.

The subarea histograms are normalized by the size of each subarea. Combination of subarea histograms is thus not equivalent to the case when a single histogram is generated from the whole pattern. Hence the impact of different subareas in the metrics is not the same. The reason for this is that information from a single subarea is hidden within a global statistics described by the histogram of full pattern comparing to the case when combined histogram is created with each subarea statistics kept separately. Therefore, treating the subareas separately with their own histograms will magnify the contribution of those subareas in the similarity measure. It can also be seen that this contribution will be equivalent to incorporating certain amount of structural information in the description due to the fact that subareas have specific locations within the image. Generally one could say that the smaller the size of subarea, the bigger is its structural information and its impact on the overall similarity measure.

In the example shown in Figure 16, there are three histograms H_1, H_2 and H_3 which can be combined in the following way

$$[\text{Combined\_3Histogram}] = [\text{Histogram\_1} \quad \text{Histogram\_2} \quad \text{Histogram\_3}] \qquad (20)$$

Each of the histograms also describes some structural information since it is taken from specific subarea. In this way combined histogram of subareas has both statistical and structural information about an image. Structural information is described in a way which allows to combine it with statistical information and can be very flexibly controlled by changing the subareas. In this framework, feature histogram of full pattern has no structural information and the smaller the subarea is the more structural information it carries.

3.5 Luminance Adjustment

A common problem in image processing is variation in local light intensity level called luminance. Different luminance levels hinder the operation of algorithms which aim for comparison of images. This will also concern local features based on block transforms and feature histograms described in previous chapters, especially the ones using the DC coefficients. When applying quantization one has to consider the luminance levels since same quantization will produce different effects in an image taken from a scene at low luminance than from the same scene at higher luminance. To reduce this impact, the luminance level has to be adjusted for images to be processed so that they can be properly compared.

Two types of luminance adjustments were used in the research for this thesis: Histogram Equalization and DC Normalization. The former is used in Publications V – X while the latter is used in Publications II - IV. They are described next.

Histogram Equalization (HE) is applied in the pixel domain before taking block transforms [2,3]. HE modifies the dynamic range and contrast of images by altering it so that the luminance histogram acquires a desired shape. Consider a discrete grayscale image, and let n_i be the number of occurrences of gray level i. The probability of occurrence of a pixel of level i in the image is

$$p(x_i) = \frac{n_i}{n}, i \in 0,\ldots,L-1 \tag{21}$$

where L is the total number of gray levels in the image, n is the total number of pixels in the image.

Let us also define c as the cumulative distribution function corresponding to p, defined by:

$$c(i) = \sum_{j=0}^{i} p(x_j) \tag{22}$$

also known as the image's accumulated normalized histogram.

The goal of HE is to create a monotonic, non-linear transformation of the form $y=T(x)$ that will produce luminance level y for each luminance x in the original image, such that the cumulative probability function of y will be linearized across the value range. The transformation is defined by:

$$y_i = T(x_i) = c(i) \tag{23}$$

Notice that the T maps the levels into the domain of [0,1]. The non-linear mapping reassigns the intensity values of pixels in the input image such that the output image contains a uniform distribution of intensities (i.e. a flat histogram).

This technique allows for areas of lower local contrast to gain a higher contrast without affecting the global contrast, which is useful in images with backgrounds and foregrounds that are both bright or both dark [2].

DC Normalization. The second method used for luminance adjustment is block-based, we call it DC Normalization (DCN). It normalizes the luminance of images by rescaling the block coefficients according to the average luminance level. The average luminance level is calculated based on the DC coefficients of the DCT blocks. Assume there are M images need to be processed, and there are N blocks in image j, and the DC value for each block is denoted by $DC_i(j)$, $1 \le i \le N$. From these DC values, we can calculate the mean DC value for this image:

$$DC_{mean}(j) = \frac{1}{N}\sum_{i=1}^{N} DC_i(j) \tag{24}$$

Next, the average luminance DC_{all} of all images in an image database is calculated by:

$$DC_{all}(j) = \frac{1}{M}\sum_{n=1}^{M} DC_{mean}(j) \tag{25}$$

The rescaling factor of luminance for image j is calculated through:

$$F = \frac{DC_{all}}{DC_{mean}(j)} \tag{26}$$

Next the coefficients of block transform, Eq. (2,6), are rescaled by

$$\overline{BLOCK_{i,j}} = BLOCK_{i,j} \times F, \quad 1 \le i \le N, \qquad 1 \le j \le M \tag{27}$$

After this rescaling, the block coefficients can be quantized by a common quantization coefficient QP

$$\overline{\overline{BLOCK_{i,j}}} = \frac{\overline{BLOCK_{i,j}}}{QP}, \quad 1 \le i \le N, \qquad 1 \le j \le M \tag{28}$$

We found that this normalization is not sensitive to the exact value of rescaling so whenever images are of perceptually tolerable quality (not strongly under- or over-

exposed) the rescaling works well. Figure 17 shows an example image and the corresponding luminance adjustment results by the above two methods.

(a) (b) (c)

Figure 17: The examples of luminance adjustment: (a) Original image, (b) After histogram equalization, (c) After DC Normalization.

Chapter 4

System Model for Image Database Retrieval

In previous chapters, new methods for generating local features and feature histograms were described. Feature histograms are basis for the image database retrieval system considered in the thesis. They have free parameters which can be adjusted within the retrieval framework to ensure its best performance. This is done by parameter optimization using training dataset taken from the domain of retrieval. Optimization is performed under criterion which is used in the standardized performance evaluation method for database retrieval.

4.1 Image Database Retrieval Problem

The image database retrieval problem has become a major research topic recently. The interest in this area is still growing due to the rapid growth of the World Wide Web. Despite of intensive research works undertaken in the past years to design efficient image retrieval systems, there are still no universally accepted feature extraction, indexing and retrieval techniques available [35].

Image retrieval is within an area which is also known as Query by Image Content (QBIC). QBIC means that one needs description of image content. This description can be based on features like colors, shapes or textures which are extracted from the image itself, which can be used as a characterizing representation of the image content. The image retrieval system uses such description, rather than text or metadata, to complete the retrieval task [36-42]. An extension of this research towards a broader scope is the research of content-based multimedia retrieval. In order to provide a generic Multimedia Content Description Interface for these retrieval activities, ISO/IEC has created a multimedia content description standard called MPEG 7 [72].

The image database retrieval problem considered in this thesis is stated as follows. A user provided query image is used as an input to the retrieval system. The retrieval process is carried out by comparing the database image to the query image

by producing description of the query image and comparing it with the descriptions of images stored in the database. A retrieval algorithm decides which images have content similar to the query image based on a similarity measure.
The system performance is measured by the number of correct retrieval decisions. For the best performance the system has free parameters which are tuned by training on a dataset for which correct decisions are known. After the tuning, the system is tested on a dataset and its final performance is evaluated.

4.2 System Description

The main idea of the retrieval system studied in this thesis is the reduction of redundant image information which is irrelevant to the retrieval process. This is done on both local level and global levels. On a local level, the reduction of redundancy is achieved by using quantized block transforms from the video compression area. Such reduction of redundancy means in this case preservation of the minimum perceptually relevant information within the block by quantization and thresholding applied in feature vectors. On the global level, reduction of redundancy is done by adjusting the length of the feature histograms and image subareas. Figure 18 shows the diagram of the proposed retrieval system.

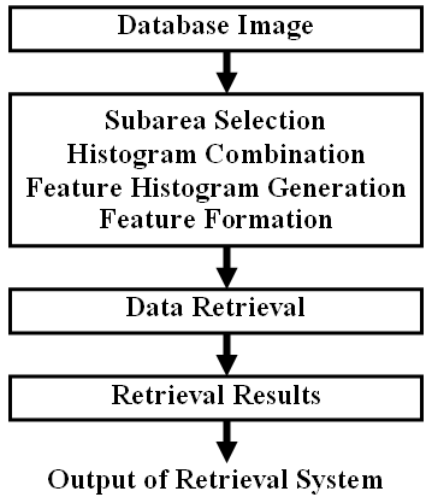


Figure 18: Diagram of the retrieval system. The optimized parameter set obtained from training system is used with the test set for performance evaluation.

There are several parameters described in Chapters 2 and 3 which are used for the optimization of retrieval. They are listed below, with the corresponding notation at the beginning of each item.

QP: Quantization Parameter (as described in Chapter 2.3)

If this parameter is too low, perceptually non-relevant information (e.g., noise) will act towards diminishing the performance. If it is too high, the perceptually relevant information will be unexpectedly removed, thus diminishing the retrieval performance. Therefore, adjustment of this parameter has a goal to find best compromise between the both aspects.

γ: Length of DC Direction-Vector (as described in Chapter 2.4.2)

In Chapter 2.4.3, we have mentioned that the first γ direction-values ($1 \leq \gamma \leq 9$) with largest differences form the DCDV. Properly adjusting γ will optimize the information carried by the DCDV.

C/M: Thresholding method of Binary Feature Vectors (as described in Chapter 2.4.3)

Every BFV is generated from nine coefficients from neighboring blocks. We mentioned in Chapter 2.4.4 that the threshold can be determined either by the central values or mean value of these nine coefficients. The choice between them is made through training process.

***f*: Threshold values of Ternary Feature Vectors (as described in Chapter 2.4.4)**

In Chapter 2.4.5, we have mentioned that there are two thresholds needed to be determined in Eq. (10) which is effectively a selection of the value of f. Proper description of local information can only be achieved when f is neither too high nor too low. It shall adapt to the character of targeted image data which is obtained through the training process.

On the global level the free parameters which are used in the performance optimization process are:

S: Size of the feature histogram (as described in Chapter 3.1)

Feature histograms are constructed by sorting the occurrence of features. Such information about probability of occurrence is obtained from the training data. The resulting histogram has bins arranged in descending order. The length of the ordered histogram is set as a free parameter which is adjusted during the optimization process by successively removing bins from the tail of the histogram. One should expect that removing some of the tail bins from histograms may improve the retrieval performance due to the elimination of irrelevant features. On the other hand

removing too many bins will result in decreasing performance. There should be thus an optimal length of the feature histogram for retrieval.

α,β: Weighting parameters for the combinations of histograms (as described in Chapter 3.3)

Combination of histograms based on different local features has one weight in the case of two histograms and two weights in the case of three histograms which control the impact of particular histogram on the overall performance.

G: The number, size and locations of subarea histograms (as described in Chapter 3.4)

Using combined histogram of subarea histograms can be seen as a way to include certain extent of structural information to the final representation. In general, the higher the number of histograms and the smaller their area, the more structural information is captured by combined histogram. There is great variety of possible selections for the number, size and locations of subareas. They can be used as the optimization parameters but this should be done carefully to manage the computational complexity. In this thesis the question studied is: what the smallest number of subareas is, with simple shapes which results in a very good retrieval performance comparing to other research results. As it is shown later, in practice two and three subareas are sufficient.

The overall optimization is achieved by the training process. Figure 19 shows diagram of the training system. The training process can be characterized as supervised machine learning process [43]. A set of query images and database images are used as the input to the training system, which contains several functional blocks: subarea selection, feature generation, feature histogram formation, histogram combination and database retrieval. The training images are classified in advance by the human operator. The listed optimization parameters are tuned sequentially, until the best retrieval performance is achieved over the training data. After that, they are applied to the test image set for performance evaluation.

4.3 Validation of Training Results

A common problem in the training-based machine learning system is the possible bias caused by the insufficient amount of training data. The optimized parameters obtained from the training set might not be valid for the test data set. In order to validate the robustness of our system, two alternative testing protocols are

applied. The common point among them is: multiple retrieval experiments shall be carried out; within each of them a different training data set shall be used. The final performance is obtained by jointly evaluating these results.

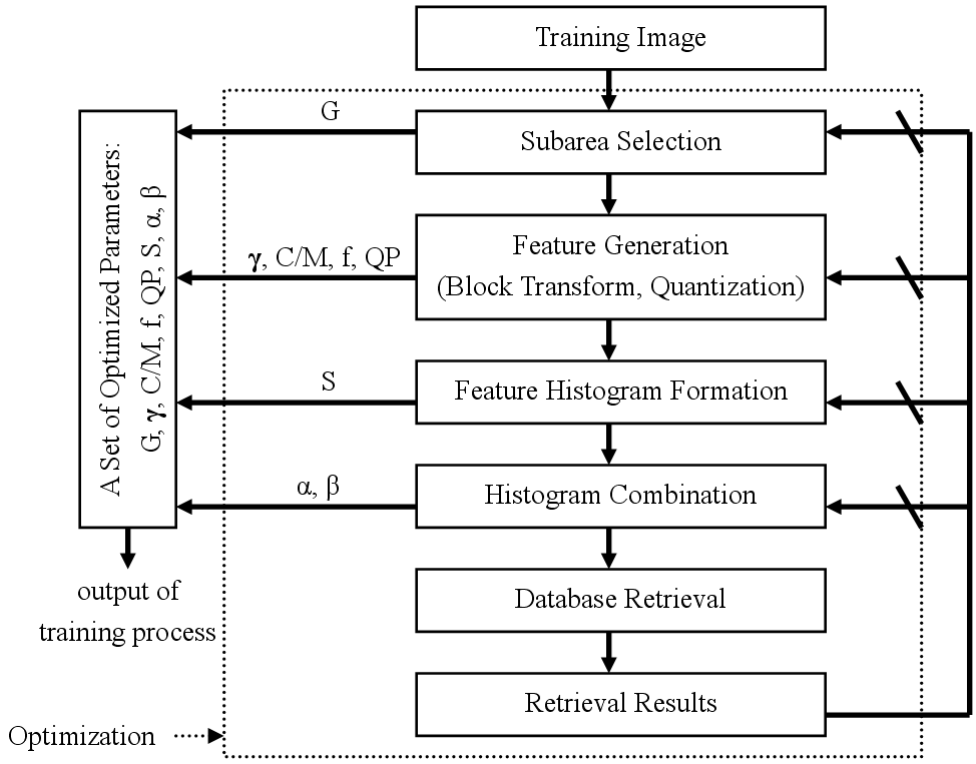


Figure 19: Diagram of the training system. A set of optimized parameters is obtained from the training process, and is subsequently applied in performance evaluation.

Protocol I. Given a large set of images, several different small subsets are selected to be the training sets. The remaining images are organized as a database where the test retrieval is carried out. Each training set will produce one set of optimized parameters which are subsequently used with the test data for performance evaluation. The results obtained from all training sets will be averaged to give the final performance.

Protocol II. Given two large sets of images (collected by different methods, under different environments, or for different purposes), one of them is used as a training data set and the other as a test set, and vice versa. In this way, two retrieval results can be obtained and jointly evaluated. In such case, these two

data sets may be quite different in the sense of properties of image. This kind of robustness validation is highly challenging.

4.4 Quantitative Evaluation of Retrieval Performance

The ideal retrieval result is when the retrieved image is exactly the one user is looking for. Images of the same object share some common characters, while at the same time have certain variations. For any given query image, a correct retrieval is defined as the retrieved image obtained from a database representing the same object as the query image. Based on this specification, the simplest way to evaluate the retrieval performance is the retrieval Accuracy Rate (AR). This rate is the ratio between the number of correct retrievals and number of total retrievals and it changes between 0% and 100%. However, in many practical retrieval systems, the produced retrieval result may be multiple images, rather than a single image. They are ordered according to their corresponding similarity to the query image, and listed sequentially. It is considered sufficient for the correct retrieval if the image being sought is listed within the first N images on the list, and it is then said to have rank N. This rank N retrieval is a looser criterion compared to the AR. In fact, the AR is a special case of it for $N = 1$. But rank N criterion is better for many practical situations. From this point of view, in the thesis research two common performance measures were adopted as the evaluation tool of the retrieval system considered in this thesis: Equal Error Rate (EER) and Cumulative Match Score (CMS). EER was described and used in the Publications II and III. It treats any retrieval result as a decision of acceptance and rejection. EER is achieved when False Rejection Rate (FRR) and False Acceptance Rate (FAR) take equal values, it has been adopted by some researchers as evaluation tool [44,45]. The FRR indicates the possibility of the case that the correct image gets high rank (and thus it will be rejected to be correct retrieval output). The FAR indicates the possibility of the case that the incorrect image gets low rank (and thus it will be accepted as the correct retrieval output). The Cumulative Match Score (CMS) was proposed in [46] as a standardized protocol for the evaluation of retrieval performance. CMS has been widely used by a large number of researchers to compare their results. CMS is a curve which shows the ascending variation of correct retrieval as a function of the rank N. Horizontal axis of the CMS plot is the retrieval rank N and the vertical axis is the percentage of correct retrieval cases. This lets one know how many images have to be examined to get a desired level of performance. For simplicity, many researchers use the CMS at the first rank $N = 1$ as the most tight performance

indicator. We refer to it as "Rank-1 CMS". However, this is not always sufficient for comparison when two CMS curves have a crossing for the first several ranks. As shown in Figure 20, curve A is generally regarded as better than B, since it outperforms B only except in the region with very low ranks [47].

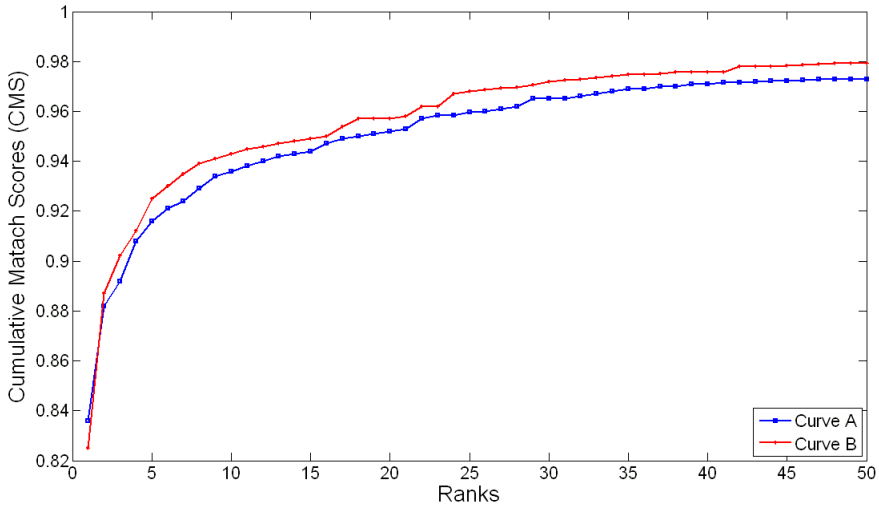


Figure 20: An example when the Rank-1 CMS is not sufficient enough to compare the retrieval performance of tests A and B.

For completeness, the author would also like to mention the well known evaluation measure "precision and recall" [73], which has also been widely used in image retrieval tasks. However, since it is not commonly used for evaluating face image databases, especially for ORL and FERET databases which are used in this research, this method is not utilized here.

Chapter 5

Retrieval Performance Evaluation

In the previous chapter, image database retrieval system based on histograms of local features has been described. Results of the system performance will be summarized now in all publications which this thesis is based on.

Despite of the fact that image database retrieval has been intensively investigated and practical systems have been built, the objective performance evaluation and comparison of results from different methods are generally difficult. Reason for this is that evaluation should be done using identical experimental conditions and performance measures. In the context of image retrieval problem this means that the same test data sets and evaluation procedures should be applied. Ideally, there should be available standardized image data sets covering large variety of domains for exhaustive testing but such sets are not available at present.

Performance evaluation is unsatisfactory if it can not be properly compared with the results of other research. In this thesis the performance is studied using test data and evaluation measures used in the area of face image retrieval since it is an important and widely investigated area. Moreover, common test face image data sets and evaluation methods are available, well documented and widely used. This thesis is not specifically targeted towards face retrieval as the methods developed are completely general. Reasons for choosing face databases as the evaluation platform is the data sets and methods available as well as the wide body of research results which can be used for comparison.

5.1 Face Image Database and Corresponding Evaluation Methodology

Face image retrieval is usually called face identification. This is a one-to-many matching process that compares a query face image against all the images in a face database to determine the identity of the query face. The identification of the test image is done by locating images in the database which have highest similarity with the test image. The identification process is a 'closed' test, which means the retrieval systems locates an individual that is known to be in the database. The

test subject's similarity features are compared to the other features in the system's database and a similarity score is found for each comparison. These similarity scores are then numerically ranked in a descending order. A correct retrieval is defined as the dissimilarity between the query image and its corresponding gallery image is smallest among all the gallery images. Here the gallery means the targeted large image database. Face retrieval is a difficult task for retrieval systems due to large variations of face images even for a single person. Many factors contribute to these variations such as face expression, lighting conditions, pose, characterization (haircut, glasses), time span between taking pictures, etc.

Figure 21: Appearance variations of the same subject under different lighting conditions and different facial expressions.

Figure 22: Appearance variations of the different subject with same facial expressions.

Human face image appearance has potentially very large intra-subject variations due to the factors like: head pose, changing illumination, facial expression, occlusion, facial hair and aging. On the other hand, the inter-subject variations should be smaller due to the similarity of individual appearances. Figure 21 gives examples of intra-subject appearance variations of one subject, and Figure 22 illustrates examples of inter-subject appearance variations among different subjects with the same face expression.

There have been many face retrieval methods proposed and there are also commercial systems used. For instance, the Principal Component Analysis (PCA) [48-50] and the Independent Component Analysis (ICA) [51] are widely used in face retrieval applications, emphasizing on dimension reduction and feature extraction. PCA

transforms a number of correlated variables into a (smaller) number of uncorrelated (orthogonal) variables called principal components. The use of PCA allows the number of variables in a multivariate data set to be reduced, whilst retaining as much as possible of the variation present in the data set. On the other hand, the goal of ICA is to recover independent sources given only observations that are unknown linear mixtures of the unobserved independent source signals. PCA restricts the distribution of the sources to be Gaussian, whereas ICA does not, in general, restrict the distribution of the sources. Among the body of other methods, there are those based on: Artificial Neural Networks (ANN) [52,53], Kernel PCA [54], Isometric Mapping (ISOMAP) [55] and Locally Linear Embedding (LLE) [56].

The face image retrieval techniques can be mainly categorized into two groups [57] based on the face representation which they use: (i) appearance-based which uses holistic texture features; (ii) model-based which employ shape and texture of the face, along with 3D depth information. For the first type, the face recognition problem has been transformed to a face space analysis problem, where a number of well known statistical methods can be tried out. Our proposed method can be classified into the first type. The second type of method relies on models, which have intrinsic physical relationship with real faces and integrated prior human knowledge. This requires explicit modeling of face variations, such as pose, illumination and expression.

The retrieval system performance is evaluated in this thesis using two popular face image databases. First of them is a large face database FERET which is thoroughly documented and with its own evaluation protocol [46]. Another database used is ORL which is much smaller, but has also been widely used in research. These databases are described in detail below together with the respective evaluation protocol.

FERET database is of special importance since it contains more than 10,000 images from more than 1000 individuals taken in largely varying circumstances. Among them, the standardized FA and FB image sets are used here. FA set contains 994 images from 994 different objects, FB contains 992 images. FA serves as the gallery set, while FB serves as the probe set. National Institute of Standard and Technology (NIST) have published several releases of FERET database. The release which we used in the testing is the one published at October 2003, called Color FERET Database, this is important to be noticed since many reference publications are based on other FERET releases (e.g. 2001) so the results are not fully comparable. Some example images after light histogram equalization are shown in Figure 23.

The advantage of using FERET database, apart of its size, is the standardized evaluation method based on performance statistics reported as Cumulative Match Scores (CMS) [46], which have been described before. Horizontal axis of the graph is the retrieval rank and vertical axis is the probability of identification (or percentage of correct matches). The percentage of times that "the highest similarity score is the correct match for all individuals" is referred to as the "top match score". This lets one know how many images have to be examined to get a desired level of performance.

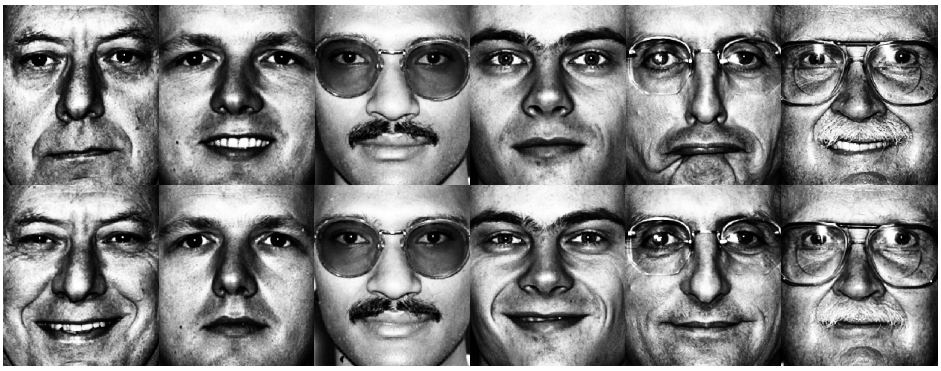


Figure 23: Example images of FERET Database.

In addition, The FERET database also provides the position data of the eyes, nose and mouth for each image. According to whether this information is used, the tests can be divided into two types: fully automatic and partially automatic test. For the partially automatic test, the FERET database provides some general tools for preprocessing of the face images. The images can be cropped by the provided geometrical information about eye, nose and mouth. They can be subsequently aligned, and adjusted by illumination normalization. Some researchers also apply certain mask over the face images, to remove the marginal background. Such preprocessing may improve the performance but it was not applied in the thesis. In our evaluation, the images were simply cropped to certain sizes, which roughly contain the face area. After that the histogram equalization is applied over the cropped image.

The **Olivetti Research Laboratory (ORL)** database [58] contains 10 different images of 40 persons. For some of the persons, the images were taken at different times, with slightly varying lighting, various facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). The ORL has thus more variations for images taken from one person. All the images have dark homogeneous background and the subjects are in up-right, frontal position (with tolerance for

some side movement). For experiment, we store the first 6 images of each person in the database and the remaining 4 images serve as key images. Therefore, the total number of stored images is 240 and the total number of query images is 160. Some example images are shown in Figure 21 and 24.

Figure 24: Example images of ORL Database.

5.2 Results

The retrieval system performance described in Chapter 4 has been evaluated with different local features and system configurations during the course of research which led to this thesis. Results of the research were described in Publications II – X and they are summarized below in the order corresponding to the development of research.

5.2.1 Results for the ACBP and DCDV histograms

The ACBP and DCDV local features were described in Chapter 2.4.1. The retrieval system based on them has been introduced in Publications II – IV, where details of the method are presented.

The database images are processed with 4x4 bock transform. The feature histograms are formed and combined based on the ACBP and DCDV described in Chapter 2.4. They are used for retrieval experiments, which have been evaluated using Protocol II from Chapter 4.3, i.e. the database is split into two sets: one is the training and the other is the testing set, and vice versa. Thus, the cross-validation has been conducted between them. Several CMS curves (explained in Chapter 4.4) produced by above experiments over the FERET database are shown in Figure 25, which represent the number of correct retrievals when N first retrieved images are listed (N<50). For visual evaluation, the higher curve represents better performance. The corresponding Rank-1 CMS results are shown in Table 1, which list the percentage of correct retrieval for the first image provided by the system.

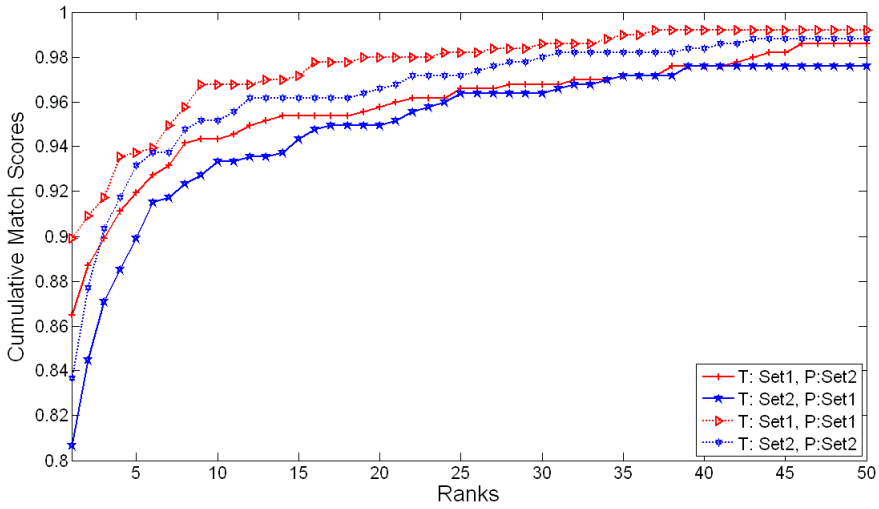


Figure 25: Cumulative match scores results of combined histogram of ACBP + DCDV. P – The parameter evaluation set. T – The testing test used in cross-validation.

Training Set	Testing set	Rank-1 CMS (%)
Set1	Set2	86.49
Set2	Set1	80.65

Table 1: Rank-1 CMS performances over FERET database. P – The parameter evaluation set. T – The testing test used in cross-validation.

5.2.2 Results for the BFV and TFV histograms

The BFV and TFV feature vectors were described in Chapter 2.4.3 and 2.4.4. The retrieval system based on them has been introduced in Publications V-VII where details of the method are presented. In Figure 26 it is shown that combined histogram of ACBP, AC-TFV and DC-TFV histograms provide best results. Comparing to the results for ACBP and DCDV histograms from Figure 25 one can see that the combination of ACBP and TFV histograms provides significantly better retrieval performance than using them individually, with 91% Rank-1 CMS correct retrieval, the corresponding Rank-1 CMS results are shown in Table 2. The training was based on the Protocol I from Chapter 4.3. As one can see, combining the ACBP and TFV histograms provides the best CMS performance. In addition, the TFV performs better than BFV histogram.

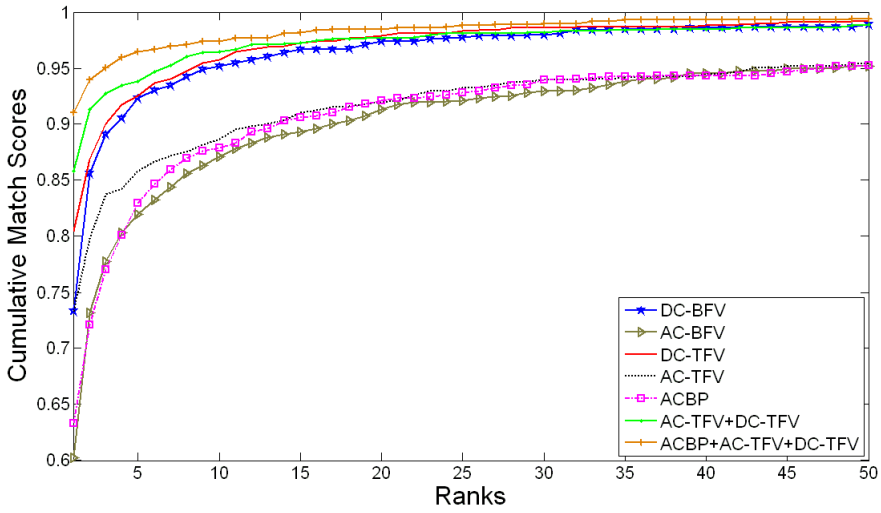


Figure 26: CMS over FERET database using different feature histograms.

FEATURE	DC-BFV	AC-BFV	DC-TFV	AC-TFV	ACBP	AC-TFV+DC-TFV	ACBP+ AC-TFV+DC-TFV
Rank-1 CMS (%)	73.29	60.28	80.44	73.29	63.31	85.79	91.03

Table 2: Rank-1 CMS performances over FERET database.

Another performance evaluation test has been conducted following the training Protocol II from Chapter 4.3. The cross-validation was performed between the FERET and ORL database. The Rank-1 CMS performance results are shown in Table 3. When training and evaluation is performed using ORL our method gives best results with 97.2% correct retrieval. Using the FERET training set and testing on the ORL set the retrieval level of 96.2% is only slightly diminished. Thus indeed the approach is giving very good and robust performance even if the training is performed on another database set.

Rank-1 CMS (%)		Training Set	
		FERET	ORL
Testing Set	FERET	91.03	89.10
	ORL	96.20	97.22

Table 3: Rank-1 CMS performances of Protocol II testing.

5.2.3 Results for the Subarea Histograms

The method of image subarea histograms was introduced in Publications VII - X and described in Chapter 3.4. As indicated there, splitting images into subareas is a way of taking into account some structural information about location of image features as opposed to the statistical information which is picked by full image histogram. Hence this approach should result in best performance when suitable local features are selected. Results will however be strongly influenced by the selection of subareas and the number, location, and size of possible subareas is very large. The goal of the research presented in the thesis was not to search for an optimal set of subareas, but to indicate that simpler approach based on selecting of just few subareas can lead to very good results. To confirm that the splitting into subareas indeed can lead into performance improvements, the approach in Publications VII - X was to generate a sample set of 512 subareas defined randomly and covering the image.

In the training phase using the protocol I the best performing subareas were selected. In the second step, the cases when two subareas are combined together were studied. 216 pairs of subarea were utilized for retrieval. Most of these selected subareas can provide relatively better performance than in the previous experiment and the best pairs were identified. Finally, one additional subarea from different region was added to the above 216 two-subarea pairs creating the case of three subareas with total of 432 combinations of three subareas.

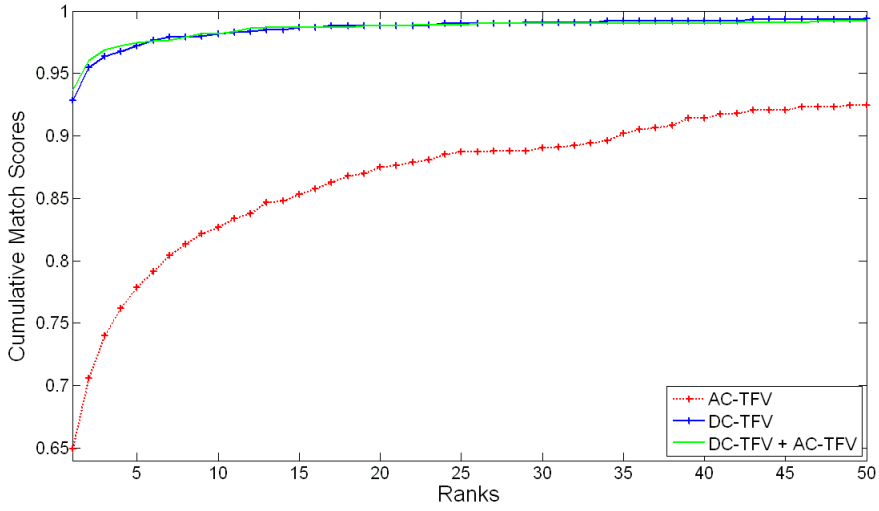


Figure 27: CMS over FERET database using different TFV histograms.

In Figure 27 and Table 4, the CMS results for full image histogram are presented. This is the starting point and reference for the following results.

	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Rank-1 CMS score (%)	92.84	64.31	93.65

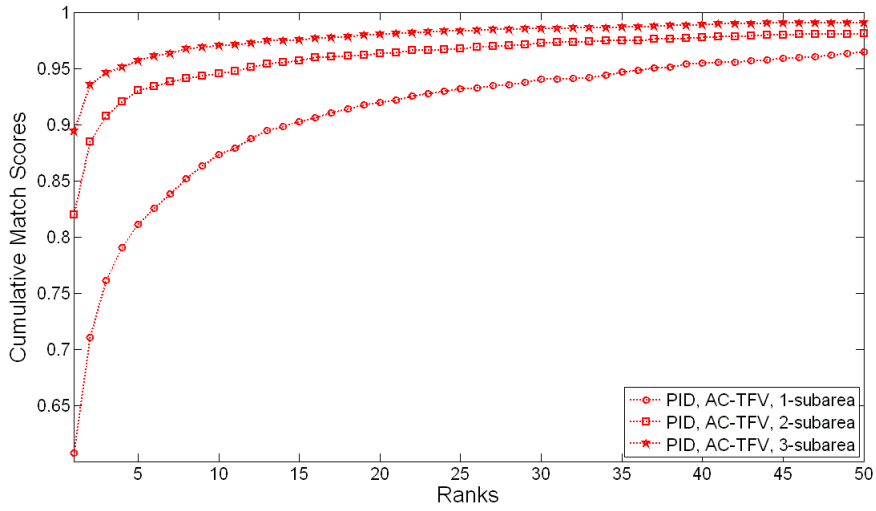
Table 4: Results using full image histogram.

Next, the test is performed over the 512 subareas generated. The maximum, minimum and mean of the resulted 512 CMS scores is shown in Table 5. One can see from it that there is very wide performance variation for different subareas. The DC-TFV subarea histograms always perform markedly better than AC-TFV histograms but their combination performs still better in the critical high performance range. Comparing to the case of full image histograms before, one can see that performance for best subareas is better both for DC-TFV and combination of DC-TFV and AC-TFV histograms. The selection of subarea is thus critical for the performance which can be achieved.

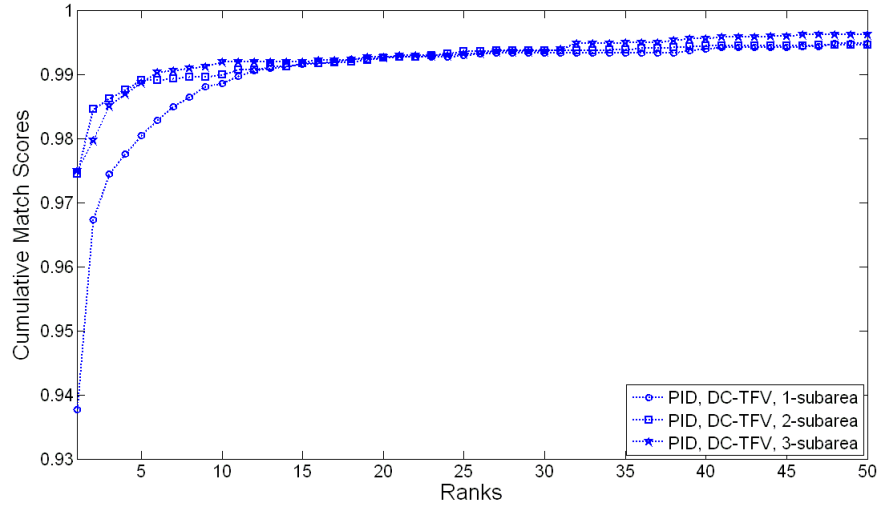
Maximum Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
1-subarea	93.77	60.77	95.30

Table 5: Results using single subarea.

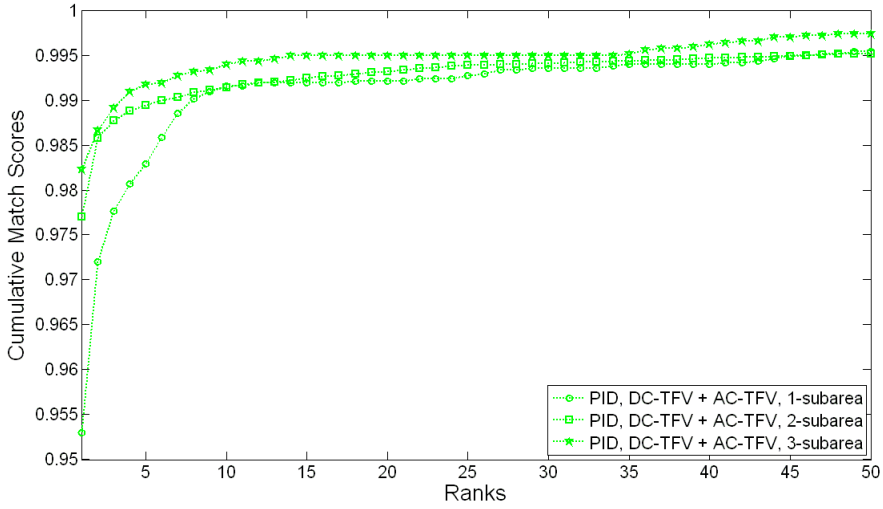
The cases of two and three image subareas are shown in Table 6. Figure 28 gives the CMS curves for corresponding cases.



(a)



(b)



(c)

Figure 28: CMS results over FERET using different number of subareas: (a) using AC-TFV histograms, (b) using DC-TFV histograms, (c) using DC-TFV + AC-TFV combined histograms.

Maximum Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
2-subarea	97.76	81.94	97.70
3-subarea	97.50	89.19	98.23

Table 6: Results using 2 and 3 subareas.

In the above experiments, only the selected image subareas were used, the rest of the image is skipped. It may be argued that this does not use full image information

and may result in diminished performance. Due to this reason the case when subarea(s) histogram(s) is/are combined with the histogram of the rest of the image was also considered. We call this case the Full Image Decomposition (FID) case, in distinction to the previous Partial Image Decomposition (PID) case. An example of 2-subarea FID and PID cases is illustrated in Figure 29.

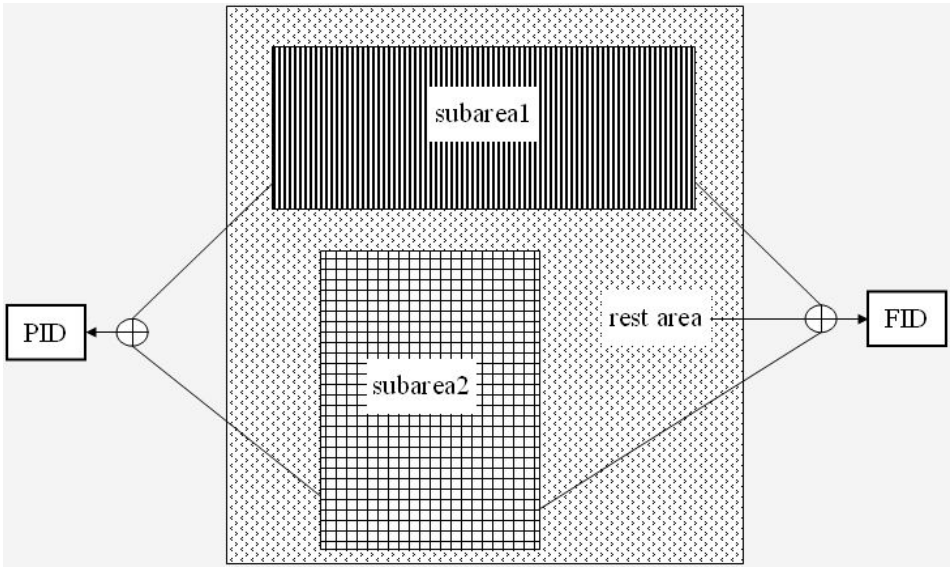


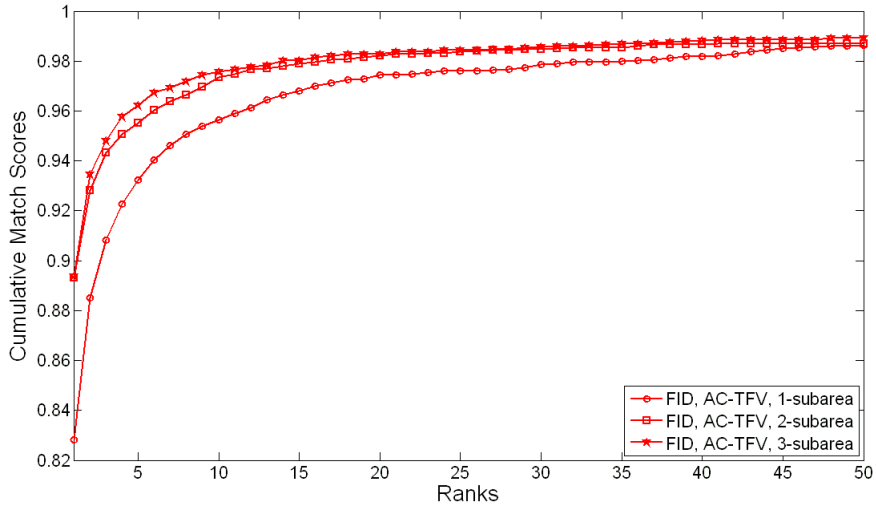
Figure 29: An example of 2-subarea FID and PID case.

The FID case can also be compared to the retrieval with the histogram of full image. In the full image histogram all features have the same impact for the similarity measure, while in the FID case selection of a subarea means increasing the impact of its features in the similarity measure. The corresponding details are described in Publication IX.

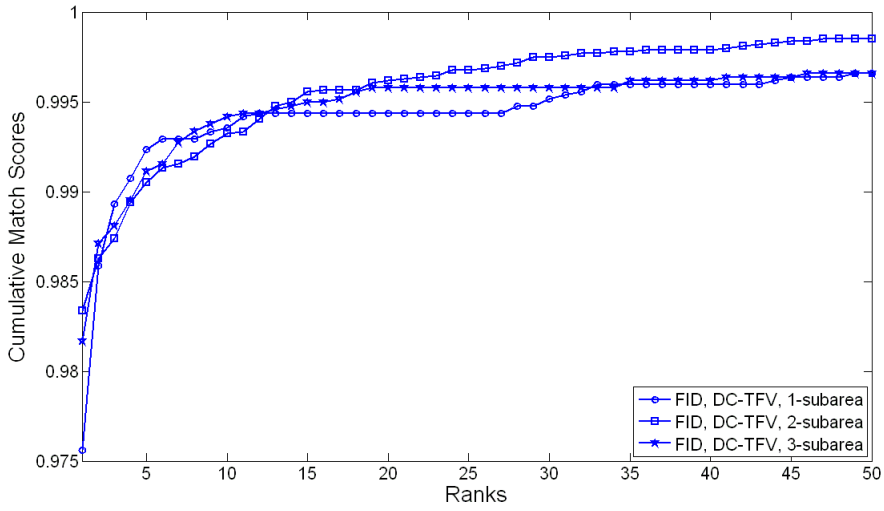
Maximum Rank-1 CMS (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
FID, 1-subarea	97.94	82.82	98.06
FID, 2-subarea	98.43	89.31	98.71
FID, 3-subarea	98.17	89.38	98.63

Table 7: Results of using 1-subarea, 2-subarea and 3-subarea for FID case.

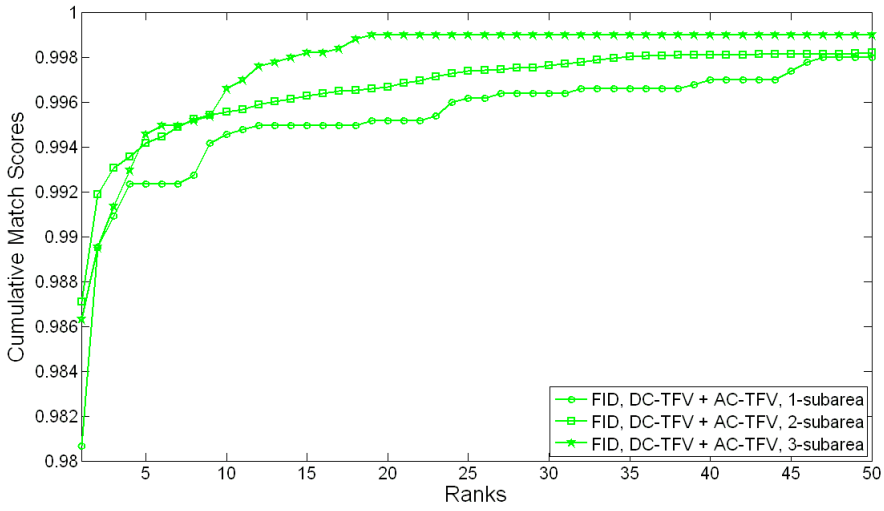
The retrieval performance results of the FID case are shown in Table 7 and Figure 30, which allow us to compare them with the PID case. One can see that again the results of the FID case are better than the results of PID. The reason is that the subareas are selected in a way to emphasize the areas which contribute to more retrieval discriminating ability. In other words subareas in the FID case add structural information to the statistical information which can be obtained from the processing of the whole image.



(a)



(b)



(c)

Figure 30: CMS results over FERET using different number of subareas and (FID case). (a) Using AC-TFV histograms. (b) Using DC-TFV histograms. (c) Using DC-TFV + AC-TFV combined histograms

5.2.4 Faster searching for the optimal subarea

From the above results one can see that by combining subarea histograms it is possible to achieve very good retrieval results when the proper subareas are selected. However, the number of possible subareas is virtually unlimited which makes searching for the optimal ones rather tedious. In order to speed up the search procedure, while at the same time keeping the similar performance, we proposed in Publication X a three-step searching method over the training sets. This method is matched to the face image structure where it can be expected that horizontal rectangular subareas will be highly informative. The searching procedure is thus as follows:

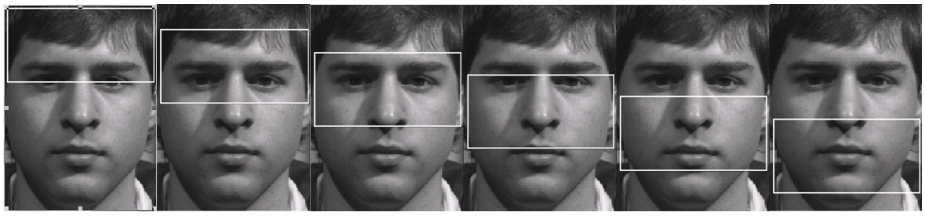
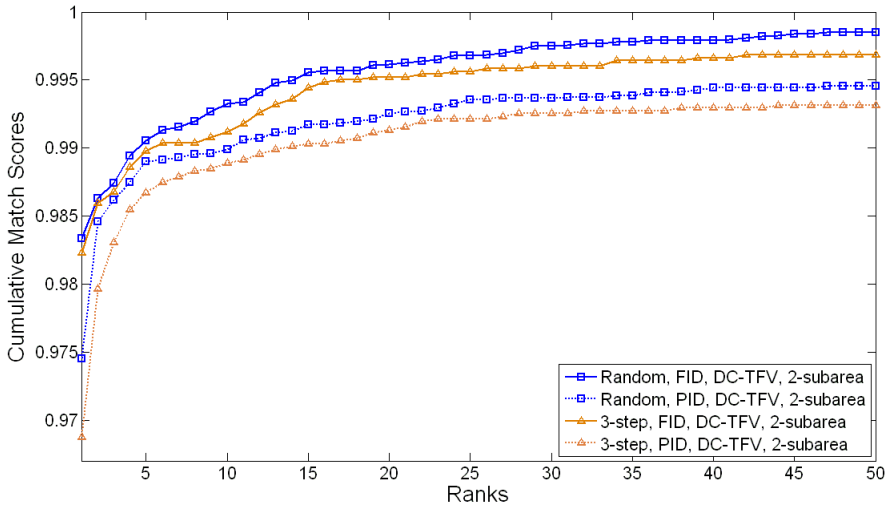


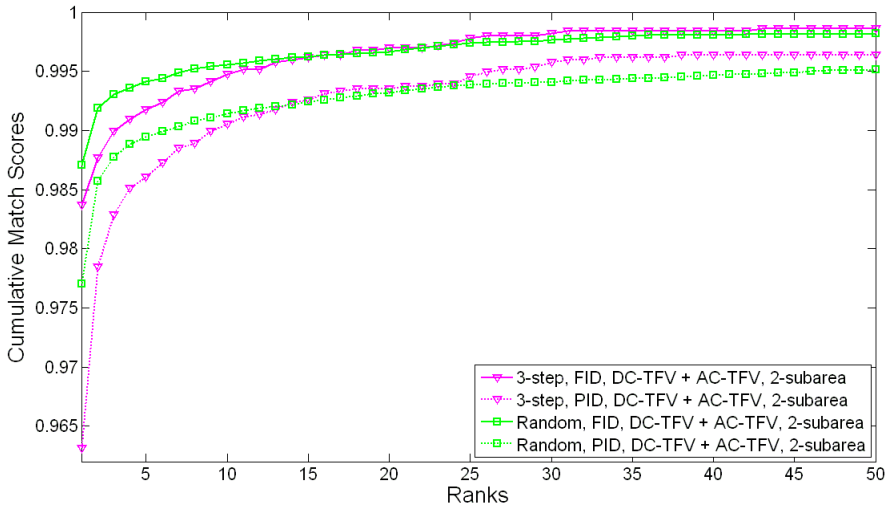
Figure 31: Example subareas from the 1st step of searching.

1. Rectangular areas covering the width of images with different height are considered in the first step. For example, in our experiments with images of size 412x556 pixels, the height of areas is ranging from 40 to 160 pixels, with the width fixed at 400 pixels. The rectangular areas are swept over the picture height in steps of 40 pixels, as shown in Figure 31. From here we have 32 subareas, which is a small subset of above 512 subareas. The subarea giving best result is selected as the candidate of for the next step.
2. The vertical position of the above candidate is fixed and now its width is changed. A number of widths are tested with the training data set and the one with best performance is selected. Here the number of tested widths is 16. After this, the subarea giving best result is selected as the candidate of for the next step.
3. Searching is performed within the small surrounding area of the above candidate. The one giving best result is selected as the final optimal subarea.

The results from the three-step searching are compared (as shown in Figure 32 and Table 8) with the previous results using randomly defined subareas. As one can see, the difference between their performances in the sense of CMS is less than one percent, which is a very good result considering large savings in the computation time and small size of the training set.



(a)



(b)

Figure 32: Comparison between two training methods: randomly defined subareas vs. 3-step searching. Experiments are conducted by using 2-subara decomposition: (a) Using DC-TFV histograms (b) Using DC-TFV + AC-TFV combined histograms.

Rank-1 CMS score (%)	Random Subareas		3-Step Searching	
	DC-TFV	DC-TFV + AC-TFV	DC-TFV	DC-TFV + AC-TFV
PID case	97.76	97.70	96.83	96.31
FID case	98.43	98.71	98.23	98.37

Table 8: Comparison between the results using 3-Step Searching and the results using randomly defined subareas (2-subarea). The difference between the resulting CMS scores is less than one percent.

5.2.5 Comparison to other Research Results

In order to compare the performance of our system with other methods, we list below some reference results from other research for the FERET database. These results are all obtained by using the FA and FB set of the same release of FERET database. In [59], the eigenvalue weighted bidimensional regression method is proposed and applied to biologically meaningful landmarks extracted from face images. Complex principal component analysis is used for computing eigenvalues and removing correlation among landmarks. An extensive analysis of this method is conducted in [60], with comparison of the effectiveness of four similarity measures including the typical L_1-norm, L_2-norm, Mahalanobis distance and eigenvalue-weighted cosine (EWC) distance. In [61] a simple template matching method is used to complete a

verification task. The input and model faces are expressed as feature vectors and compared using a distance measure between them. A combined subspace method is proposed in [62], using the global and local features obtained by applying the LDA-based method to either the whole or part of a face image respectively. The combined subspace is constructed with the projection vectors corresponding to large eigenvalues of the between-class scatter matrix in each subspace. The combined subspace is evaluated in view of the Bayes error, which shows how well samples can be classified. Table 9 lists the result of above papers, as well as the result of 2-subarea FID (2-FID) case of our method. The results are expressed by the way of Rank-1 CMS score.

REFERENCE	[59]	[60]	[61]	[62]	proposed
METHOD	Landmark Bidimensional Regression	Landmark	Template Matching	Combined Subspace	2-FID Method
Rank-1 CMS (%)	79.4	60.2	73.08	97.9	98.71

Table 9: List of the referenced results based on release 2003 of FERET database.

In addition, Table 10 lists the performances of some other methods which use different version of the FERET database.

REFERENCE	[63]		[64]	[65]
METHOD	PCA-L1	ICA-Cosine	Boosted Local Features	JSBoost
Rank-1 CMS (%)	80.42	60.2	73.08	97.9

Table 10: List of some other referenced results based different release of FERET database.

Furthermore, we also list in Table 11 the referenced results over ORL database, which use the same testing procedure as used in the thesis.

REFERENCE	[66]	[67]	[68]	[69]	[70]	[71]	proposed	proposed
METHOD	IPCA_PCA	VQ	Local Gabor Feature	ARENA	SVM+DWT	Template Matching	Training ORL	Training FERET
Rank-1 CMS (%)	88.3724	97	85	96.2	90.8	92.5	97.22	96.2

Table 11: Performance over ORL database in comparison to references.

5.2.6 Results of Detecting Facial Landmarks

Two landmark detection methods were proposed in Publications XI and XII. They have been briefly described in Chapter 2.5, some results are presented below. Figure 33 is obtained using CDM and Figure 34 is obtained using BDM described in Chapter 2.5.

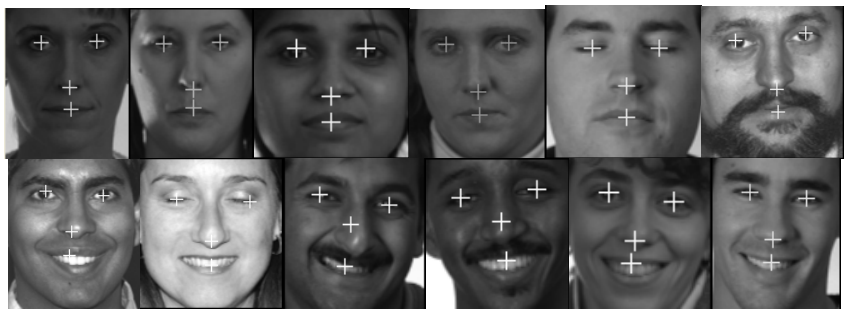


Figure 33: Some example detection results using CDM.

Figure 34: Some example detection results using BDM.

In Table 12 quantitative results of CDM are shown as the number of false face feature detections among the overall 360 detection tests. The performance is shown for the combinations of up to four different AC coefficients have. The best accuracy rate achieved is about 91.7%.

AC Coefficients	12	12+8	12+8+10	12+8+10+2+3
False Detections	51	33	30	42

Table 12: Number of false detections among 360 detection tests using the CDM.

As one can see, both methods are able to detect rough locations of the facial landmarks even though the face expressions and lighting conditions vary quite significantly. Even in some challenging cases, e.g. when the eyes are closed, the proposed method may still work satisfactorily.

5.2.7 Discussion of results

The results presented in this chapter show performance of a simple retrieval system. The system is based on local features constructed from block transforms. The features are combined into histograms and optimized by a training process providing good retrieval performance. This is achieved by using statistical information for the retrieval. Increasing the sophistication of local features and combining the feature histograms leads to consistent performance improvement. Finally, the best results are achieved by including limited structural information by considering combination of two and three image subareas. The results are then on the level of best other methods which are remarkable when taking into account the simplicity of the proposed approach comparing with many other methods.

Chapter 6

Conclusions

The image database retrieval is nowadays an active research topic, with very wide scope of methods. In this thesis the algorithms for image database retrieval based on block transforms and histograms are proposed. Experiments based on public image databases and standardized testing protocols are carried out to prove the efficiency and robustness of the proposed methods.

The proposed approach starts from the generation of flexible sets of local features which can characterize the content of images. They are formed based on coefficients of quantized block transforms. The block transforms with quantization are effectively able to remove perceptually redundant and irrelevant information. In addition, the transform and feature vectors make the description of local features compact, which facilitate their use. In the thesis, the feature vectors like ACBP, DCDV, BFV and TFV are proposed to describe the local features.

For the description of statistical information of images feature histograms are proposed. The feature histograms are used for comparison of images using similarity measure. Histograms of different feature vectors are combined together to improve the retrieval performance. The information content of feature histograms is optimized by changing their length.

In addition to the statistical information represented by combined feature histograms, structural information is incorporated by applying image subarea-based scheme. The feature histograms are generated for image subareas, and subsequently used in combination histograms histogram. The number of the subareas, as well as their size, shape and locations, reflects the complex nature of structural information. In this thesis, it is shown that very good retrieval performance results can be achieved when just two or three properly selected subareas are combined.

Based on the above developments, hierarchical retrieval system is proposed in the thesis. In this system, free parameters for feature vectors, histograms, and subareas are tuned using the training data set. As a result, a set of parameters

ensuring the best performance of the retrieval system is obtained. Results of performance evaluation using face databases show that the proposed system has performance better than many other existing methods and on the level of the best results to date. On the other hand, the system is simpler, and has not been specifically developed with focus on face images. The performance evaluation shows that the proposed system achieves about 99% accuracy with the FERET database, which is sufficient for practical applications.

The approach presented in the thesis combines statistical and structural information of images in a novel way which is reducing the necessary information content to achieve specific retrieval performance level. Further research in this direction is required using multi-resolution image representation and enhanced description of structural information.

BIBLIOGRAPHY

[1]. N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete Cosine Transform", IEEE Trans. Computers, C-23, pp. 90-93, Jan 1974.

[2]. R. Gonzalez and R. Woods, "Digital image processing (3rd Edition)", Addison-Wesley Publishing Company, 2007. ISBN 0-201-50803-6.

[3]. John C Russ, "The Image Processing Handbook", CRC Press, 2002. ISBN 0-8493-7254-2.

[4]. Seong-O Shim, Tae-Sun Choi , "Edge color histogram for image retrieval", Proceeding of International Conference on Image Processing, pp. 957-960, vol.3, June 2002.

[5]. Zhou X.S and Huang T.S., "Edge-based structural features for content-based image retrieval", Pattern Recognition Letters, 22(5):457-468, April 2001.

[6]. Canny, J., "A computational approach to edge detection", IEEE Trans. Pattern Analysis and Machine Intelligence, 8:679-714, 1986.

[7]. Ziou, D. and Tabbone, S., "Edge detection techniques: an overview", International Journal of Pattern Recognition and Image Analysis", 8(4):537--559, 1998.

[8]. S. Mallat, "Wavelets for a Vision", Proceeding of the IEEE, Vol. 84, No. 4, pp. 604-614, April 1996.

[9]. Keun-Chang Kwak and Witold Pedrycz, "Face recognition: a study in information fusion using fuzzy integral", Pattern Recognition Letters, Vol. 26, Issue 6, pp. 719 - 733, 2005.

[10]. J.T. Chien and C.C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12), 1644–1649, 2002.

[11]. Ekenel H K and Sankur B., "Multiresolution face recognition". Image and Vision Computing, 23 (5):469-477, 2005.

[12]. Gonzalo Pajares and Jesus Manuel de la Cruz, "A wavelet-based image fusion tutorial", Pattern Recognition, Vol. 37, 1855 – 1872, 2004.

[13]. C. Garcia, G. Zikos and G. Tziritas, "Wavelet packet analysis for face recognition", Image and Vision Computing, 18(4), 289–297, 2000.

[14]. Annalisa Franco, Alessandra Lumini, Dario Maio and Loris Nanni, "An enhanced subspace method for face recognition", Pattern Recognition Letters, Vol. 27, No. 1, pp. 76-84, 2006.

[15]. Loris Nanni and Dario Maio, "Weighted Sub-Gabor for face recognition", Pattern Recognition Letters, Volume 28, Issue 4, 1, Pages 487-492, 2007

[16]. H. Abrishami Moghaddam, T. Taghizadeh Khajoie, A.H. Rouhi, and M. Saadatmand Tarzjan, "Wavelet correlogram: An approach for image indexing and retrieval", Pattern Recognition, Vol. 38, 2506 – 2518, 2005.

[17]. Phillips, P.J., "Matching pursuit filters applied to face identification", IEEE Transactions On Image Processing, Vol. 7, No. 8, 1998 .

[18]. Chengjun Liu and Harry Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition", IEEE Transactions On Image Processing, Vol. 11, No. 4, 2002.

[19]. Y. Zhong and A. K. Jain, "Object localization using color, texture and shape", Pattern Recognition, Vol.33, No.4, pp.671-684, 2000.
[20]. R. E. Frye and R. S. Ledley, "Texture discrimination using discrete cosine transformation shift-insensitive (DCTSIS) descriptors", Pattern Recognition, Vol.33, No.10, pp.1585-1598, 2000.
[21]. S. Eickeler, S. Muller and G. Rigoll, "Recognition of jpeg compressed face images based on statistical methods", Image and Vision Computing, Vol. 18, No. 4, pp. 279-287, 2000.
[22]. D. Ramasubramanian and Y. V. Venkatesh, "Encoding and recognition of faces based on the human visual model and DCT", Pattern Recognition, Vol.34, No.12, pp. 2447-2458, December 2001.
[23]. Jie. Wei, "Image segmentation based on situational DCT descriptors", Pattern Recognition Letters, Vol. 23, No.1-3, pp. 295 - 302, January 2002.
[24]. J. Jiang, A. Armstrong and G. Feng, "Direct content access and extraction from JPEG compressed images", Pattern Recognition, Vol.35, No.11, pp.2511-2519, November 2002.
[25]. C. Sanderson and K. K. Paliwal, "Fast features for face authentication under illumination direction changes", Pattern Recognition Letters, Vol. 24, No.14, pp. 2409 - 2419, 2003.
[26]. M. S. Kim, D. Kim and S. Y. Lee, "Face recognition using the embedded HMM with second-order block-specific observations", Pattern Recognition, Vol. 36, No. 11, pp.2723-2735, November 2003.
[27]. G. Feng, J. Jiang, "JPEG compressed image retrieval via statistical features", Pattern Recognition, Vol.36, No. 4, pp. 977-985, April 2003.
[28]. C. Sanderson and K. K. Paliwal, "Features for Robust Face Based Identity Verification", Signal Processing, Vol.83, No.5, pp.931-940, May 2003.
[29]. JPEG Standard (JPEG ISO/IEC 10918-1 ITU-T Recommendation T.81)
[30]. ISO/IEC 13818-2:2000 Information Technology - Generic Coding of moving pictures and associated audio information: Video.
[31]. ITU-T recommendation H.264 | ISO/IEC 14496-10 AVC, Draft ITU-T recommendation and final draft international standard of joint video specification.
[32]. Ojala, T., Pietikäinen and M. Harwood, D., "A comparative study of texture measures with classification based on feature distributions", Pattern Recognition, Vol. 29, No. 1, pp. 51-59, 1996.
[33]. JTC1/SC29/WG11; MPEG-4, Final Draft of International Standard, Part 2 (Visual). Doc. No. N2502 of ISO 14496-1, 1998.
[34]. Thomas Deselaers, Daniel Keysers, Hermann Ney, "features for image retrieval: a quantitative comparison", Pattern Recognition, 26th DAGM Symposium, 2004.
[35]. Goodrum, A.A., "Image information retrieval: An overview of current research", Informing Science, Vol. 3, Iss. 2, pp. 63-67, 2002.
[36]. Remco C. Veltkamp and Mirela Tanase, "Content-Based Image Retrieval Systems: A survey", Technical report, Available at: http://www.aa-lab.cs.uu.nl/cbirsurvey/cbir-survey/.
[37]. 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, Vol. 22, Iss 12, pp. 1349-1380, 2000.
[38]. Del Bimbo, "Visual information retrieval", Morgan Kaufmann Publishers, San Francisco, 1999.
[39]. Ritendra Datta, Dhiraj Joshi, Jia Li and James Z. Wang, "Image retrieval: ideas, influences, and trends of the new age", Technical report, Available at: infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta.pdf

[40]. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele and P. Yanker, "Query by image and video content: the QBIC system", IEEE Computer, Vol. 28, Iss. 9, pp. 23-32, 1995.
[41]. N. Fuhr, "Information retrieval methods for multimedia objects", State-of-the-Art in Content-Based Image and Video Retrieval, pp. 191-212, Kluwer, Boston, 2001.
[42]. Yong Rui, Thomas S. Huang and Shih-Fu Chang, "Image retrieval: Current techniques, promising directions and open issues. Journal of Visual Communication and Image Representation, 10(1):1–23, March 1999.
[43]. Christopher M. Bishop (2007) Pattern Recognition and Machine Learning, Springer ISBN 0-387-31073-8.
[44]. R. M. Bolle, S. Pankanti and N. K. Ratha, "Evaluation techniques for biometrics-based authentication systems (FRR)", Proc. International Conference on Pattern Recognition, Vol. 2, pp. 831 – 837, Sept 2000.
[45]. K. Kotani, C. Qiu and T. Ohmi, "Face Recognition Using Vector Quantization Histogram Method", International Conference on Image Processing, II-105, Sep. 2002.
[46]. P. J. Phillips, H. Moon, P. J. Rauss and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, October 2000.
[47]. Delac, K., Grgic, M., Grgic, S., "Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set", International Journal of Imaging Systems and Technology, Vol. 15, Issue 5, pp. 252-260, 2006.
[48]. H. Hotelling, "Analysis of a complex of statistical variables into principal components", Journal of Educational Psychology, 24:417–441, 1933.
[49]. M.A. Turk and A.P., "Pentland. Face recognition using eigenfaces", In Proceedings of the Computer Vision and Pattern Recognition, pages 586–591, 1991.
[50]. M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces", IEEE Trans. Pattern Anal. Mach. Intel., Vol 12, No. 1, pp. 103-108, 1990.
[51]. A.J. Bell and T.J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution", Neural Computation, 7(6):1129–1159, 1995.
[52]. H. A. Rowley, S. Baluja and T. Kanade, "Neural Network-Based Face Detection", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp. 23-38, Jan. 1998.
[53]. K.-K. Sung and T. Poggio, "Example-Based Learning for View-Based Human Face Detection", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp. 39-51, Jan. 1998.
[54]. B. Scholkopf, A. Smola and K. Muller, "Nonlinaer component analysis as a kernel eigenvalue problem", Neural Computation, vol. 10, no. 5, pp. 1299–1319, 1998.
[55]. J.B. Tenenbaum, V. de Silva and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol. 290, pp. 2319–2323, 2000.
[56]. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", Science, vol. 290, pp. 2323–2326, 2000.
[57]. X. Lu, "Image Analysis for Face Recognition – A brief survey," personal notes, May 2003, Avalainle at: http://www.cse.msu.edu/~lvxiaogu/research/FaceRcg_survey.htm.
[58]. ORL Face Database, 2005. Available at: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. AT&T Laboratories Cambridge.
[59]. Jiazheng Shi, Ashok Samal and David Marx, "Face recognition using landmark-based bidimensional regression", Proceeding of IEEE International Conference on Data Mining, 765-768, 2005.

[60]. Jiazheng Shi, Ashok Samal and David Marx, "How effective are landmarks and their geometry for face recognition", Computer Vision and Image Understanding, 102(2): 117-133, 2006.
[61]. Roure, J. and Faundez-Zanuy, M., "Face recognition with small and large size databases", Proceeding of 39th Annual International Carnahan Conference on Security Technology, 153-156, 2005.
[62]. Chunghoon Kim, Jiyong Oh and Chong-Ho Choi, "Combined subspace method using global and local features for face recognition", Proceedings of International Joint Conference on Neural Networks, 2030-2035, 2005.
[63]. Kyungim, B., Bruce, A.D., J.Ross, B. and Kai, S., "PCA vs. ICA: A comparison on the FERET data set", Proceeding of International Conference on Computer Vision, Pattern Recognition and Image Proceeding, North Carolina:Durham, 2002.
[64]. Jones, M. and Viola, P., "Face recognition using boosted local features", Proceeding of International Conference on Computer Vision, 2003.
[65]. Xiangsheng, H., Li, S.Z. and Yangsheng, W., "Jensen-Shannon boosting learning for object recognition", Proceeding of CVPR 2005, Vol. 2, 144- 149, 2005.
[66]. Issam, D. and Rabih, N., "Face Recognition Using IPCA-ICA Algorithm", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 28, No. 6, June 2006.
[67]. Kotani, K., Qiu, C. and Ohmi, T., "Face Recognition Using Vector Quantization Histogram Method", Proceeding of International Conference on Image Processing, II-105, 2002.
[68]. Erik, H., "Biometric Systems: A Face Recognition Approach. Proceeding of the Norwegian Conference on Informatics, pp. 189-197, 2002.
[69]. Rahul, S., Matthew, M. and Shumeet, B., "Memory-based face recognition for visitor identification", Proceeding of IEEE Face and Gesture, 2000.
[70]. Travieso, C.M., Alonso, J.B. and Ferrer, M.A., "Strategy for improving the reliability in the facial identification2, Proceeding of 39th Annual, International Carnahan Conference on Security Technology, 2005.
[71]. Roure, J. and Faundez, Z.M., "Face recognition with small and large size databases", Proceeding of 39th Annual International Carnahan Conference on Security Technology, 2005.
[72]. Shih-Fu Chang, Thomas Sikora and Atul Puri, "Overview of the MPEG-7 Standard," IEEE Transactions on Circuits and Systems for Video Technology, 11 (6), June 2001, 688-695.
[73]. Makhoul, John; Francis Kubala; Richard Schwartz; Ralph Weischedel: Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999.

Publication I

DaiDi Zhong, Irek Defée, "Pattern Recognition in Compressed DCT Domain", in Proceedings of IEEE International Conference on Image Processing (ICIP2004), pp. 2031-2034, October 2004

PATTERN RECOGNITION IN COMPRESSED DCT DOMAIN

Daidi Zhong, Irek.Defée

Institute of Signal Processing
Tampere University of Technology
P.O. Box 553,
FIN-33101 Tampere , FINLAND.
E-mail: daidi.zhong@tut.fi irek.defee@tut.fi

ABSTRACT

Images and video are currently predominantly handled in compressed form. Block-based compression standards are by far the most widespread. It is thus important to devise information processing methods operating directly in compressed domain. In this paper we investigate this possibility on the example of simple face information processing method based on the DCT (Discrete Cosine Transform) blocks. We use patterns of quantized 4x4 DCT blocks for representing local picture information. These patterns at different quantization levels provide very flexible representation of picture information. We represent global information in pictures by histograms of quantized DCT pattern distributions. The approach is tested on database of face images and it is shown that despite its simplicity provides good results in the face recognition problem.

1. INTRODUCTION

Images and video are handled nowadays mostly in compressed format. This is due to the reduced size of the data which diminishes the need for storage and communication bandwidth. The compression methods used are highly optimized for the removal of all non relevant information while preserving high perceptual quality. Compression leaves only perceptual content and this is desirable not only from the data size reduction point but also for the visual information extraction. Hence visual information extraction in compressed domain should be advantageous but this point has not been widely followed due to the lack of proper approach. Such an approach is proposed in this paper and initial study illustrates certain important aspects of it. We show that recognition results are very good considering the simplicity of method and that with adaptive selection of quantization results are excellent.

In this paper we study the problem of information extraction in compressed domain on the example area of face detection and recognition. This topic has been extensively studied in the past from a large variety of viewpoints. Face detection and recognition technology has also been widely used in many fields, such as personal identification, video surveillance system, human computer interaction, etc. In recent years, many peoples are working in this field to find efficient ways to recognize face images fast and accurately [1],[2]. Recently face information became important in multimedia information retrieval systems. Instead of the original linguistic description of images using some key words, these methods is more powerful for image indexing and face recognition, as it automatically considers the information from images themselves. This information is so-called visual features, a number of key visual descriptors of facial content, which can distinguish one face from other faces.

Generally, face information processing is based on extraction of local visual features, such as color, texture, shape and integrating them into data structure describing face. Recently simple histogram based methods were proposed for the face detection and recognition problems [3],[4]. In these methods histogram of basic features extracted is formed. Despite of simplicity of this approach results are surprisingly good. In [3] histogram method is applied to features extracted by block Vector Quantization (VQ). In [4] histograms based on skin color features were applied for face detection problems. Both these approaches give good results considering simplicity of methods used.

In this paper we use histogram method with feature extraction based on 4x4 DCT blocks. This has advantage

0-7803-8554-3/04/$20.00 ©2004 IEEE.

of enabling direct extraction in the compressed domain where 8x8 DCT is used. Scaling from the 8x8 DCT size to 4x4 DCT is straightforward and has been used in the past for picture downsizing [5]. Using the DCT for feature extraction has many advantages. The transform is easy to calculate, it is well preserving perceptual information under quantization in wide range of values. Using the DCT of size 4x4 is good compromise between the amounts of detail block size; same size is often used in application of vector quantization to image compression [4]. The difference between the VQ and DCT approach lays in much more flexible control of feature extraction by quantization and good properties of DCT in matching perceptually important features.

The approach presented illustrates the strength of the extraction of shape information in the DCT domain on the example of face recognition.

2. FEATURE DESCRIPTION USING DCT

For the feature extraction we use 4x4 DCT blocks. The 2-D DCT can be calculated directly by:

$$G[m,n] = a(m)a(n)\sum_{i=0}^{N-1}\sum_{k=0}^{N-1} g(i,k)\cos\left[\frac{\pi(2i+1)m}{2N}\right]\cos\left[\frac{\pi(2k+1)n}{2N}\right] \quad (1)$$

$$a(0) = \sqrt{\frac{1}{N}} \quad \text{and} \quad a(m) = \sqrt{\frac{2}{N}},\ 1 \le m \le N$$

Here, g is the source block and the G is the DCT transformed block. N is the dimension of the blocks.

In compression standards 8x8 DCT is commonly used. The 4x4 DCT coefficients can be obtained from the 8X8 DCT blocks. One possible way is as shown as below:

$$DCT(B') = DCT(H)DCT(B)DCT(H^T) \quad (2)$$

Here, HT is the transpose matrix of H. B is the 8x8 source block and the B' is the 4x4 output block. H is defined as below, [5]

$$H = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix} \quad (3)$$

In compression applications the DCT is quantized. Quantization eliminates very effectively non-perceptual information overhead. 8x8 blocks of DCT coefficients are quantized using quantization matrices. Methods for deriving quantization matrices based on the properties of human visual system were also proposed [7]. The 4x4 DCT quantization has been studied only recently in the context of new H.264 video compression standard [8]. In this standard scalar quantization, so-called QP factor is applied to the coefficients of DCT blocks. For the purpose of this paper the QP factor is adopted as quantization method.

Elimination of perceptually irrelevant information by quantization can be considered desirable from the pattern recognition viewpoint. There is however a question what is the impact of QP factor in this regard. The effect of quantization can be seen as reducing the variety of DCT blocks, the number of different blocks images is reduced when the QP value increases. This point is illustrated in Fig. 1 where this relation is shown for typical picture.

It can be expected that for certain range of QP values recognition based on the DCT blocks will be facilitated since the number of blocks will be reduced while relevant information will be still preserved.

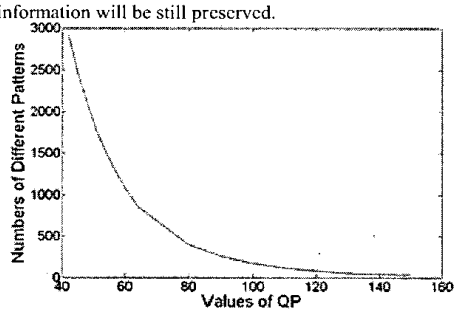


Fig.1 QP vs. Number of Patterns

3. APPLICATION TO PATTERN RECOGNITION

We will apply quantized DCT blocks to pattern recognition. The main idea is to study the effect of quantization on the recognition performance. As emphasized above increasing quantization results, up to a point, in the elimination of perceptually non-relevant information. Obviously, too strong quantization will remove relevant information too. Thus, there should be optimal level of quantization maximizing pattern recognition performance. Moreover, the overall performance should be good because of the feature-preserving properties of the DCT.

To check this reasoning we adopted simple pattern recognition scheme based on histograms [3]. For a database D of pictures D={P1....Pn} histograms of quantized DCT blocks are calculated for each picture Pi. Pattern recognition for pictures in the database is based on comparison of histograms, using sum of absolute differences measure between the histograms. Minimum absolute difference (MAD) corresponds to best recognition candidate

$$MADi = \min | Hi - Hj| \quad j=1...n \quad (4)$$

Fig. 2 shows the recognition process. Firstly, we divide the image into 4x4 blocks. Then we perform 4x4 DCT transform to them. If the source image has already been compressed into 4x4 DCT blocks, such as H.264, we can use these DCT blocks directly. The third step is quantization, aiming to limit the coefficients to a small range. After that, we can use some generic patterns to match the face images, count for each pattern to achieve a histogram, which representing the times of each pattern appearing in this image. At the registration period, the histogram of each input image is calculated at certain QP. At the searching period, we process the target image in the same way and calculate the difference of its histogram to other histograms. The image in the database, which has the lowest difference to the target one, is our final choice.

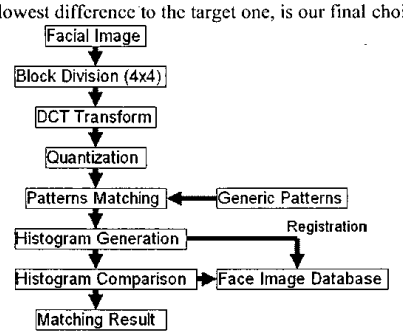


Fig.2 Face Recognition Process

5.1. 4x4 DCT blocks and quantization

In order to extract the local pattern information, we focus on the 15 AC coefficients of the 4x4 DCT block. After quantization, the 15 AC coefficients are limited to certain numbers of patterns and the total number of the patterns in a picture is dependent on the quantization factor QP. The type of source images has also impact on the type of DCT patterns. If the number of patterns is too small, then the difference of histogram between each image will also be too small for us to distinguish them; but if it is too big, the computation complexity is quite large. The overall question is about the dependence of performance on the number of patterns.

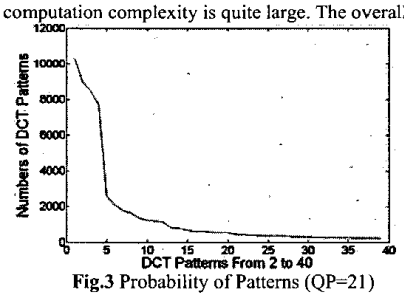


Fig.3 Probability of Patterns (QP=21)

We can generate different patterns from a number of images, but a lot of them scarcely appear in other images. Fig.3 shows the frequency of these patterns appearing in typical image. For pattern recognition a set of patterns appearing most often can be selected. This set will depend on the type of picture, quantization factor QP, size of the picture database and required recognition performance.

The optimization problem is thus given a an image database D find quantization factor QP and DCT patterns set providing best recognition performance based on the histograms of DCT blocks and minimum absolute difference measure. One can notice that both best QP and DCT pattern set may be image- dependent. The question is what ultimately achievable performance of such system is. The performance has been studied in this paper on the example of face recognition from a face database.

4. EXPERIMENTAL SYSTEM AND RESULTS

A system for testing ultimate performance for face recognition problem based on histograms of quantized DCT blocks was developed. Block DCT transformation of images are calculated and quantized. Next a set of DCT patterns is selected for each image providing histogram for which the number of images found by the MAD measure is minimized as a function of the DCT pattern set and quantization factor QP.

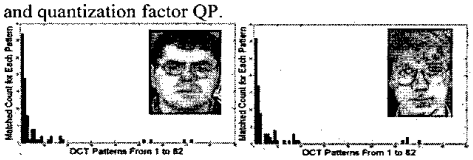


Fig.4 Histograms of Patterns For Two Persons

For experiments we used the Georgia Tech Face (GTF) Database [6]. We selected 3 images for each person, which are all front face images. Therefore, totally we got 150 images in our database. Fig.4 shows two

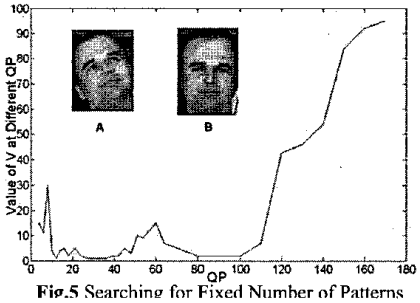


Fig.5 Searching for Fixed Number of Patterns

histogram of patterns for different persons.

If we use one target image to search among other 149 images in this database, we use a value Vi, {1≤Vi≤149, i=1, 2...149}, to evaluate the searching results. Vi =1 means that there is only one image which has the difference of histogram less or equal than the ith image. So, Vi =1 is the best result for ith image; Vi =149 is the worst result.

When the number of patterns is fixed to certain value, we can evaluate the value of Vi as a function of quantization factor QP This is shown in Fig. 5 indicating that there are local flat minima in the function. QP can be thus selected in broad range of values for good results.

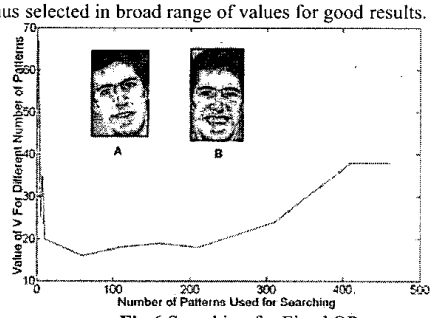


Fig.6 Searching for Fixed QP

In the second test, we fix the quantization factor QP and change the number of patterns. The result is shown in Fig. 6 indicating for smooth behavior with the existence of minimum.

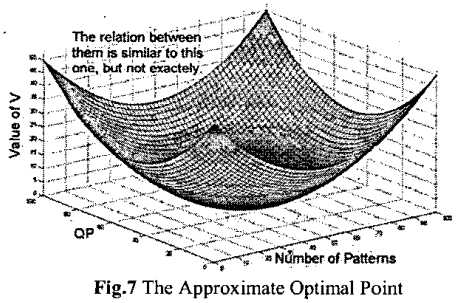


Fig.7 The Approximate Optimal Point

In the third experiment we perform the searching by varying both QP and pattern number. As shown in Fig. 7 we found that, the relation between result, QP and pattern number is as a whole a concave function. This indicates that there is global minimum resulting in very good database search result but depending on both QP and the number of patterns. For most face images, the best performance is obtained when QP is set to from 18 to 36.

For illustration, we select QP = 21, and use 30 basic patterns, with one target image A1 to search for images A2 and A3, which are all taken from the same person in the database. The result shows that we can always find the right person. It means that either A2 or A3 or both has Vi ≤ 12 (8% of 150). The best match is dependent on the detail of this target image, such as face expression, hair type, beard type and background, etc. Table.1 shows us the searching result for two persons A and B:

Table 1. Searching Result for Person A and B

	A_1-A_2	A_1-A_3	B_1-B_2	B_1-B_3
V_i	9	1	1	2

* A_1-A_2 means: Use A_1 to search for A_2.

5. CONCLUSIONS

In this paper, it is shown that quantized DCT and selection of proper pattern set results in very informative description for pattern recognition, The approach is illustrated using histograms of quantized 4x4 DCT blocks on face database images. By adjusting quantization level and DCT pattern set very good face recognition results are obtained. The method is computationally efficient and can be used directly for information extraction from compressed video. Further research is needed for complementing histograms with other measures of global pattern similarity based on histograms, and optimizing the patterns set by normalizing the bins of histograms.

11. REFERENCES

[1] Ara V. Nefian, "Embedded Bayesian Networks for Face Recognition," *in IEEE International Conference on Multimedia and Expo*, vol. 2, p. 133 -136, August 2002.

[2] F. S. Samaria. "Face Recognition using Hidden Markov Models," PhD thesis, University of Cambridge, 1994.

[3] Koji Kotani, Chen Qiu, and Tadahiro Ohmi, "Face Recognition Using Vector Quantization Histogram Method," in *International Conference on Image Processing*, II-105, Sep. 2002.

[4] Dapang Chen, and Alan C. Bovic, "Visual Pattern Image Coding," in *IEEE Trans. on Communications*, pp. 2137, Dec. 1990.

[5] Shih-Fu Chang; Messerschmitt, D.G., "Manipulation and compositing of MC-DCT compressed video", *IEEE Journal on Selected Areas in Communications*, Jan. 1995.

[6] Georgia Tech Face Database, Available: ftp://ftp.ee.gatech.edu/pub/users/hayes/facedb/.

[7] Joint Video Team(JVT) of ISO/IEC MPEG&ITU-T VCEG, "Draft ITU-T Recommendation and Final Draft International Standardd Joint Video Specification", Document JVT-G050,March 2003

[8] H. A. Peterson, H. Peng, J. H. Morgan, W. B. Pennebaker, "Quantization of color image components in the DCT domain", in *Human Vision, Visual Processing, and digital Display II, Proc. SPIE*, vol. 1453, pp. 210-222, 1991.

Publication II

DaiDi Zhong, Irek Defée, "Global Pattern Selection For Compression Histogram Database Retrieval", in Proceedings of International Workshop on Systems, Signals and Image Processing (IWSSIP 2004), pp. 239-242, September 2004.

International Workshop on Systems, Signals and Image Processing
Poznań, POLAND
September 13-15, 2004
IWSSIP'04
ambient multimedia

Global Pattern Selection For Compression Histogram Database Retrieval

Daidi.Zhong, Irek Defée

Tampere University of Technology, Institute of Signal Processing
P.O. Box 553, FIN-33101 Tampere, FINLAND.
E-mail: daidi.zhong@tut.fi irek.defee@tut.fi

Abstract – **Compression of visual information and pattern recognition are two classical and widely studied topics, each with its own foundations and core syllabus. There has been surprisingly little influence between these both areas. The present paper is a step into investigating links between them indicating for an underlying potential. We study histogram-based pattern recognition in compressed images. Relation between the compression and recognition capabilities is rev-ealed and optimization procedure suggested. Experimental results show very good performance comparing to the simplicity of the approach. This shows that compression can play important role in pattern recognition.**

Keywords: Pattern, Histogram, DCT, Retrieval

I. INTRODUCTION

Pattern recognition is nowadays a classical area with huge body of knowledge which has been collected over the years. Despite this, there are still puzzling discrepancies between the capabilities of biological systems and those running in hardware. Visual recognition is a highly overdimensioned problem which is seen easily if one would try to consider images as matrices in NxN space. Only extremely limited sets of such matrices carry useful information. The recognition should thus by necessity use highly effective preprocessing to limit the amount of input information in the first place. Evidence for such preprocessing is visible in the architecture of biological systems and also in technical systems. From the principles point of view one of the important goals of preprocessing should be elimination of any redundant information which is not indispensable for recognition. Here is the junction at which pattern recognition can meet compression. Compression has a goal of minimizing the amount of information while preserving perceptual properties. This is fully compatible and desirable for pattern recognition. Indeed one could think that elimination of perceptually redundant information should be very beneficial for the efficiency of pattern recognition process. It seems however that this point has not been fully exploited before.

This paper develops one way of approaching pattern recognition from the compression perspective and resulting benefits. Our approach is simple and does not pretend to provide full solution to the recognition complexities. It should be rather treated as a direction towards a paradigm linking both areas.

Linking compression and pattern recognition has additional interesting aspect stemming from the fact that images and video are nowadays handled in compressed formats. Developing methods operating directly on these formats is very desirable. Compression methods based on block transforms are highly efficient basis for major compression standards. Several methods were used in the past to for recognition in compressed pictures. In [1] recognition of faces based on JPEG compressed images was studied. The approach used fifteen DCT block coefficients for training two-dimensional Hidden Markov Model (HMM). In [2] object localization technique using DCT is presented. It is based on the feature vectors representing color and texture and based on the energy of quantized DCT components summed for all blocks over a region. This is complemented with shape-based matching and good overall performance is reported. A method of image retrieval from database using DCT is described in [3]. This method uses histograms of DC coefficients of the DCT and moments in the DCT domain. It is suggested there that the DCT gives good results at low computational complexity. In [4] image extraction from the database based on DC DCT coefficients is proposed and compared with the use of uncompressed images. It is shown that performance is similar but reduction in complexity is very significant. In [5] image segmentation is studied by combining PCA (Principal Component Analysis) and k-means algorithm operating on quantized

DCT coefficients with good results. In [6], [7] method is proposed for face authentication under varying illumination conditions. Combination of DCT with HMM (Hidden Markov Model) for face recognition was studied in [8].

This quick survey of literature reveals interesting point. All the approaches proposed rely on a kind of preprocessing of data provided by block transform. Subsequent recognition is done by optimization and/or learning algorithms based on principles which can be equally well applied to the non-compressed images too. However, it has been noticed, e.g. in [1] that processing in the compressed domain results in better results. The question is why it is so and to what extent the compression itself impacts the performance. Issues of this kind are addressed in this paper by a very simple model in which we try to separate the compression part from the subsequent recognition part.

This is done by investigating the performance of quantized block transform histograms in image database retrieval. Histograms use is well-known in pattern recognition, especially color histograms. In this paper we investigate histograms of quantized blocks and perform experimental tests of retrieval capabilities. Histograms are of particular interest in our context as they collect purely statistical information without any reference to structure. They just pickup global information about local features which in turn in our case have content optimized from the point of perceptual relevance. As a basis for experimental tests we use face database retrieval. On this example we show that histogram optimization can provide surprisingly good results.

II. RETRIEVAL SYSTEM

2.1. Block transform histogram

Histograms are often used for gray level or color statistics as it is easy to represent frequency of occurrence of these properties. Here we use histograms of block transforms taken over images. The DCT (Discrete Cosine Transform) is commonly used transform robustly preserving perceptual features under quantization. However, the transform block size or type is not relevant for our development. The key point for histogram application is in quantization.

Let P be an image and $S=\{S_i, 0<i<N\}$ be a set of blocks transformed. Let $\{B_j, 0<j<N\}$ be a partition of S into equivalence classes, $S_k, S_l \in B_j$ if and only if $S_k=S_l$. The block histogram is $H_p = \{|B_j|, 0<j<N\}$ where $|B_j|$ denotes cardinality of set B_j. For nonquantized images the histogram will be large and 'flat', that is there will be very many equivalence classes with low cardinality. Quantization will reduce the number of classes and even more importantly will make the histogram concentrated. The higher the quantization the more prominently this will be manifested.

We will formulate the database retrieval problem as follows For a database D of pictures $D=\{P_1 \ldots P_n\}$ histogram Hp_i of quantized DCT blocks are calculated for each picture P_i. We shall compare similarity of two pictures by using block distance measure

Pattern recognition for pictures in the database is based on comparison of histograms, using sum of absolute differences measure between the histograms. Minimum absolute difference (MAD) corresponds to best

$$| H_i - H_j| \quad j=1...n \qquad (1)$$

Then similarity matching of pictures is based on ranking the differences (1). Minimum absolute difference (MAD) corresponds to best recognition candidate.

2.2. Light intensity normalization

The retrieval approach to work practically has to compensate for overall light intensity variations. This can be simply dealt with by rescaling coefficients of block transform by average light intensity level. The average light intensity level is evaluated from DC values of block transform coefficients. For image database we normalize all pictures by taking average light intensity level of all images and rescaling DC values accordingly.

2.3. Quantization optimization

The most interesting question is what is the impact of quantization and histogram on the retrieval performance. It is evident that low quantization produces flat histograms which will not produce required ranking. On the other hand too high quantization will produce too similar histograms. There should be thus range of most suitable values of quantization providing best retrieval. This prediction can be tested experimentally in a system shown in Fig.1.

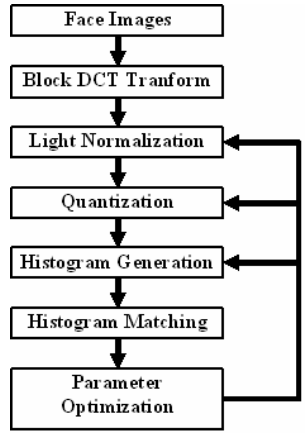


Fig. 1 System for retrieval optimization

In this system we look for the optimal quantization

value and histogram formation by iterative procedure in which the parameters are changed until best performance is achieved.

III. HISTOGRAM FORMATION

3.1 Block transforms patterns and quantization

While there can be many different types of transforms and ways of histogram formation in this paper we use a simple approach to illustrate the relation between the compression and retrieval capability. We use the 4x4 DCT transform (which can be obtained from downsampling the standard 8x8 DCT). The method of quantization in image compression standards is based on quantization matrix but for the 4x4 DCT we use quantization, similarly to the H.264 standard [9], quantization by a scalar called Quantization Factor (QF). This is sufficient for the 4x4 case. As mentioned later the DCT blocks are rescaled to compensate for different illumination levels.

3.2 Histogram formation problem

After the DCT blocks are calculated and quantized, the frequency of occurrence of blocks is computed for the formation of histograms. DCT patterns occur in an image with different probabilities. The block histogram for specific QF=2 and first 40 block patterns with highest frequency of occurrence is shown in Fig.2. It can be seen that pattern distribution has long tail. There is quite limited number of patterns which appear in large quantities and significant number of patterns which appear rarely.

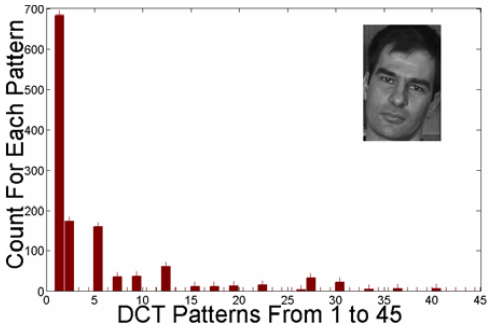


Fig. 2 Example of histogram

As there may be quite large number of different DCT blocks still left after quantization the histograms can be quite sizable. The histograms are used for the retrieval based on the MAD criterion (1). The problem to be investigated is how the histogram should be composed. This can be formulated as finding the minimum size set of frequencies which provides best retrieval performance. This will depend on the selection of Quantization Factor QF and hence optimization can be done with respect of both the histogram and QF.

3.3 Importance Normalization

From Fig.2 we can see that different bins in histogram have different importance. The patterns with high frequency of occurrence may have bigger impact on the difference measure between two images. Therefore, one can normalize patterns in the histogram by their importance. For example, given histograms Hp_A and Hp_B of picture A and B respectively, the absolute difference of

$$D(A,B) = Hp_A - Hp_B = \{|B_{jA} - B_{jB}|, 0<j<N\} = \{D_j\} \quad (2)$$

Subsequently the difference is normalized by the value of each bin.

$$D_N(A,B) = D(A,B)./Hp_B = \{D_j/B_{jB}, 0<j<N\} \quad (3)$$

Here B is the stored picture and A is the test image.

3.4 Measuring of the retrieval performance

For performance evaluation of a retrieval system, we use False Acceptance Rate (FAR) and False Reject Rate (FRR) measures [10]. Given a certain classification threshold, an input face image of person A may be falsely classified to person B. If the target person is person A, then the ratio of how many images of person A have been classified into other persons are called FRR, while the ratio of how many images of other persons have been classified into person A is called FAR. From FAR and FRR, an Equal Error Rate (EER) can be achieved where both measures achieve equal values, and this indicates well the overall retrieval performance. The lower the EER is, the better is the system's performance, as the total error rate which is the sum of the FAR and the FRR at the point of the EER decreases. Fig.3 shows the FAR, FRR and EER of example retrievals.

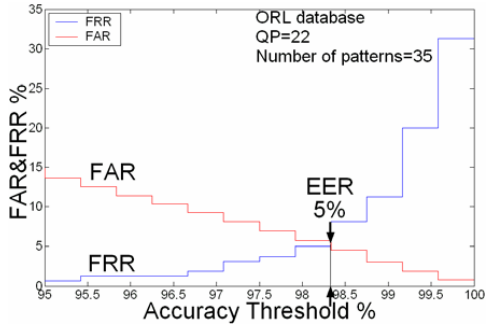


Fig. 3 Equal Error Rate

IV. EXPERIMENTAL SYSTEM AND RESULTS

Here we will describe in detail our experiments and performance achieved with histograms. For experiments we used face image databases ORL [11].

The first problem is selection of patterns for

histograms. The goal is to select minimum pattern set providing best retrieval in the equal error sense. In a basic approach the patterns are firstly ordered by their probabilities, and then a subset of them is chosen to form a histogram.

For example, one pattern set can be made of patterns 1 to 20, descending in the probability, while the second pattern set can be made of every second pattern from 1 to 39. Both sets have 20 patterns, but the EER result of the first set is 0.05, while for the second set it is 0.1125.

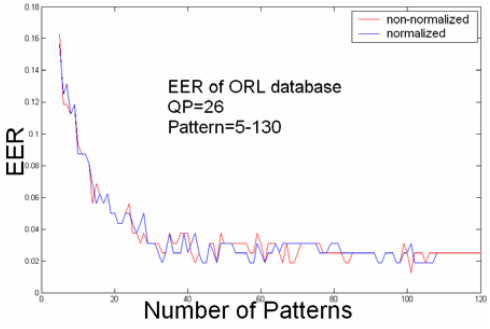


Fig. 4 Dependence of EER on the number of patterns

In addition, there is a trade-off here between the performance and the number of patterns. Small number of patterns leads to worse retrieval accuracy, but also has smaller computation complexity. Fig.4 shows the performance of different number of patterns when QF=26, it can be seen that performance stabilizes in a narrow range with more than 40 patterns.

The ORL (Olivetti Research Laboratory) database contains 10 different images of 40 distinct subjects. All the images have dark homogeneous background and the subjects are in up-right, frontal position. For experiment, we store the first 6 images of each person in the database and the remaining 4 images serve as test images. Therefore, the total number of stored images is 240 and the total number of test images is 160.

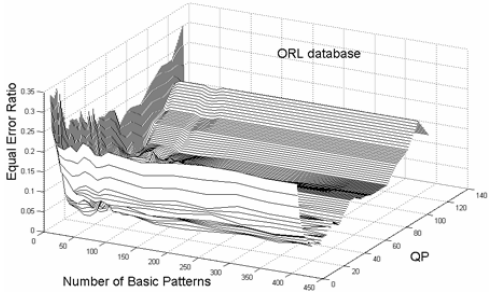


Fig. 5 Results of optimization search

In order to give find optimal retrieval parameters we performed exhaustive search among different values of quantization parameters QF, the number of basic patterns and default light intensities. In addition, the 50% DCT block overlapping is used in the calculation of histograms.

Results of the search are shown in Fig.5 as a three dimensional plot. For ORL database, we get the lowest EER of 1.25%, when the first 101 patterns are selected and QF=26. However, if we use the first 42 patterns, we will get EER=1.875%, which still a very good result, with much less computation. For comparison, in [12] EER= 2.6% was achieved.

CONCLUSIONS

We have shown that looking to pattern recognition from compression perspective results in interesting results. Our approach to the investigation of image database retrieval using histograms reduces the problem to the formation of optimal local pattern set without inclusion of structural information. The pattern set optimization is done by quantization and selection the best subset. Experimental results performed on image database were done using the 4x4 DCT blocks. Best pattern subset and quantization were searched for ORL database. The performance for optimal values if very good and close for best systems using sophisticated structural and learning approach. This shows that optimal compression of local features has very significant impact on the pattern recognition performance

REFERENCES

[1] Y. Zhong and A. K. Jain, "Object localization using color, texture and shape", *Pattern Recognition*, Vol.33, No.4, pp.671–684, Apr. 2000

[2] R. E. Frye, R. S. Ledley, "Texture discrimination using discrete cosine transformation shift-insensitive (DCTSIS) descriptors", *Pattern Recognition*, Vol.33, No.10, pp.1585-1598, October 2000

[3] S. Eickeler, S. Muller and G. Rigoll, "Recognition of JPEG Compressed Face Images Based on Statistical Methods", *Image and Vision Computing*, Vol. 18, No. 4, pp. 279-287, 2000

[4] J. Wei, "Image segmentation based on situational DCT descriptors", *Pattern Recognition Letters*, Vol. 23, No.1-3, pp. 295 - 302, January 2002

[5] J. Jiang, A. Armstrong and G. Feng "Direct content access and extraction from JPEG compressed images", *Pattern Recognition*, Vol.35, No.11, pp.2511-2519, November 2002

[6] C. Sanderson and K. K. Paliwal, "Fast Features for Face Authentication under Illumination Direction Changes", *Pattern Recognition Letters*, Vol. 24, No.14, pp.2409 - 2419, 2003.

[7] M. S. Kim, D. Kim and S. Y. Lee, "Face recognition using the embedded HMM with second-order block-specific observations", *Pattern Recognition*, Vol. 36, No. 11, pp.2723-2735, November 2003

[8] G. Feng, J. Jiang, "JPEG compressed image retrieval via statistical features", *Pattern Recognition*, Vol.36, No. 4, pp. 977-985, April 2003

[9] Joint Video Team(JVT) of ISO/IEC MPEG&ITU-T VCEG, "Draft ITU-T Recommendation and Final Draft International Standardd Joint Video Specification", Document JVT-G050,March 2003

[10] R. M. Bolle, S. Pankanti, N. K. Ratha, "Evaluation techniques for biometrics-based authentication systems (FRR)", *Proc. International Conference on Pattern Recognition*, Vol. 2, pp. 831 - 837, Sept 2000

[11] AT&T Laboratories Cambridge, "The ORL Database of Faces", Available at: *http://www.cam-orl.co.uk/facedatabase.html*.

[12] Koji Kotani, Chen Qiu, and Tadahiro Ohmi, "Face Recognition Using Vector Quantization Histogram Method," in *International Conference on Image Processing*, II-105, Sep. 2002.

Publication III

DaiDi Zhong, Irek Defée, "DCT Histogram Optimization for Image Database Retrieval", Pattern Recognition Letter, Vol. 26, Iss. 14, pp. 2272-2281, 2005

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 2272–2281
Pattern Recognition Letters
www.elsevier.com/locate/patrec

DCT histogram optimization for image database retrieval

Daidi Zhong, Irek Defée *

Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland

Received 6 May 2004; received in revised form 8 March 2005
Available online 22 June 2005

Communicated by L. Bottou

Abstract

Information extraction from images and video has been traditionally done in the pixel domain. Currently great majority of pictures and video are available in compressed form with compression based on block DCT transform. Compression removes significant amount of information leaving only perceptually important part and this has potential advantage from the information retrieval point. Optimization of compression for retrieval purposes is thus of interest but topic has not been much emphasized in the past. In this paper we study the problem of image database retrieval from the compression perspective. The approach is based on histograms of quantized DCT blocks. We show how these histograms can be optimized in order to achieve best retrieval performance by optimizing the selection of quantization factor and the number of DCT blocks under normalization of luminance. Results of experiments on face databases show that optimized histograms are robust in retrieval tasks. This indicates that selection and local feature compression optimization is an important step for effective pattern retrieval.

Keywords: Histogram; DCT; Face retrieval; Quantization

1. Introduction

Information extraction from images and video has been intensively studied in the past resulting in large variety of techniques. In recent years the technology of handling images and video has undergone significant change and great majority of content is nowadays handled in compressed form. Lossy compression based on quantized block discrete cosine transform (DCT) is a proven, highly efficient technique used in major compression standards. This technique enables reduction of the size of the content to a small fraction of the original size while preserving well its perceptual quality. Since perceptual quality is also of

* Corresponding author. Tel.: +358 400736612; fax: +358 33653087.
E-mail addresses: daidi.zhong@tut.fi (D. Zhong), irek.defee@tut.fi (I. Defée).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.04.012

major importance for pattern recognition one can conclude that compression may provide interesting perspective for studying recognition problems. By compression the pattern data size can be greatly reduced and this should be advantageous from the pattern recognition system efficiency and complexity point of view. This has not been emphasized in previous work and it is not entirely clear how to incorporate compression optimization into the pattern recognition framework. The present paper is a step in this direction considering global optimization of compression of local features and their selection for database retrieval.

Recently, there were numerous approaches based on using the DCT block processed images for information extraction from pictures. In (Zhong and Jain, 2000) recognition of faces based on JPEG compressed images was studied. The approach used fifteen DCT block coefficients for training two-dimensional Hidden Markov Model (HMM). This led to a system which is complex but has excellent recognition capabilities reaching sometimes 100%. It was also noticed (Zhong and Jain, 2000) that the same HMM trained on original pixel data is able to reach only 94.5% correct recognition rate. The improvement in using DCT is explained by its decorrelation property resulting in diagonalization of covariance matrices for the probability density function of the HMM. In (Frye and Ledley, 2000), object localization technique using DCT is presented. It is based on the feature vectors representing color and texture and based on the energy of quantized DCT components summed for all blocks over a region. This is complemented with shape-based matching and good overall performance is reported. A method of image retrieval from database using DCT is described in (Eickeler et al., 2000). This method uses histograms of DC coefficients of the DCT and moments in the DCT domain. It is suggested there that the DCT gives good results at low computational complexity. In (Ramasubramanian and Venkatesh, 2001) it has been noticed that application of DCT to pattern recognition has a problem since the block transform is not shift-invariant. Instead of using DCT coefficients it is suggested to use their sum of squares, i.e. the power spectrum. It is shown that the shift-invariant power spectrum offers better texture discrimination then the DCT coefficients itself. In (Wei, 2002) image extraction from the database based on DC DCT coefficients is proposed and compared with the use of uncompressed images. It is shown that performance is similar but reduction in complexity is very significant. In (Jiang et al., 2002) image segmentation is studied by combining principal component analysis (PCA) and k-means algorithm operating on quantized DCT coefficients with good results. A method for face authentication under varying illumination conditions is proposed in (Sanderson and Paliwal, 2003a; Kim et al., 2003). This method is based on so-called delta polynomial coefficients derived from neighboring DCT blocks. The effect of applying delta is equivalent to taking the differences between DCT coefficients in neighboring blocks which reduces impact of noise and changes in light intensity. The method shows better performance than best other methods in these respects. In (Feng and Jiang, 2003) a method of compressed image retrieval based on statistical features is presented. In (Sanderson and Paliwal, 2003b) face authentication method based on differential DCT and principal component analysis is presented and shown to be robust against changes in illumination.

From the review above it is evident that the use of DCT block transformation in combination with other techniques results in good retrieval performance. In most cases DCT can be seen as a preprocessing step followed by a more or less sophisticated method for extracting structural features, e.g. the HMM model. The advantage of using DCT can be attributed to its decorrelation properties, noise filtering and feature preservation. But it is difficult to point out the exact contribution of the DCT, either what is the maximum achievable impact, or its contribution to the overall performance in case of complex systems, although it has been shown in some cases that the DCT based processing gives significantly better results than direct pixel based processing (Zhong and Jain, 2000). There is thus general question about the informativness of the DCT processing alone, separated from subsequent processing of structural information. Another related question is what the overall system complexity is. If DCT

based features are highly informative then perhaps the overall information extraction system does not need to be highly sophisticated.

In this paper we are approaching these issues by presenting evidence that informativness of the compressed DCT features is indeed high. This is done by studying optimization of DCT information for image database retrieval purposes using compression and block DCT histograms... Histograms are relevant for this problem since they are based only on the overall statistics and no structural information about the location of features is considered. We illustrate the optimization of retrieval on face databases which is considered difficult problems in pattern recognition. The area has been widely studied in the past but our approach shows that using only globally optimized local feature statistics gives good performance even without structural information. We are thus able to show that optimization of data by perceptual lossy compression is very beneficial and should be used for designing retrieval systems.

2. Informativness of DCT coefficients

The DCT is well-known for its robust behavior under quantization. Perceptually important features are preserved even at high quantization levels, the fact which is used extensively in image and video compression. This feature of the DCT should also be advantageous from the information extraction point since removing non-relevant information could make the retrieval easier. However, the relation between the quantization of DCT and its retrieval capabilities is not clear. This is because in images, single DCT blocks may contribute minimally to the overall information content. The content is represented by local features encoded in many DCT blocks and by spatial arrangement of the blocks.

In this paper we separate both aspects of the representation by looking into the information content in the DCT blocks alone without any spatial/structural information. This is done by evaluating histograms of DCT block distribution. Histograms are relevant here since they use only frequency of occurrence of different blocks without any regard of their spatial distribution. By using the histograms it is then possible to study how much information content is carried by quantized DCT blocks at different quantization levels disregarding their spatial arrangement.

3. Block DCT and histograms

In this section we shall review basic concepts used in this paper. Block discrete cosine transform (DCT) is widely used in standard image and video compression algorithms. This transform is originally defined in the 1-D form, and can be used to construct 2-D separable transform. It has been found useful for source coding, especially image and video coding. The 2-D DCT can be calculated directly by

$$\begin{aligned} G(m,n) = a(m)a(n)\sum_{i=0}^{M-1}\sum_{k=0}^{N-1} & g(i,k) \\ & \times \cos[\pi(2i+1)m/2M] \\ & \times \cos[\pi(2k+1)n/2N] \\ & 0 \leqslant m \leqslant M-1,\ 0 \leqslant n \leqslant N-1 \end{aligned} \quad (1)$$

$$a(m) = \begin{cases} 1/\sqrt{M}, m = 0 \\ \sqrt{2/M}, 1 \leqslant m \leqslant M-1 \end{cases}$$

$$a(n) = \begin{cases} 1/\sqrt{N}, n = 0 \\ \sqrt{2/N}, 1 \leqslant n \leqslant N-1 \end{cases}$$

In image compression typically the 8×8 DCT blocks are used. When considering DCT blocks, the first uppermost DCT coefficient $G(0,0)$ in (1) is called DC and it corresponds to average light intensity level of a block, other coefficients are called AC and they correspond to components of different (cosinusoidal) frequencies.

Fig. 1 shows the average magnitudes of the 8×8 DCT coefficients for a test image. One can see that the energy of the DCT coefficients drops quickly as the frequency index increase. Unique feature of the DCT is that it is very robust against quantization, in the perceptual sense. Quantization of the DCT is done by rounding the transform coefficients $G(m,n)$. Rounding parameters in the form of quantization matrices or quantization

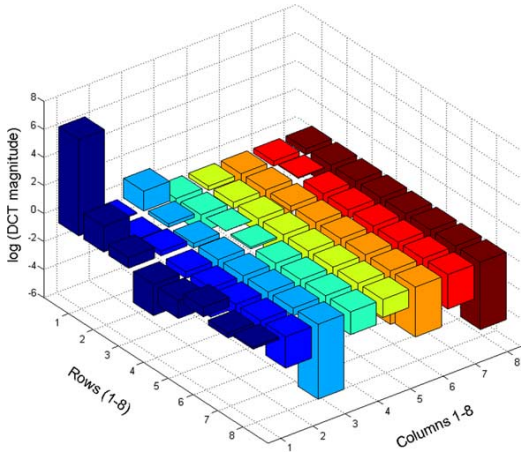


Fig. 1. Distribution of DCT coefficients for typical 8×8 DCT block.

factors (Richardson, 2003) are carefully selected in order to minimize perceptual loss. High level of quantization will result in the elimination of many frequency components from the DCT representation which, when measured in bits, minimizes the information content of DCT blocks. Typical image block can often be well-represented perceptually with a few low-frequency DCT coefficients.

In this paper, a DCT block without DC coefficient is referred as AC-Pattern. We shall use histograms of quantized AC-Patterns for image database retrieval. At low quantization, the number of different AC-Patterns in image will be very high which will make histograms broad and flattened. At very high quantization, the AC-Pattern will have only a small number of non-zero AC coefficients (in the limit, all AC coefficients will be zero). Increasing the quantization up to some level removes perceptually non-relevant information which may be impairing the retrieval process. Beyond certain level, however, further increase of quantization will impair perceptually important information.

One can thus expect that there should be a range of quantization levels which will provide best retrieval performance based on the statistics of quantized AC-Patterns. The question is how to design a retrieval system based on this idea and to evaluate its performance. To deal with this problem we formulate first the retrieval method and optimization problem. Next, experiments on several face databases are done to illustrate the procedure and evaluate the results.

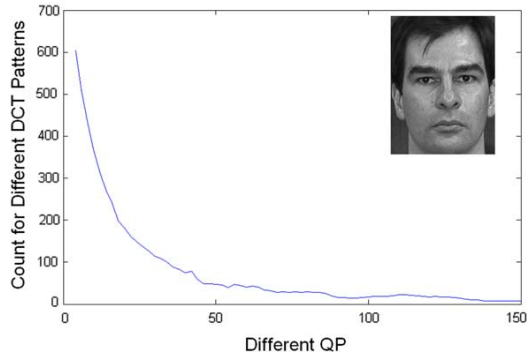


Fig. 2. Number of different AC-Pattern for different quantization levels.

We shall define the AC-Pattern histogram as spectrogram of frequencies of appearance of AC-Pattern in an image. For constructing histograms we shall use 4×4 DCT blocks quantized by single parameter called quantization parameter QP, as in the H.264 video compression standard (Joint Video Team of ITU-T and ISO/IEC JTC 1, 2003). As an illustration, for a test image of size 181×241, the number of different 4×4 AC-Pattern as a function of the quantization parameter QP is shown in Fig. 2. We can see that the total number of different blocks is quickly diminishing when the QP increases. The block histogram

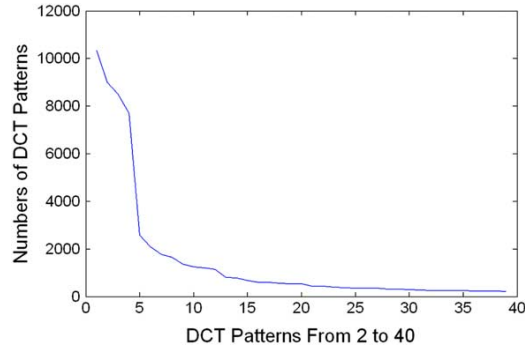


Fig. 3. Histogram of 39 AC-Patterns.

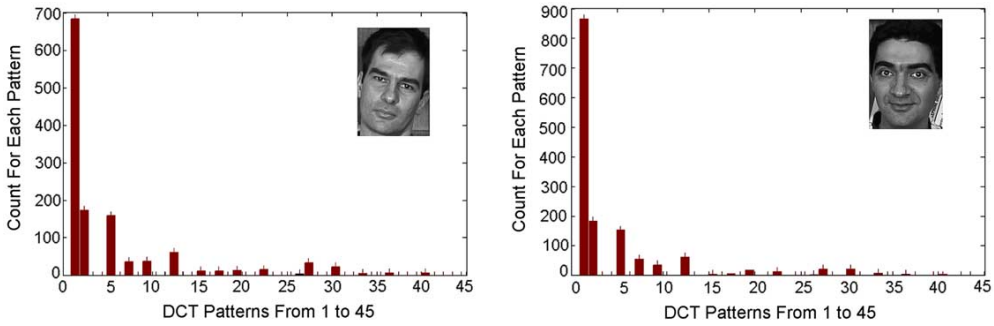


Fig. 4. Histogram of 39 AC-Patterns.

for specific QP = 22 and first 40 AC-Patterns with highest frequency of occurrence is shown in Fig. 3.

It can be seen that the distribution of AC-Patterns has long tail. There is quite limited number of patterns which appear in large quantities and significant number of patterns which appear rarely. We use histograms based on the most often appearing blocks, two typical examples of such histograms based on 45 AC-Patterns with QP = 22 are shown in Fig. 4.

We shall now consider the retrieval scheme used in this paper based on AC-Pattern histograms. First, histogram similarity is evaluated using the city block measure

$$B_{i,j} = \sum_{k=1}^{m} |H_i(k) - H_j(k)| \quad i,j \in D \tag{2}$$

where k is the bin number in the histograms of two pictures i,j belonging to the database D.

4. DC coefficients and direction vectors

The DC coefficients carry information about average luminance within the DCT blocks. This information becomes meaningful when considering how the luminance is changing between the blocks. To account for it we complement the AC-Pattern histograms with histograms of luminance differences between the DC blocks.

We consider full connectivity of blocks and calculate nine differences between the DC coefficient of the current block and its neighbors. The difference for Direction 9 is the difference between current DC and the mean of the all the nine neighboring DCs (including the current DC). The differences are ordered according to their absolute values and the first γ direction-values ($1 \leqslant \gamma \leqslant 9$) with largest differences are taken to form direction vectors, γ is a parameter which can be adjusted in the optimization process. The process of forming direction vectors is illustrated in Fig. 5(a)–(d).

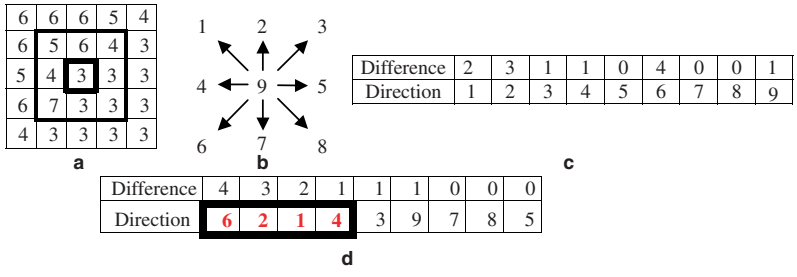


Fig. 5. Forming of a direction vector: (a) image area with DC values, (b) directions, (c) differences for directions, (d) direction vector formation.

The direction vectors for all blocks are used for calculation of DC direction vector histograms. We denote it as DC-DirecVec histogram.

Subsequently the histograms based on AC-Pattern and DC-DirecVec are combined together by the following formula

$$H = [H_{\mathrm{AC}}, \alpha \times H_{\mathrm{DC}}] \tag{3}$$

where α is a parameter controlling the relative impact of AC and DC component on the retrieval and can be adjusted in the optimization process.

5. Luminance normalization

The distribution of patterns in quantized block histograms depends on the overall luminance. Same quantization will produce different DCT blocks from a scene taken at low luminance than from the same scene at higher luminance. To eliminate this impact, we normalize the luminance of images by rescaling the DCT coefficients according to the average luminance level. The average luminance level is calculated based on the DC coefficients of the DCT blocks.

Assume there are N DCT blocks in an image j, and the DC value for each block is denoted by $\mathrm{DC}_i(j)$, $1 \leqslant i \leqslant N$. From these DC values, we can calculate the mean DC value for this image

$$\mathrm{DC}_{\mathrm{mean}}(j) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{DC}_i(j) \tag{4}$$

Next, in similar way the average luminance $\mathrm{DC}_{\mathrm{all}}$ of all images in a database is calculated based on (4). The ratio of luminance rescaling for image j is calculated through:

$$R = \frac{\mathrm{DC}_{\mathrm{all}}}{\mathrm{DC}_{\mathrm{mean}}(j)} \tag{5}$$

Next the, AC coefficients of a block are rescaled by

$$\overline{\mathrm{DCT}_{i,j}} = \mathrm{DCT}_{i,j} \times R, \quad 1 \leqslant i \leqslant N, \ 1 \leqslant j \leqslant M \tag{6}$$

After normalization, the DCT coefficients are then quantized by a quantization coefficient QP

$$\overline{\overline{\mathrm{DCT}_{i,j}}} = \frac{\overline{\mathrm{DCT}_{i,j}}}{\mathrm{QP}}, \quad 1 \leqslant i \leqslant N, \ 1 \leqslant j \leqslant M \tag{7}$$

We found that system performance is not sensitive to the exact value of rescaling so whenever images are of perceptually tolerable quality (not strongly under- or overexposed) the rescaling works well.

6. Optimization of a database retrieval system

Let image database D be given. When retrieval of pictures is based on the measure of histogram similarity (2), the minimum absolute difference (MAD) corresponds to the best retrieval candidate.

$$\mathrm{MAD}_i = \min_j (CB_{i,j}) \quad 0 < j < m + 1 \tag{8}$$

the first k smallest values of $CB_{i,j}$ for picture j will correspond to best k retrieval candidates.

The optimization problem is now formulated as follows. Given a database D find the DCT block quantization parameter, the number of DCT block bins included in the histograms, the parameters α in (3) and γ for direction vector, to obtain retrieval result with minimum errors.

With such optimization we can find the limits of retrieval based on DCT histograms. This in turn will indicate the limits of retrieval without global structural information, based only on statistics of local features with minimized information content due to quantization and size of the histograms.

The overall scheme of the system for the optimization of parameters is shown in Fig. 6. We have

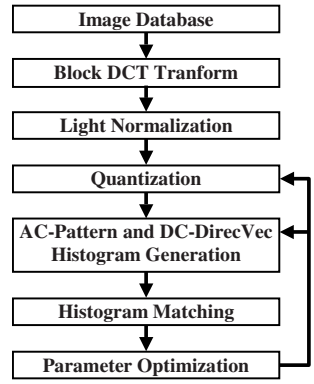


Fig. 6. Optimization of histogram parameters.

used 4×4 DCT. The optimization problem has four parameters, since only identical scalar quantization of DCT blocks by QP is used, in addition to the histogram bins size and the parameters α and γ. One should emphasize that quantization is done in a way typical for image compression that is same value of QP is used for all blocks in the picture. For the whole database, a single optimal value of QP is evaluated. This severely limits the typical QP search space.

Both the cases of non-overlapping and 50% overlapping DCT blocks were investigated though we found that in the case of overlapping blocks results are approximately two times better, so only results for overlapping blocks are presented here. The histograms are constructed from sets of AC-Pattern and DC-DirecVec grouped into bins which are ordered according to the number of patterns. Initially, the histograms are calculated for all images in the database, the values of MAD are calculated for the retrieval process and the error rates are calculated. The quantization parameter QP and number of histogram bins selected from the histograms are changed iteratively to establish their values for best retrieval results.

Computational complexity of the method is quite low. Image processing consists of calculation of block transforms which can be done efficiently using fast algorithms and quantization. AC-Pattern and DirecVec vectors formed are used in the evaluation of the vector difference measure. All these operations are fairly standard and can be implemented very efficiently using multimedia instructions available in standard processors. The optimization of parameters involves iterative calculations on the subset of picture database. The iteration steps are discrete and their number is strictly limited by the quantization, size of the histograms and vectors. In practice for a test set of 200 pictures and standard non-optimized software evaluation of optimal parameters takes several minutes on standard PC.

7. Experiments with face databases

The retrieval method proposed does not use structural information about position of features. It relies only on global quantization and formation of histograms. Information about local features and their statistics is then optimized for best retrieval. The question is what level of performance can be achieved and how it compares to other methods which use sophisticated structural information and learning. To establish the retrieval performance of the proposed system we performed experiments on several face databases. Face databases were selected since they are considered challenging for retrieval due to varying face expressions and poses. Three face databases used commonly in research community were used. The GTF (Georgia Tech Face Database), and ORL (AT&T Laboratories Cambridge), are small databases used commonly in research and the FERET (Face Database) is a large database used for very comprehensive evaluations of practical systems. Pictures in the ORL and GFT databases have small sizes about 120×90 pixels which give compelling argument for considering block overlapping.

For the evaluation of performance we use two standard methods. The first one is based false acceptance rate (FAR) and false reject rate (FRR) measures to establish the equal error rate (EER), Fig. 7 (Bolle et al., 2000). This approach reflects well the process of face retrieval when the system under evaluation provides response in the form of several candidate faces. Given a certain classification threshold, an input face image of person. A may be falsely classified to person B. If the target person is person A, then the ratio of how many images of person A have been classified into other persons is called FRR, while the ratio of how many images of other persons have been classified into person A is called FAR. From the FAR and FRR, an equal error rate (EER) is achieved when both measures take equal values. The lower the EER is, the better is the system's performance, as the total error rate which is the sum of the FAR and the FRR at the point of the EER decreases. The second method, adopted from the FERET database methodology, is based on the cumulative match scores which includes leave-one-out measure as first rank (Phillips et al., 2000).

One should emphasize that our method does not depend on training or learning but it is rather searching for optimal data processing parameters.

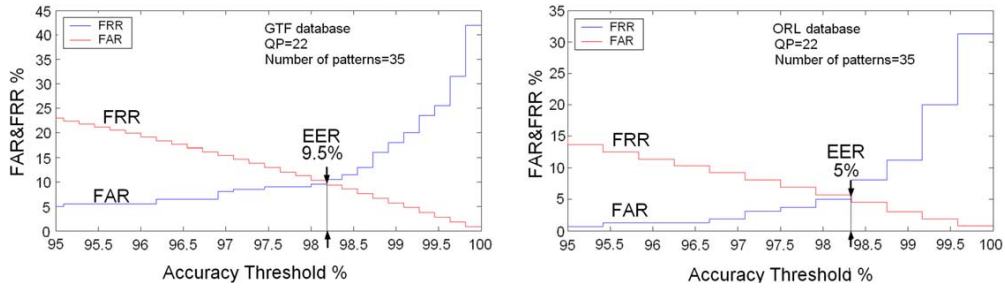


Fig. 7. Typical EER histogram performance for two face databases.

These parameters depend on the statistics of the data and can be estimated using valid subset of the data. To illustrate this point we evaluated the optimal parameters both for the complete databases and also by using cross-validation. The cross-validation was made by splitting the databases into two equal sets; one set was used for evaluating the parameters which were used for testing with the second test.

7.1. Results for the ORF and GTF databases

The Georgia Tech Face (GTF) database contains the face images of 50 people, from both male and female, each with 15 images. Most of the images were taken in two different sessions to account for the variations in illumination conditions, facial expression, and appearance. In addition to this, the faces were captured at different scales and orientations. For test, we store the first 11 images of each person in the database and the remaining 4 images serve as key images for retrieval. Therefore, the total number of stored images is 550 and the total number of key images is 200.

The Olivetti Research Laboratory (ORL) database contains 10 different images of 40 persons. For some of the persons, the images were taken at different times, with slightly varying lighting, various facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). The ORL has thus more variations for images taken from one person. All the images have dark homogeneous background and the subjects are in up-right, frontal position (with tolerance for some side movement). For experiment, we store the first 6 images of each person in the database and the remaining 4 images serve as key images. Therefore, the total number of stored images is 240 and the total number of key images is 160.

Table 1
EER results based on AC-Pattern, DC-DirecVec and combined histogram for ORL and GTF

	AC-Pattern histograms (%)	DC-DirecVec histograms (%)	Combined histogram (%)
EER-ORL	1.25	3.125	0.625
EER-GTF	7	7	4.5
EER-FERET	4.6371	7.06	3.43

To illustrate the contribution of AC-Pattern and DC-DirecVec histograms we performed separate optimizations for each of them. Next, optimization for combined histograms was done. Results are shown in the top three lines of Table 1. The best result of ORL is obtained when: QP_AC = 36, number of AC-Pattern = 80, QP_DC = 75, number of DC-DirecVec = 300 and $\alpha = 0.7$, $\gamma = 7$. The best result of GTF is obtained when: QP_AC = 10, number of AC-Pattern = 250, QP_DC = 20, number of DC-DirecVec = 400 and $\alpha = 0.9$, $\gamma = 5$.

The parameters and performance were also evaluated using cross-validation providing nearly the same results. In this case, For the ORL and GTF databases by using half of the data set for parameter estimation and the other half for retrieval we found the EER = 0% for ORL, EER = 4% for the GFT. For comparison, the EER result for the ORL reported in (Kotani et al., 2002),

where the retrieval is based on Vector Quantization Histogram Method, is 2.6%.

8. Experiments with the FERET database

ORL and GTF are relatively small databases. In order to provide some general guidance for estimating optimal parameter values, we also test our method based on a large FERET database (Phillips et al., 2000). The FERET database contains overall more than 10,000 images from more than 1000 individuals taken in largely varying circumstances. The FERET database images are divided into several sets which are formed to match its methodology of evaluation. Here we made a test based on the sets **fa** and **fb**. In both of them, each face has one picture with picture in **fb** taken seconds after the corresponding picture in **fa**. The **fa** set which has size of 994 images and serves as the database, the **fb** set which has sizes of 992 images, is used as key images for retrieval from the **fa**.

The FERET database also provides the position data of the eyes, nose and mouth for each image. According to whether this information is used, the tests can be divided into two types: *fully automatic* and *partially automatic* test. Here we consider the latter one only as we normalize location of the face based on the data of eye positions. The performance on the FERET database has been optimized by two methods. The first is the EER described above. The EER results of different measurement are shown in the bottom line of Table 1. The best EER result is obtained when: QP_AC = 12, number of AC-Pattern = 400, QP_DC = 12, number of DC-DirecVec = 400 and $\alpha = 0.5$, $\gamma = 4$.

The EER result for FERET is in line with corresponding results for ORL and GTF above which shows that our method provides consistent results for different databases and retrieval conditions.

The evaluation method of FERET presented in (Phillips et al., 2000) is based on performance statistics reported as *cumulative match scores*, which are plotted on a graph. The horizontal axis of the graph is rank and the vertical axis is the probability of identification (P_I) (or percentage of correct matches). In a similar way to the EER this lets one know how many images have to be examined to get a desired level of performance since the question is not always "is the top match correct?", but "is the correct answer in the top *n* matches?" Fig. 8 shows the *cumulative match scores* results of our method and other algorithms used in (Phillips et al., 2000). They are all partially automatic tests.

We also evaluated the optimal parameters by cross-validation experiment for the FERET database. The data set was split into two parts, denoted

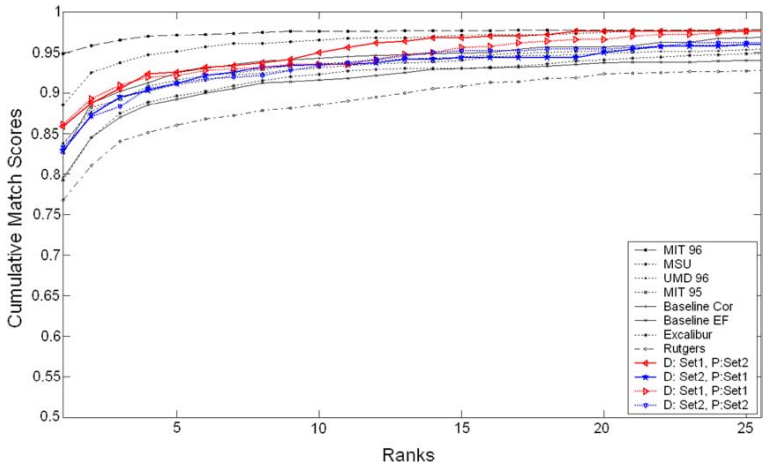


Fig. 8. Cumulative match scores results. *P*—the parameter evaluation set. *D*—the database test set used in cross-validation.

by Set1 and Set2, which were used both for parameter evaluation and testing the retrieval performance for each of them and for full FERET database. The optimal parameters evaluated from Set1, and Set2 set are the same for QP_AC (= 12), QP_DC (= 10), the number of AC-Pattern (= 400) and DC-Pattern (= 390). There differences are for α and γ parameters, for Set1, $\alpha = 1.5$, $\gamma = 3$ and for Set2, $\alpha = 2.5$, $\gamma = 4$. These differences have small impact since as can be seen from Fig. 8, the results are very similar for all cases. This indicates that optimal set of parameters is reliably evaluated.

9. Conclusion

In this paper, it is shown that image retrieval without structural information using only optimized histograms of DCT blocks results in performance which is on par with very sophisticated methods. The approach presented is based on optimizing only a few global parameters: scalar quantization, size of the histograms, combination of AC-Pattern and DC-DirecVec histograms and size of the difference vector. The method is not very sensitive to the values of optimal parameters. The parameters can be evaluated on a subset reflecting statistics of the database and used for images fitting to the statistics. In such case the image database retrieval system becomes extremely simple since optimization of parameters is done only once. The results can be contrasted with many other approaches which achieve good performance by sophisticated training and optimization. One can think that in these approaches there are two problems. First, elimination of perceptually non-relevant information is not done robustly. Second, contributions of relevant statistical and structural information are mixed which adds to the complexity and does not necessarily guarantee results superior to the presented here. Our results show that by only optimizing of perceptual information from local features and global statistical information about them one can get comparable results. This indicates that further reduction in complexity and improvement in performance of image retrieval systems might be possible by combining the optimization of statistical information with subsequent robust structural approach. This will be the topic of future research.

References

AT&T Laboratories Cambridge, ORL Database of Faces. Available at: <http://www.cam-orl.co.uk/facedatabase.html>.

Bolle, R.M., Pankanti, S., Ratha, N.K., 2000. Evaluation techniques for biometrics-based authentication systems (FRR). In: Proc. Internat. Conf. on Pattern Recognition, vol. 2, September 2000, pp. 831–837.

Eickeler, S., Muller, S., Rigoll, G., 2000. Recognition of JPEG compressed face images based on statistical methods. Image Vision Comput. 18 (4), 279–287.

Feng, G., Jiang, J., 2003. JPEG compressed image retrieval via statistical features. Pattern Recognition 36 (4), 977–985.

FERET Face Database. Available at: <http://www.itl.nist.gov/iad/humanid/feret/>.

Frye, R.E., Ledley, R.S., 2000. Texture discrimination using discrete cosine transformation shift-insensitive (DCTSIS) descriptors. Pattern Recognition 33 (10), 1585–1598.

Georgia Tech Face Database. Available at: <ftp://ftp.ee.gatech.edu/pub/users/hayes/facedb/>.

Jiang, J., Armstrong, A., Feng, G., 2002. Direct content access and extraction from JPEG compressed images. Pattern Recognition 35 (11), 2511–2519.

Joint Video Team of ITU-T and ISO/IEC JTC 1, 2003. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264|ISO/IEC 14496-10 AVC). Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050, March 2003.

Kim, M.S., Kim, D., Lee, S.Y., 2003. Face recognition using the embedded HMM with second-order block-specific observations. Pattern Recognition 36 (11), 2723–2735.

Kotani, K., Qiu, C., Ohmi, T., 2002. Face recognition using vector quantization histogram method. In: Internat. Conf. on Image Processing, II-105, September 2002.

Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S., 2000. The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Machine Intell. 22 (10).

Ramasubramanian, D., Venkatesh, Y.V., 2001. Encoding and recognition of faces based on the human visual model and DCT. Pattern Recognition 34 (12), 2447–2458.

Richardson, I.E.G., 2003. H.264 and MPEG-4 Video Compression. John Wiley & Sons, ISBN 0-470-84837-5. August 2003.

Sanderson, C., Paliwal, K.K., 2003a. Fast features for face authentication under illumination direction changes. Pattern Recognition Lett. 24 (14), 2409–2419.

Sanderson, C., Paliwal, K.K., 2003b. Features for robust face based identity verification. Signal Process. 83 (5), 931–940.

Wei, J., 2002. Image segmentation based on situational DCT descriptors. Pattern Recognition Lett. 23 (1-3), 295–302.

Zhong, Y., Jain, A.K., 2000. Object localization using color, texture and shape. Pattern Recognition 33 (4), 671–684.

Publication IV

DaiDi Zhong, Irek Defée, "Pattern Retrieval Using Optimized Compression Transform", in Proceedings of SPIE Visual Communications and Image Processing (VCIP2005), pp. 1571-1578, July 2005.

Pattern Retrieval Using Optimized Compression Transform

Daidi.Zhong, Irek.Defée*

Institute of Signal Processing, Tampere University of Technology
P.O.Box 553, FIN-33101 Tampere, Finland

ABSTRACT

Images and video are currently predominantly handled in compressed form. Block-based compression standards are by far the most widespread. It is thus important to devise information processing methods operating directly in compressed domain. In this paper we investigate this possibility on the example of simple face information processing method based on the H.264 AC Transformed blocks. We use patterns of quantized 4x4 transformed blocks for representing local picture information. These patterns at different quantization levels provide very flexible representation of picture information. By combining both the AC and DC information, we represent global information in pictures by histograms of quantized block pattern distributions. The approach is tested on FERET database of face images and it is shown that despite its simplicity provides good results in the face recognition problem.

Keywords: H.264 Transform, Quantization, Histogram, FERET, Face Image

1. INTRODUCTION

Pattern recognition is nowadays a classical area with huge body of knowledge which has been collected over the years. Despite this, there are still puzzling discrepancies between the capabilities of biological systems and those running in hardware. Visual recognition is a highly overdimensioned problem which is seen easily if one would try to consider images as matrices in NxN space. Only extremely limited sets of such matrices carry useful information. The recognition should thus by necessity use highly effective preprocessing to limit the amount of input information in the first place. Evidence for such preprocessing is visible in the architecture of biological systems and also in technical systems.

Currently great majority of pictures and video are available in compressed form with compression based on block transform. Compression has a goal of minimizing the amount of information while preserving perceptual properties and this goal is fully compatible with and desirable for pattern recognition. The problem is whether there exist how to use efficiency of compression benefiting the pattern recognition task which would lead to best retrieval results. Indeed one could think that elimination of perceptually redundant information should be very beneficial for the efficiency of pattern recognition process. It seems however that this point has not been fully exploited before.

Recently, there were numerous approaches based on using the compressed block processed images for information extraction from pictures. A method[1] of image retrieval from database using Discrete Cosine Transform (DCT) is described. This method uses histograms of DC coefficients of the DCT and moments in the DCT domain. It is suggested there that the DCT gives good results at low computational complexity. The image segmentation[2] is studied by combining PCA (Principal Component Analysis) and k-means algorithm operating on quantized DCT coefficients with good results.

In this paper, a novel recognition method based on information extracted from compressed domain is proposed. First, the 4x4 transform from H.264 standard[3] is utilized to remove the redundancy. Second, the quantization is performed to further control the precision of the information extraction. Finally, a simple combination of several the most significant block features are used to form a so called “feature histogram”. For the second step, two different quantization methods are considered separately. The example results are shown based on the well-known public face recognition database – FERET[4]. The proposed methods can achieve a good result with low computation complexity.

2. 4X4 H.264 AC TRANSFORM

The transform we used in this research is introduced from the H.264 standard. This transform is a 4x4 integer transform,

* irek.defee@tut.fi; phone +358 4 00736612; fax +358 3 3653087

which is originally aim to encode the coefficients of inter blocks. Overall, this transform performs in a similar way with the widely-used DCT. The first uppermost coefficient after transform is called DC and it corresponds to average light intensity level of a block, other coefficients are called AC and they correspond to components of different frequencies. The AC coefficients provide us some useful information about the texture detail of this block. The ability of integer calculation allows rapid process. In addition, it makes the information compact, which greatly facilitates the information extraction.

The forward transform matrix of H264 AC Transform is $\boldsymbol{B_f}$ and the inverse transform matrix is $\boldsymbol{B_i}$.

$$B_f = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \qquad B_i = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix}$$

For simplicity, here we removed the '1/2' in the matrix. The 4x4 pixel block $\boldsymbol{P}$ is forward transformed to block $\boldsymbol{H}$ using (1), and block $\boldsymbol{R}$ is subsequently reconstructed from $\boldsymbol{H}$ using (2). The 'T' means linear algebraic transpose here.

$$H = B_f \times P \times B_f^T \tag{1}$$

$$R = B_i^T \times H \times B_i \tag{2}$$

Figure 1 show the results after applying H.264 AC transform and DCT transform to one sample 4x4 block.

164	161	159	162
162	162	163	160
162	162	164	160
160	161	160	161

(a)

2583	10	-1	5
7	15	13	0
-7	0	13	-15
6	5	4	5

(b)

645.75	1.63	-0.25	0.67
1.17	1.5303	2.09	-0.07
-1.75	-0.16	3.25	-2.36
0.86	0.42	0.48	0.47

(c)

Figure1. (a) A sample 4x4 pixel block; (b) Result of 4X4 H.264 AC transform; (c) Result of 4X4 DCT transform

We perform 4x4 H.264 block transforms over more than thousand different blocks, and the results are further averaged. After applying the transform, one could see from Figure 2(a) that the main energy is distributed around the DC coefficient. Since there are big differences between the values of coefficients, here we used natural logarithm to express the data. After quantization, some non-important information can be removed while the main information is still preserved.

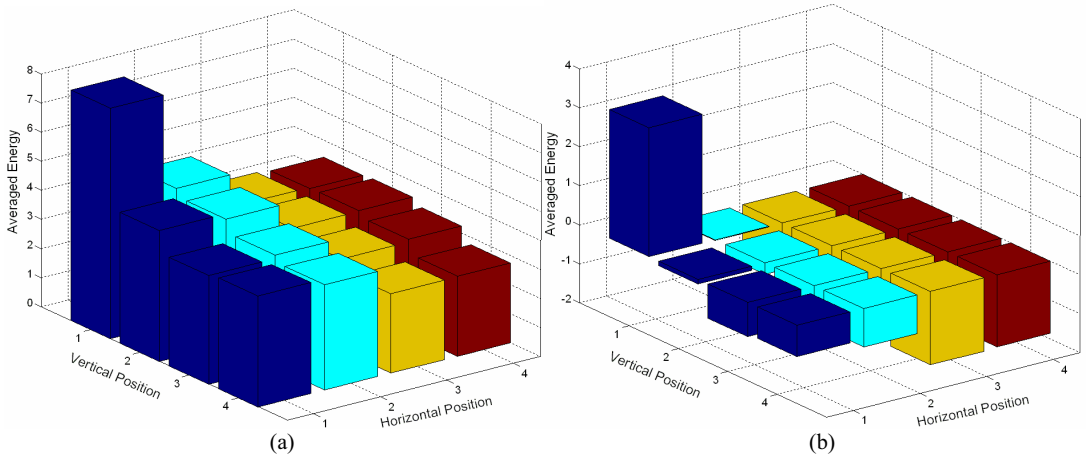


Figure2. (a) Natural logarithm of averaged distribution of energy after transform; (b) Natural logarithm of averaged distribution of energy after quantization (QP=100)

3. QUANTIZATION

After transform, quantization is used to further remove the redundant coefficients. As the energy is mostly presented at the upper-left corner, quantization can make most of the high-frequency coefficients to zero. This is shown by Figure 2(b). Therefore, if we consider the AC coefficients only, there must be some blocks in the whole image are totally the same.

Figure 3(a) shows the logarithmic distribution of different 4x4 quantized block patterns from 994 images. Among all of these different kinds of blocks, only a part of them are most often appearing in an image database. From the retrieval point of view, considering only the most common blocks can be helpful to reduce both the redundancy and computation complexity.

Furthermore, Figure 3(b) shows that the larger quantization step leads to a low number of different blocks. From here we know that quantization can be used to control the number of different blocks, in the mean while, control the amount of information remained.

Based on above observation, one may argue that the total amount of information we can extract can be represented by a function ***I(q,n)***. ***q*** is the value of Quantization Step (QP) and ***n*** is the number of quantized block pattern. In our previous work[5], we have found that by applying quantization with appropriate scale, one may reach the optimal point that most useful information is preserved while most useless information is removes. Moreover, such optimal point is not limited to a specific value, but a consecutive range of changeable values.

In this research, two kinds of quantization method are used. One is to quantize the coefficients with a uniform QP; the other is to quantize the 4x4 block with a 4x4 quantization matrix. The matrix is extracted from the quantization table originally designed for JPEG compression[6]. It took the human visual system into account, applying larger quantization to high-frequency coefficients.

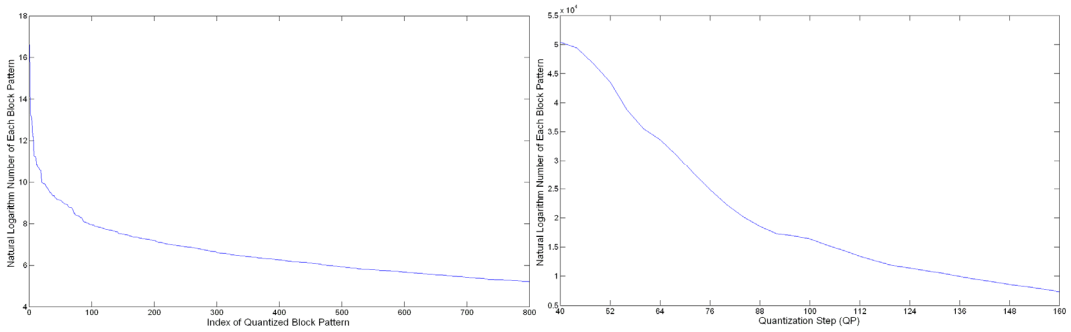


Figure3. (a) Natural logarithmic distribution of different quantized block pattern (QP=100); (b) Distribution of different quantized block pattern over different QP (Block Pattern Index = 30)

4. AC/DC VECTOR

AC and DC coefficients can provide different information of the same block. Combining the 15 AC coefficients together, we can get an AC-Vector. After quantization, the number of each different AC-Vector is changing according to the QP. It gives us a useful hint about the statistical features of the whole image.

On the other hand, the DC coefficients carry information about average luminance within the DCT blocks. This information becomes meaningful when considering how the luminance is changing between the blocks. Both Difference and Direction information can be considered.

6	6	6	5	4					
6	5	6	4	3					
5	4	3	3	3					
6	7	3	3	3					
4	3	3	3	3					

(a)

Difference	2	3	1	1	0	4	0	0	1
Direction	1	2	3	4	5	6	7	8	9

(b)

1 2 3
4 9 5
6 7 8

(c)

Difference	4	3	2	1	1	1	0	0	0
Direction	6	2	1	4	3	9	7	8	5

(d)

Figure 4. Forming of a Direction-Vector
(a) image area with DC values; (b) directions; (c) differences for directions; (d) Direction-Vector for

Since the DC value is sensitive to the light intensity and light direction. It is better to consider the neighboring DC values jointly rather than individually. Through this way, we can get another kind of knowledge about the image. According to our tests, the differences between neighboring blocks can serve as a useful feature. They are used to generate a DC Direction-Vector. According to out experiments, such vector has similar distributions over QP and Index as the AC-

Vector.

We consider full connectivity of blocks and calculate nine differences between the DC coefficient of the current block and its neighbors. The difference for Direction 9 is the difference between current DC and the mean of the all the nine neighboring DCs (including the current DC). The differences are ordered according to their absolute values and the first γ direction-values ($1 \le \gamma \le 9$) with largest differences are taken to form direction vectors, γ is a parameter which can be adjusted in the optimization process. The process of forming direction vectors is illustrated in Figure 5 a) –d). The direction vectors for all blocks are used for calculation of DC Direction-Vector histograms.

5. HISTOGRAM

Histograms are widely used in the image processing filed[7]. They are used in this research to carry the information and allow comparisons. Different from the conventional histograms based on pixels, the histograms we used here is based on quantized transformed blocks. Each bin of the histogram is representing the number of each type of feature vector in one image, which has been explained before. The basic unit here for counting is a vector, not a pixel. By calculating the Absolute Sum of Difference (SAD) between two histograms, we can measure the similarity between them. Larger SAD indicates larger difference between them.

In order to combine the AC-Vectors and DC Direction-Vector, we use a parameter α to join the two histograms like below:

$$[\text{Combined\_Histogram}] = [\ \text{Histogram\_AC} \quad \alpha \times \text{Histogram\_DC}\ ] \tag{3}$$

6. LUMINANCE NORMALIZATION

The distribution of patterns in quantized block histograms depends on the overall luminance condition. Same quantization will produce different blocks from a scene taken at low luminance than from the same scene at higher luminance. To eliminate this impact, we normalize the luminance of images by rescaling the coefficients according to the average luminance level. The average luminance level is calculated based on the DC coefficients of the transformed blocks.

Assume there are N transformed blocks in an image j, and the DC value for each block is denoted by $DC_i(j)$, $1 \le i \le N$. From these DC values, we can calculate the mean DC value for this image

$$DC_{mean}(j) = \frac{1}{N}\sum_{i=1}^{N} DC_i(j) \tag{4}$$

Next, in similar way the average luminance DC_{all} of all images in a database is calculated based on (4). The ratio of luminance rescaling for image j is calculated through:

$$R = \frac{DC_{all}}{DC_{mean}(j)} \tag{5}$$

Next the, AC coefficients of a block are rescaled by

$$\overline{DCT_{i,j}} = DCT_{i,j} \times R, \quad 1 \le i \le N, \qquad 1 \le j \le M \tag{6}$$

After normalization, all the coefficients are then quantized by a quantization coefficient QP

$$\overline{\overline{DCT_{i,j}}} = \frac{\overline{DCT_{i,j}}}{QP}, \quad 1 \le i \le N, \qquad 1 \le j \le M \tag{7}$$

We found that system performance is not sensitive to the exact value of rescaling so whenever images are of perceptually tolerable quality (not strongly under- or overexposed) the rescaling works well.

7. EXPERIMENTAL SYSTEM

To show the above ideas, a public face recognition database – FERET is used[4]. The FERET database contains overall more than 10,000 images from more than 1000 individuals taken in largely varying circumstances. The FERET database images are divided into several sets which are formed to match its methodology of evaluation. Here we made a test based on the sets **fa** and **fb**. In both of them, each face has one picture with picture in **fb** taken seconds after the corresponding picture in **fa.** The **fa** set which has size of 994 images and serves as the database, the **fb** set which has sizes of 992 images, is used as key images for retrieval from the **fa**.

All the images are transformed and quantized. For each image, the most common AC-Vectors and Direction-Vectors are extracted. According to the numbers of each vector, two histograms are generated. The basic unit here for counting is a vector, not a pixel. After this, the SAD is calculated between images, as a measurement of the similarity. For two images taken from the same person, the ideal case is that the SAD between them is zero.

The evaluation method of FERET[4] is based on performance statistics reported as *cumulative match scores*, which are plotted on a graph. The horizontal axis of the graph is rank and the vertical axis is the probability of identification (or percentage of correct matches). Each image in **fb** is used to retrieve the image from the same person in **fa**. The N results with the lowest SAD values will be returned for each retrieve. This lets one know how many images have to be examined to get a desired level of performance since the question is not always "is the top match correct?", but "is the correct answer in the top ***N*** matches?" We hope that the image from the same person in **fa** is appearing as top as possible in these ***N*** results. Figure 6 shows the *cumulative match scores* results of our method and other algorithms used in that paper. They are all partially automatic tests.

The recognition results of combined histograms are shown in Figure 6. As we can see, the JPEG Quantization method achieves a slightly better performance than Uniform Quantization method when $N \leq 17$. While when $N > 17$, their performances are quite similar.

In order to exam the sensitivity of the choices to the parameters, we also evaluated the optimal parameters by cross-validation experiment for the JPEG Quantization method. The data set was split into two parts, denoted by Set1 and Set2, which were used both for parameter evaluation and testing the retrieval performance for each of them and for full FERET database. The optimal parameters evaluated from Set1, and Set2 set are the same for QP_AC (=116), QP_DC (=40), the number of AC-Pattern (=400), DC-Pattern (=400) and γ (=3). There differences are only for α, for Set1, $\alpha=1$; and for Set2, $\alpha=3$. These differences have small impact since as can be seen from Figure 6, the results are very similar for all cases. This indicates that optimal set of parameters is reliably evaluated.

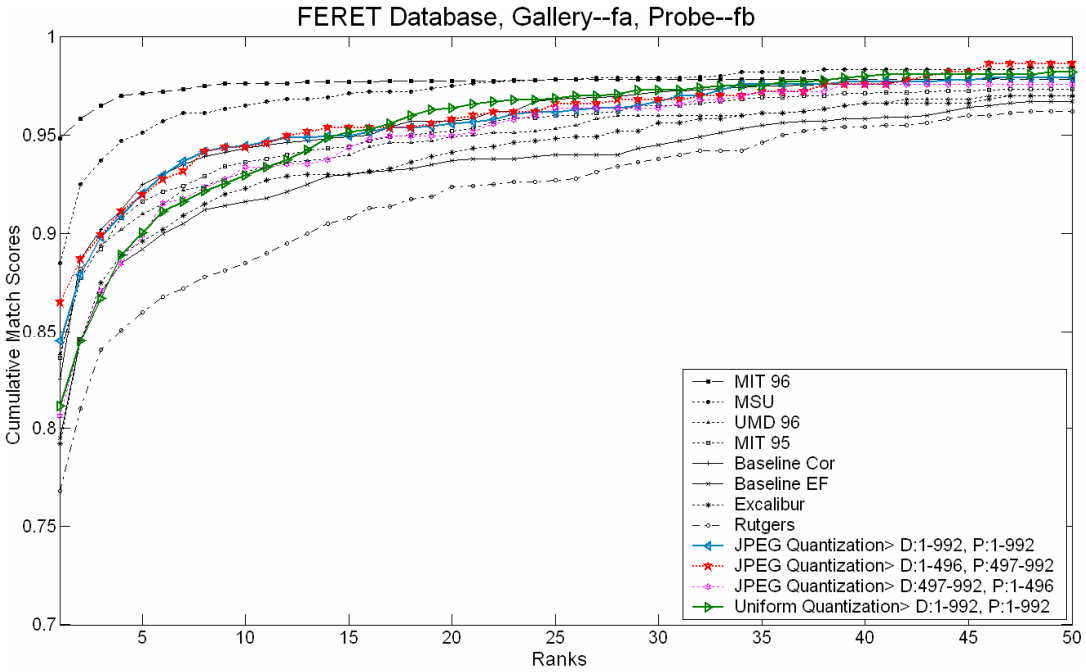


Figure 5. Cumulative match scores results of combined histograms over FERET database.
P – The parameter evaluation set. D – The database test set used in cross-validation.

8. CONCLUSIONS

In this paper, it is shown that image retrieval without structural information can be done by using optimized information from compressed domain. H.264 AC transform is used to make the useful information compact and convenient for retrieval. Quantization is subsequently used to remove the redundant information. Two different quantization methods are used separately, and their final performances are relatively similar.

The approach presented is based on optimizing only a few global parameters: scalar quantization, size of the histograms, combination of AC-Pattern and DC Direction-Vector histograms and size of the difference vector. The method is not very sensitive to the values of optimal parameters. The parameters can be evaluated on a subset reflecting statistics of the database and used for images fitting to the statistics. In such case the image database retrieval system becomes extremely simple since optimization of parameters is done only once. The results can be contrasted with many other approaches which achieve good performance by sophisticated training and optimization.

One can think that in these approaches there are two problems. First, elimination of perceptually non-relevant information is not done robustly. Second, contributions of relevant statistical and structural information are mixed which adds to the complexity and does not necessarily guarantee results superior to the presented here. Our results show that by only optimizing of perceptual information from local features and global statistical information about them one can get comparable results. This indicates that further reduction in complexity and improvement in performance of image retrieval systems might be possible by combining the optimization of statistical information with subsequent robust structural approach. This will be the topic of future research.

REFERENCES

1. S. Eickeler, S. Muller and G. Rigoll, *Recognition of JPEG Compressed Face Images Based on Statistical Methods*, Image and Vision Computing, Vol. 18, No. 4, pp. 279-287, 2000.

2. J. Jiang, A. Armstrong and G. Feng, *Direct content access and extraction from JPEG compressed images*, Pattern Recognition, Vol.35, No.11, pp.2511-2519, November 2002.
3. Joint Video Team of ITU-T and ISO/IEC JTC 1, *Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC)*, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050, March, 2003.
4. FERET Face Database, Available at: http://www.itl.nist.gov/iad/humanid/feret/.
5. Zhong. Daidi, Defée. Irek**,** *Global pattern selection for compression histogram database retrieval,* Proceedings of 11th International Workshop on Systems, Signals and Image Processing (IWSSIP'04), Ambient Multimedia, pp. 239-242, 13-15 September 2004
6. Long-Wen Chang, Ching Yang Wang, Shiuh Ming Lee, *Designing JPEG Quantization Tables Based on Human Visual System*. IEEE International Conference on Image Processing (ICIP), 1999
7. Michael. Swain, Dana. Ballard, *Color indexing*, International Journal of Computer Vision, 1991.

Publication V

DaiDi Zhong, Irek Defée, "Study of image retrieval based on feature vectors in compressed domain", Proceedings of 7th Nordic Signal Processing Symposium (NORSIG 2007), pp. 202-205, June 2006

Study of Image Retrieval Based on Feature Vectors in Compressed Domain

Daidi. Zhong[1†], *Irek. Defée*[1]

[1]Tampere University of Technology
Institute of Signal Processing
P.O. Box 553, FIN-33101 Tampere
FINLAND
[†]Tel. +358-3- 31154503, Fax: +358-3-3653087
[†]E-mail: Daidi.Zhong@tut.fi, Irek.Defee@tut.fi

ABSTRACT

An image retrieval method is proposed in this article, exploiting information of frequency components in compressed blocks. In this method images are first processed with block transform and quantization. Subsequently, the Binary Feature Vector (BFV) is formulated to represent the local visual information. Special histograms are generated next based on BFV vectors providing statistical description of distribution of BFV vectors. The BFV concept is then extended to Ternary Feature Vector (TFV). The BFV and TFV histograms are used for the image database retrieval. Three different Feature Vector schemes are proposed and the performances are investigated. Good retrieval results are obtained for standard public face image database.

1. INTRODUCTION

Image database retrieval is important for many multimedia applications. The problem of effective and robust retrieval is difficult because of complexity of image information which involves statistical and structural components. There is a question of how to describe the both components of the information and to combine them. In this paper we are concerning an efficient description of statistical information and evaluation of its retrieving capabilities. We form description of statistical information by histograms of features. Features are generated from quantized coefficients of block transform. This paper is organized as follows: In Section 2, brief introduction to block transform and quantization are given. In section 4, three Feature Vector schemes are presented. Section 5 shows the performance results for a public face image database, and Section 5 concludes the paper.

2. BLOCK TRANSFORM AND QUANTIZATION

2.1 Block Transform

The Block Transform Coding methods are widely used in image and multimedia compression. Transform coding is an integral part of one of the most widely known standards for lossy image compression, the JPEG (Joint Photographic Experts Group) standard [1]. The transform coding method makes image data compact by representing the original signal with a small number of transform coefficients. It exploits the fact that for typical images a large amount of signal energy is concentrated in a small number of coefficients. For compression purposes, this fact allows quantization to reduce the visual redundancy; however, this property can also be used in image retrieval tasks to facilitate the informational extraction.

The transform we used in this research is taken from the H.264 standard [2]. This transform is a 4x4 integer transform, which is originally aim to encode the coefficients of inter blocks. Overall, this transform performs in a similar way with the widely-used DCT. The ability of integer calculation allows rapid processing. The first uppermost coefficient after transform is called DC and it corresponds to average light intensity level of a block, other coefficients are called AC and they correspond to components of different frequencies. The AC coefficients provide us some useful information about the texture detail of this block.

The transformation and reconstruction processes are described below.

$$B_f = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \qquad B_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix}$$

$$H = B_f \times P \times B_f^T \tag{1}$$

$$R = B_i^T \times H \times B_i \tag{2}$$

2.2 Quantization

After performing the transform, quantization is used to remove the redundant coefficients. As the energy is mostly distributed around the DC coefficient, quantization can make most of the high-frequency coefficients to zero, while the relevant information is effectively preserved. The remaining non-zero AC coefficients after the transform indicate the existence of major textural information in the block area. Elimination of perceptually

irrelevant information by quantization is also very beneficial for pattern recognition problems. Our previous work [3] has shown that: by collecting statistics of some generic quantized block patterns, we can achieve good results in image retrieval applications. This paper extends the previous study to the statistics of another set of features which is extracted from neighboring quantized blocks.

3. HISTOGRAM OF FEATURE VECTORS

From the information extraction point of view, useful visual contents within image are usually distributed in multiple regions. In addition to exploiting the global visual information (e.g. color histogram), one should also extract the local statistical and structural information. Ramin et al., proposed a Non-parametric Local Transforms method in pixel-domain to measures the local image structure [4]. We will extend this idea to transformed and quantized DCT coefficients.

3.1 Binary Feature Vector

For each non-marginal 4x4 image block, there are eight blocks surrounding it. Such a 3x3 block matrix is utilized here to generate a Binary Feature Vector (BFV). Taking the DC coefficients as an example: the nine DC coefficients within this area form a 3x3 DC coefficient matrix. By measuring and thresholding the magnitude of differences between the non-central DC's and the central DC coefficient, a binary vector length 8 is formed. The central DC coefficient is thus used as a threshold to binarize neighboring coefficients. Two different cases are considered here:

Case1:
0 – current coefficient ≤ threshold
1 – current coefficient > threshold
Case2:
0 – current coefficient < threshold
1 – current coefficient ≥ threshold

Case1 is called *Inclusive* case, while the Case2 is called *Exclusive* case. One example is shown in Fig. 1 to illustrate the generation of BFV. BFV generated from DC coefficients is called DC-BFV, while those from AC coefficients are called AC-BFV. Using the BFV, the information between neighboring blocks is expressed in compact form representing the local structure. Furthermore, by counting the BFV's for whole image a BFV histogram can be generated, as example in Fig. 2.

3.2 Central BFV And Adaptive BFV

In above section, the central coefficient is used as the threshold. However, the threshold may also be the mean or median of the 3x3 block. We call the aforementioned BFV as Central BFV, and the current one as Adaptive BFV. Their performances will be compared later.

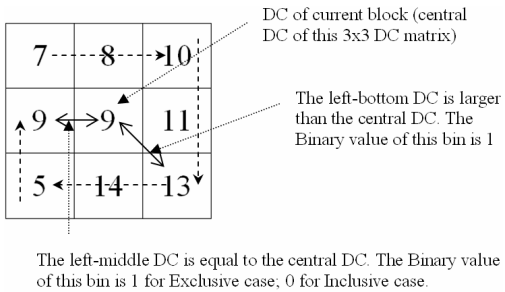


Fig. 1. Formation of BFV. Starting from the upper-left DC coefficient, the BFV is formed to be [0 0 1 1 1 1 0 0] for Inclusive case; [0 0 1 1 1 1 0 1] for Exclusive

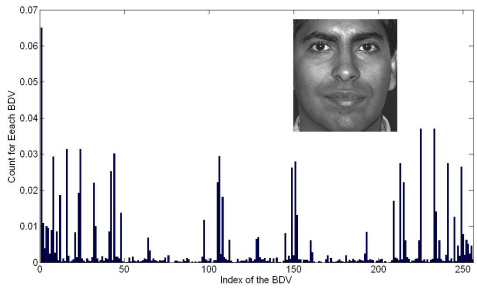


Fig. 2. Examples of DC-BFV Histogram

3.3 Ternary Feature Vector

Each BFV has bit length of 8, which gives 256 binary vectors. From our experimental results, BFV has a good ability to distinguish different visual features. However, a single threshold makes the variations unequally distributed. Certain bins may have dominant values (e.g. Fig. 2). Most BFV's are concentrated over limited amount of bins. This property is quite harmful for the retrieval ability. Therefore, we propose a Ternary Feature Vector (TFV) to allow a more spreading in DFV distributions (although not completely equally).

Similar to BFV, TFV is also calculated from a 3x3 coefficient matrix Within each matrix, assuming the max value is X, the min value is N, the mean value is M, the threshold is calculated by:

$$Threshold_{-}^{+} = M \pm (X - N) \times f \qquad (3)$$

where f is a real number in range of (0,1). In our experiment, $f = 0.4$. Again, two different cases are applied here:

Inclusive Case:
0 – current coefficient ≤ threshold$_{-}$
2 – current coefficient ≥ threshold^{+}
1 – otherwise
Exclusive Case:
0 – current coefficient < threshold$_{-}$
2 – current coefficient > threshold^{+}

1 – otherwise

One TFV with length 8 has totally 3^8=6561 variations. This seems to be a problem from complexity point of view. However, among these variations, some of them appear much more frequently than others. Therefore, only those most common ones are used in the histogram, which gives a better trade-off between the retrieval ability and complexity.

Many quantized AC coefficients are zeros, therefore, no matter which other vectors are used, there is still one dominant vector occurs in the histogram. The problem "whether should we take this bin into consideration" is studied later as *Excluding-First* mode and *Including-First* mode.

3.4 Formation of Feature Vector Histograms

The 16 coefficients in each 4x4 transform block represent different information. Our previous work [5,6] has found that, that only several AC coefficients in some specific positions are relevant for retrieval. Therefore, only some of them, rather than all of them are used to form the histograms of AC-FV, and subsequently contribute to the retrieval tasks. What we used here are the AC positions (0,1), (1,0) and (3,0). In the following experiment, the histograms of AC-FV and DC-FV are generated separately, and further combined together to give a better retrieval result. During combination, a parameter is used to weight the impact of single histogram:

$$[\text{Combined\_Histo}] = [\text{Histo\_AC} \quad \text{x Histo\_DC}] \qquad (4)$$

For AC-FV histograms alone, histograms of those three AC coefficients are combined by:

$$[\text{Histo\_AC}] = [\text{Histo1+Histo2+Histo3}] \qquad (5)$$

The retrieval task is completed by calculating the city-block distance between the histograms of different images. Smaller distance means a better match.

4. EXPERIMENTS AND RESULTS

We tested our method based on a large FERET face database [7]. We use the standard 1986 images from 994 individuals in this database, taken in largely varying circumstances. The advantage of using this database is standardized evaluation method of FERET [8] based on performance statistics reported as *cumulative match scores*, which are plotted on a graph. The horizontal axis of the graph is retrieval rank and the vertical axis is the probability of identification (PI) (or percentage of correct matches). This lets one know how many images have to be examined to get a desired level of performance since the question is not always "is the top match correct?", but "is the correct answer in the top n matches?"

4.1 Performance of BFV Histograms

We first studied the impact of the *Inclusive* and *Exclusive* case mentioned in Section 3.1, as well as the *Excluding-First* mode and *Including-First* mode mentioned in Section 3.3. Fig. 3 shows the cumulative match scores results of Adaptive AC-BFV and Adaptive DC-BFV histograms in different modes and cases. One can find out that different modes and cases have less impact on DC-BFV than AC-BFV. For AC-BFV, the *Excluding-First* mode clearly performs better than *Including-First* mode. In AC-BFV and DC-BFV, the performances of *Inclusive* and *Exclusive* case are very similar.

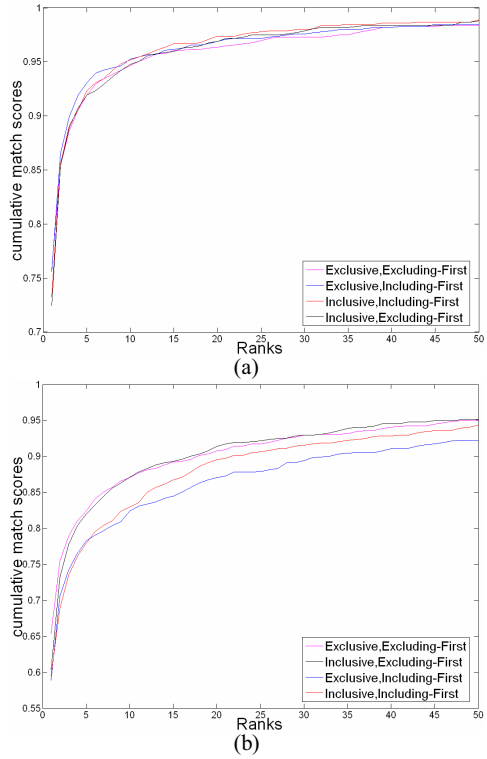
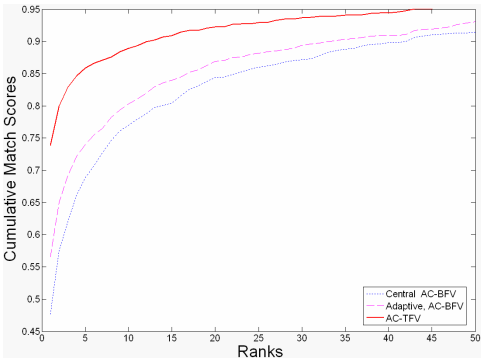


Fig. 3. Cumulative match scores results over FERET database. (a) Results of Adaptive DC-BFV. (b) Results of Adaptive AC-BFV.

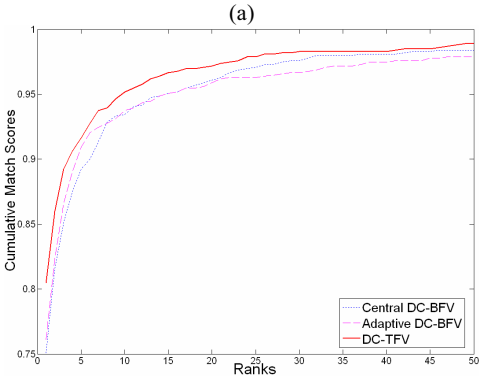
4.2 Performance of AC-FV and DC-FV Histograms

Secondly, we studied the performance of AC-FV and DC-FV histograms (Fig. 4). Clearly, one may find that:

1. Generally, DC-FV gives better results than AC-FV.
2. Adaptive BFV gives results than Central BFV.
3. TFV gives results than BFV.
4. The differences of overall performance are smaller in DC-FV than AV-FV.



(a)



(b)

Fig. 4. Cumulative match scores results over FERET database. (a) Results of DC-FV. (b) Results of AC-FV.

4.3 Performance of Combined Histograms

Finally, the performances of combined histograms are studied (Fig. 5). Unsurprisingly, the combination of AC-TFV and DC-BFV gives better result than above. Fig. 6 shows some comparative results from [8].

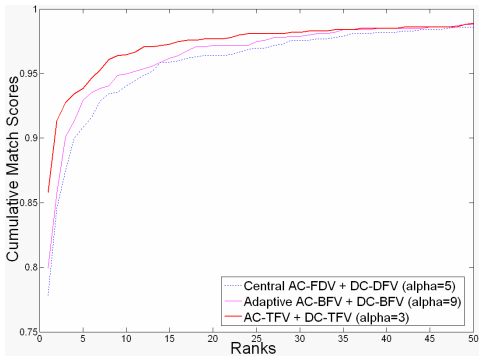


Fig. 5. Cumulative match scores results of combined schemes over FERET database.

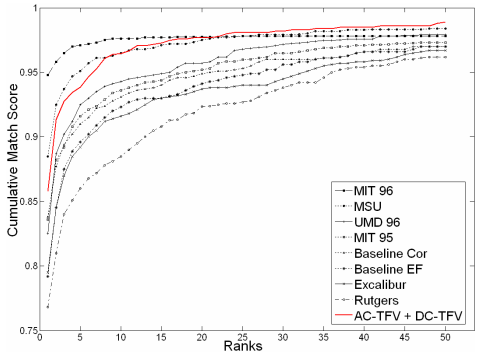


Fig. 6. Comparative results over FERET database.

ACKNOWLEDGMENT

The authors would like to thank for the financial support from Nokia Foundation Scholarship and Tampere Graduate School in Information Science and Engineering (TISE).

REFERENCES

[1] W. B. Pennebaker, J. L. Mitchell, JPEG still image compression standard, New York, Van Nostrand Reinhold 1993.

[2] Joint Video Team of ITU-T and ISO/IEC JTC 1, Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC), Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050, March, 2003.

[3] Daidi Zhong, Defée Irek, Global pattern selection for compression histogram database retrieval, Proceedings of IWSSIP'04, pp. 239-242, 2004.

[4] Zabih Ramin, Woodfill Johns, Non-parametric Local Transforms for Computing Visual Correspondence, Proceedings of the Third European Conference on Computer Vision, Stockholm, May 1994.

[5] Daidi Zhong, Defée Irek, Facial features detection by coefficient distribution map, Proceedings of 11th International Conference on Computer Analysis of Images and Patterns, CAIP 2005, Versailles, France, 2005

[6] Daidi. Zhong, Defée. Irek, Location detection of face features by DCT coefficients, Proceedings of the Fifth IASTED International Conference Visualization, Imaging, and Image Processing, Benidorm, Spain, 2005

[7] FERET Face Database, Available at: http://www.itl.nist.gov/iad/humanid/feret/.

[8] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, The FERET evaluation methodology for face recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, October 2000.

Publication VI

DaiDi Zhong, Irek Defée, "Performance of Similarity Measures Based on Histograms of Local Image Feature Vectors", Pattern Recognition Letter, Volume 28, Issue 15, pp. 2003-2010, 2007

Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 2003–2010
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Performance of similarity measures based on histograms of local image feature vectors

Daidi Zhong *, Irek Defée

Tampere University of Technology, Department of Information Technology, TF 314 Tietotalo, FIN-33101 Tampere, Finland

Received 8 January 2007; received in revised form 18 April 2007
Available online 12 June 2007

Communicated by M.-J. Li

Abstract

We investigate similarity measures for image retrieval from databases based on histograms of local feature vectors. The feature vectors are obtained from grouping quantized block transforms coefficients and thresholding. After preliminaries on block transforms we are introducing binary DC and AC feature vectors. Subsequently ternary DC and AC vectors are defined. Next we show how the histograms of vectors defined can be combined to form similarity measure for image retrieval from database. We formulate the database training and retrieval problem using the defined similarity measures. Performance results are shown using widely used FERET and ORL databases and the cumulative match score evaluation. We show that despite simplicity the proposed measures provide results which are on par with best results using other methods. This indicates that statistics based retrieval should not be underestimated comparing to structural methods.

Keywords: Pattern recognition; Face image retrieval; Block transform; Feature vector

1. Introduction

We have indicated previously that histograms of suitably formed local image features are quite efficient for the statistics based retrieval problems (Zhong and Defée, 2005). In this paper we introduce new feature vectors and measures of image similarity based on histograms and show that their retrieval efficiency is on par with best methods developed in other work.

Our approach to feature representation is related to the local binary pattern (LBP) (Ojala et al., 1996) and texture spectrum unit methods (He and Wang, 1990). There are two basic differences in this paper to these previous approaches. We use different thresholding schemes and, crucially, we do not apply them directly in the image pixel domain but to the quantized block transform coefficients which have been widely used in the image and video compression area. Block transforms have significant advantage in preserving perceptual information even under strong quantization which eliminates information non-relevant for retrieval. We define binary and ternary feature vectors based on the DC and AC quantized transform coefficients. These vectors are used for the formation of feature histograms. The histograms are describing statistical content of images using defined feature vectors. Image histograms themselves can be treated as vectors and their absolute differences is used as similarity measure. Our proposed similarity measures are based on combination of histograms for different feature vectors. We apply then these similarity measures to the image database retrieval problem and search for the similarity measure with the best performance.

While our method is general, the testing and performance evaluation is done with face image retrieval tasks. Face recognition and retrieval has been the topic of many

* Corresponding author. Tel.: +358 0 409655304; fax: +358 0 33653087.
E-mail address: daidi.zhong@ieee.org (D. Zhong).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.05.019

studies, recent overview of face recognition area can be found in (Zhao et al., 2003). We also present results in the standardized framework developed for the FERET face database (FERET, 2003) with its performance evaluation based on the cumulative match score (CMS) (Phillips et al., 2000). First the similarity measure coefficients are tuned using limited training database face sets and then evaluation of performance is conducted using large database test set. Results are presented for the FERET and also for ORL database (ORL, 2005), and compared with other methods. It is shown that the CMS performance of our method can compete with other methods which are based on much more sophisticated approaches.

2. Block transforms and quantization

We have elaborated previously on the usefulness of block transforms and quantization for pattern recognition and image retrieval (Zhong and Defée, 2005), here we provide brief summary for completeness. Block transforms methods are widely used in image and multimedia compression because of their robust preservation of perceptual information even under strong quantization. Block transform strongly eliminates the perceptually non-relevant information and this should be of advantage for the pattern recognition and image retrieval tasks too. The specific block transform we use was introduced in the H.264 standard (JVT of ITU-T, 2003) as particularly effective and simple. The transform matrix of the transform is denoted as $\boldsymbol{B}_f$ and the inverse transform matrix is denoted as $\boldsymbol{B}_i$. They are defined as

$$\boldsymbol{B}_f = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \quad \boldsymbol{B}_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix}. \tag{1}$$

A 4×4 image pixel block $\boldsymbol{P}$ can be forward transformed to block $\boldsymbol{H}$ using (2), and the quantization process $Q[\,]$ is used to remove irrelevant information, which will result to a quantized version of $\boldsymbol{H}$, $Q[\boldsymbol{H}]$. During the reconstruction process, the inverse quantization process $Q^{-1}[\,]$ is applied to the quantized block $Q[\boldsymbol{H}]$, and block $\boldsymbol{R}$ is subsequently reconstructed from the inverse-quantized block $Q^{-1}[Q[\boldsymbol{H}]]$, using (3):

$$\boldsymbol{H} = \boldsymbol{B}_f \times \boldsymbol{P} \times \boldsymbol{B}_f^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{R} = \boldsymbol{B}_i^{\mathrm{T}} \times Q^{-1}[Q[\boldsymbol{H}]] \times \boldsymbol{B}_i, \tag{3}$$

with superscript T denoting transposition.

0	**1**	**2**	**3**
4	**5**	**6**	**7**
8	**9**	**10**	**11**
12	**13**	**14**	**15**

Fig. 1. The 16 coefficients in one 4×4 transformed block. The AC coefficients are numbered.

The leading element of the matrix $\boldsymbol{H}$ is called the DC coefficient (0th in Fig. 1). All other elements are called AC coefficients. There are thus 15 AC coefficients in the matrix $\boldsymbol{H}$ but many of them will have zero value after the quantization $Q[\boldsymbol{H}]$ is applied. The power of the transform stems from the fact that despite of strong quantization, the reconstructed block $\boldsymbol{R}$ will still approximate well the original image block $\boldsymbol{P}$ (JVT of ITU-T, 2003). Quantization has the effect of limiting the number of different blocks in an image. The quantized blocks with AC coefficients can then be used for image retrieval as described in (Zhong and Defée, 2005) where they were called AC block patterns (ACBP).

3. Feature vectors

We now define new type of feature vectors (FV) based on quantized block transform coefficients. The block transform (2) can be performed for non-overlapping image blocks or for partially overlapping image blocks. In either case we can group separately DC and AC coefficients for all transform blocks of an image into matrices. From these matrices feature vectors are formed as described below.

3.1. DC binary feature vectors

From the matrix of DC coefficients, 3×3 submatrices of neighboring DC coefficients are formed. The eight DC coefficients surrounding the center one are thresholded to form a binary vector with length eight:

If the DC value $<$ threshold then put 0.

If the DC value $\geqslant$ threshold then put 1.

The value of the threshold can defined in many ways, e.g., as the value of the central coefficient in the 3×3 submatrix or as the mean value of all its nine coefficients. The thresholding approach to forming the feature vector is actually the essence of the linear binary pattern (LBP) method (Ojala et al., 1996). This method operates directly on the image pixels and uses the central coefficient value which may make it sensitive to noise. We found that in our framework using transform coefficients the mean threshold value is better and we use it in our experimental scenario. The binary feature vector formed by thresholding the DC coefficient is denoted as DC-BFV.

3.2. AC binary feature vectors

Following the procedure described for the formation of the DC-BFV above, the binary feature vectors are defined

for the AC coefficients in the same way by forming 3×3 matrices and thresholding. We shall denote such vectors as AC-BFV. Any of the AC coefficients shown in Fig. 1 can be used for constructing AC vectors but many of them will have little impact after quantization. Experiments reported later (Section 5.3) have shown that three coefficients 1st, 4th and 12th have significantly more impact than others and they are used for constructing the AC-BFV vectors.

3.3. Ternary feature vectors

In (He and Wang, 1990) a texture spectrum unit has been proposed. This corresponds to an extension of binary feature vector described above to a feature vector with ternary values which are obtained by the following thresholding:

If the DC value $<$ threshold put 0.
If the DC value $=$ threshold put 1.
If the DC value $>$ threshold put 2

and the same rules for the AC coefficients.

Here we propose the ternary feature vector (TFV) based on flexible threshold range instead of a single value of the threshold above. For a 3×3 matrix of DC or AC coefficients, the threshold range is defined as

$$T_{-}^{+} = \boldsymbol{M} \pm (\boldsymbol{X} - \boldsymbol{N}) \times \boldsymbol{f}, \tag{4}$$

where $\boldsymbol{f}$ is real number from the interval $(0, 0.5)$, $\boldsymbol{X}$ and $\boldsymbol{N}$ are maximum and minimum values in the coefficient matrix, and $\boldsymbol{M}$ is the mean value of the coefficients. The thresholded values are either 0, 1 or 2

If the pixel value $\leqslant T_{-}$ put 0.
If the pixel value $\geqslant T^{+}$ put 2.
otherwise put 1.

The resulting thresholded vectors of length eight are subsequently converted to decimal numbers in the range of $[0, 6560]$, where $6560 = 3^8 - 1$.

4. Histograms of feature vectors and similarity measure

4.1. Histograms of feature vectors

Statistical distribution of specific feature vectors in images can be represented in normalized way by histograms. Examples of such histograms are shown in Fig. 2.

The histograms can be also seen as 1-D vectors and similarity measure of images is defined as a difference between the histogram vectors using selected similarity measures. This measure counts only statistics of feature distribution and disregards structural information about location of features. Feature vectors constructed as above from the quantized block transform coefficients can be used for the formation of histograms. Since the number of such vectors is limited due to thresholding, the resulting histograms will also have limited number of bins. Moreover, it is not necessary to use all feature vectors for the formation of histograms but only those which are contributing to the overall performance. For example, in real images, after block transform and quantization, flat areas will dominate but will not contribute significant information. Feature vectors corresponding to flat areas (AC coefficients equal to zero) can be then skipped.

4.2. Similarity measures used

Feature vector histograms can be compared using various similarity measures (Delac et al., 2006; Yossi et al., 2001). In this paper we use several well-known measures based on the L_1 norm distance, L_2 norm distance, Standard Euclidean (SE) distance, Cosine (Cos) distance and Correlation (Cor) distance. For two histograms $\boldsymbol{H}_i(b)$ and $\boldsymbol{H}_j(b)$, with bins numbered as $b = 1, 2, \ldots, L$, the similarity measures are defined as follows. For the L_1 and L_2 based measures:

$$L_1(i,j) = \sum_{b=1}^{L} |\boldsymbol{H}_i(b) - \boldsymbol{H}_j(b)|, \tag{5}$$

$$L_2(i,j) = \sum_{b=1}^{L} (\boldsymbol{H}_i(b) - \boldsymbol{H}_j(b))^2. \tag{6}$$

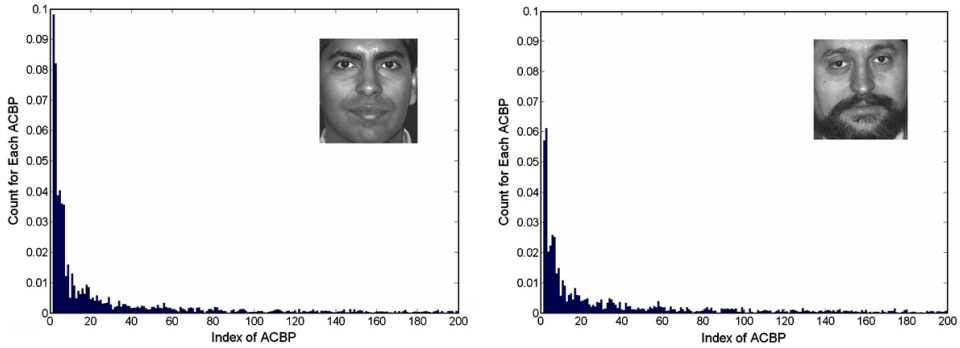


Fig. 2. Two example histograms made of quantized AC block patterns (ACBP).

Standard Euclidean measure is defined as:

$$\mathrm{SE}(i,j) = \sum_{b=1}^{L} \left(\frac{\boldsymbol{H}_i(b) - \boldsymbol{H}_j(b)}{\sigma(b)} \right)^2, \tag{7}$$

with $\sigma(b)$ denoting the variance of all histograms at bth bin.

The Cosine distance measure is defined as

$$\mathrm{Cos}(i,j) = \sum_{b=1}^{L} \frac{\boldsymbol{H}_i(b)\boldsymbol{H}_j(b)}{\sqrt{\|\boldsymbol{H}_i\|\|\boldsymbol{H}_j\|}}, \tag{8}$$

where $\|\boldsymbol{H}_i\|$ and $\|\boldsymbol{H}_j\|$ are histogram length in the L_2 norm.

Finally, the correlation distance is defined as

$$\mathrm{Cor}(i,j) = \sum_{b=1}^{L} \frac{(\boldsymbol{H}_i(b) - \overline{\boldsymbol{H}_i})(\boldsymbol{H}_j(b) - \overline{\boldsymbol{H}_j})}{\sqrt{\|\boldsymbol{H}_i - \overline{\boldsymbol{H}_i}\|\|\boldsymbol{H}_j - \overline{\boldsymbol{H}_j}\|}}, \tag{9}$$

The $\overline{\boldsymbol{H}_i}$ and $\overline{\boldsymbol{H}_j}$ in (9) represent the mean of $\boldsymbol{H}_i$ and $\boldsymbol{H}_j$, respectively.

4.3. Combination of histograms

We have formed now three types of feature histograms based on the ACBP, DC Feature Vectors (DC-FV) and AC Feature Vectors (AC-FV), respectively.

These histograms can also be combined in different ways starting with any two of the histograms as follows:

$$[\text{Combined\_2Histogram}] = [\text{Histogram\_A } \boldsymbol{\alpha} \times \text{Histogram\_B}], \tag{10}$$

where Histogram_A and Histogram_B can be any one of the histograms with ACBP, DC-FV or AC-FV. The combined histogram is thus a vector formed by concatenating the histograms A and B with coefficient $\boldsymbol{\alpha}$ introduced for controlling the relative weight of A and B in the combination. In the same way we can combine the three histograms, with two weight parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$[\text{Combined\_3Histogram}] = [\text{Histogram\_ACBP } \boldsymbol{\alpha} \times \text{Histogram\_AC } \boldsymbol{\beta} \times \text{Histogram\_DC}]. \tag{11}$$

For AC-FV histograms, we use three AC coefficients with most impact (1st, 4th and 12th, in Table 1). They are combined in the following way:

$$[\text{Histogram\_AC}] = [\text{Histogram\_coef1 Histogram\_coef2 Histogram\_coef3}]. \tag{12}$$

All the combined histogram vectors (10)–(12) can be used with any of the similarity measures (5)–(9). The values of coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are established within the framework of database retrieval system training described next.

4.4. Image database retrieval system training

Our image database retrieval problem is formulated as follows. Given an image database set $\boldsymbol{S}=\{\boldsymbol{I}_1,\ldots,\boldsymbol{I}_n\}$ we would like to establish for certain query image $\boldsymbol{I}$ if there are images similar to it in the database. For this we use the feature vector histograms of images and similarity measure defined above to find its minimum values for the image $\boldsymbol{I}$ and images from the database $\boldsymbol{S}$.

However, before this can be done the parameters used for the calculation of histograms and similarity measure need to be found using training database set. This set can be selected as a small subset of the database $\boldsymbol{S}$ and some query images selected from the subset and outside it. Knowing the correct responses for the training database allows us to tune the parameters to achieve best retrieval results. The optimal parameter set which is going to be found out during training process includes: the Quantization Scalar for (QS) for the quantization of transform blocks $Q[\boldsymbol{H}]$ in (3), coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in (10) and (11) and the parameters $\boldsymbol{f}$ in (4). The optimal parameter set is identified as the one which is maximizing the retrieval performance over training database. The resulted optimal parameter set is applied to the test database $\boldsymbol{S}$ to evaluate the actual system performance.

5. Results of retrieval performance

5.1. Image database

To train and test retrieval performance of the proposed system we used two databases of face images, FERET and ORL. The FERET database is of special importance since it contains more than 10,000 images from more than 1000 individuals taken in largely varying circumstances. Among them, the standardized FA and FB image sets are used here. FA set contains 994 images from 994 different objects, FB contains 992 images. FA serves as the gallery set, while FB serves as the probe set. National Institute of Standard and Technology (NIST) have published several releases of FERET database. The release which we used in the testing is the one published at October 2003, called Color FERET Database, this is important to be noticed since many reference publications are based on other FERET releases (e.g., 2001) so the results are not fully comparable. The advantage of using FERET database, apart of its size, is standardized evaluation method based on performance statistics reported as cumulative match scores (CMS),

Table 1
Three AC coefficients are selected based on their Rank-1 CMS performances

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1 + 4 + 12
Rank-1 CMS (%)	58.8	30.3	18.9	58.7	29.5	29.9	11.3	31.1	11.4	7.1	7.7	58.4	15.5	15.5	11.5	73.3

which are plotted on a graph (Phillips et al., 2000). Horizontal axis of the graph is the retrieval rank and vertical axis is the probability of identification (PI) (or percentage of correct matches). This lets one know how many images have to be examined to get a desired level of performance. For simplicity, many researchers use the CMS at the first rank to represent the recognition rate. We refer to it as "Rank-1 CMS".

The FERET database provides some tools for preprocessing of the face images. The images can be cropped by the provided geometrical information about eye, nose and mouth. They can be subsequently aligned, and adjusted by illumination normalization. However, we did not apply such complex preprocessing. In our evaluation, the images were simply cropped to certain sizes, which roughly contain the face area (e.g., Fig. 2). They were not aligned to the same size, and no rotation and masking are applied to them. All the images have different sizes, but these sizes are not dramatically different from each other. The differences of sizes between them are up to 30%. One should emphasize though that such techniques were used by other researchers for the presentation of the performance of their methods.

The second face database we used is ORL (Olivetti Research Laboratory). This database is much smaller since it contains 10 different images for 40 distinct subjects but it has been widely used in research. In the ORL, for some of the subjects, the images were taken at different times, with slightly varying lighting, various facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). All the images have dark homogeneous background and the subjects are in up-right, frontal position. We used ORL for additional checking of performance obtained with FERET.

5.2. Training process over FERET database

For the training of system coefficients we used first the FERET database. Here the problem is selection of the training set which on one hand should be small but on the other hand the results should not depend on particular selection of the set. To get insight into this problem we used for the training five tests sets, each composed of 50 images selected randomly. The FERET database contains only one image for each person in gallery set FA, and one image from the same person in probe set FB. Therefore, the training process is conducted between different image pairs. The "size of training set is 50 images" means 50 pairs of image from FA and FB are used and evaluated. For each of these test sets optimal parameters described in Section 4.4 were found. In result we established that coefficients found for each set are very similar and the system performance is not sensitive to the selection of the test set.

5.3. Selection of similarity measures and best AC coefficients

The first experiment aims for the selection of best AC coefficients and similarity measure for our further experiments. The selection of AC coefficients is based on their CMS scores. The training and retrieval of each coefficient are conducted independently. They are obtained from the test over FERET database using TFV and L_1 norm. Table 1 shows the Rank-1 CMS scores of these 15 coefficients. As one can see, three AC coefficients, 1, 4, 12

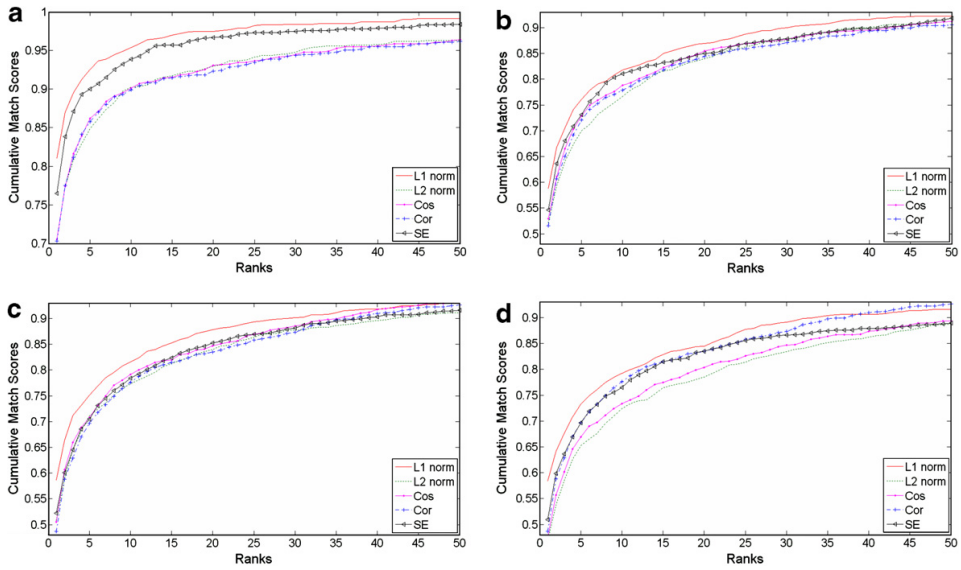


Fig. 3. The CMS of TFV for five different similarity measures. (a) 0th coefficient. (b) 1st coefficient. (c) 4th coefficient. (d) 12th coefficient. (Using the index from Fig. 1.)

provide much higher scores than the others. More coefficients may be used, but the corresponding additional cost may not be worthwhile. We thus use these combined AC histogram based on these three coefficients. Their corresponding CMS performances of using different similarity measures are shown in Fig. 3.

Results of the CMS scores will depend on the similarity measure used. In Fig. 3 the CMS performance is shown for the DC, and the histograms of three AC coefficients for the five similarity measures (5)–(9). The optimization and training process of TFV histograms are performed independently for each measure. One can conclude that L_1 norm gives overall best performance. The possible reason is that the nature of L_1 norm fit well with the intrinsic property of the optimization method in which histograms bins are ordered and their length adjusted to provide best performance. The L_1 norm also has lower computational complexity and good accuracy. Therefore, we adopt the L_1 norm for the main part of our system.

5.4. Performance of BFV histograms using FERET

Here we study the CMS performance of BFV histograms using the FERET database. Fig. 4 shows the CMS curves for AC-BFV and DC-BFV histograms using L_1 norm. From the result, one can find out that DC-BFV performs better than AC-BFV and the Rank-1 CMS are 73.3% and 60.3%, respectively.

5.5. Performance of combined ACBP and TFV histograms

Since the TFV vectors provide better results than the BFV vectors we evaluated the performance of combined histogram of ACBP, AC-TFV and DC-TFV histograms as defined in (11). In Fig. 5 the CMS curves of combined histograms are shown overlaid with results for single histograms. One can see that combination of ACBP and TFV histograms provides significantly better retrieval performance than using them individually, with 91% Rank-1 CMS correct retrieval.

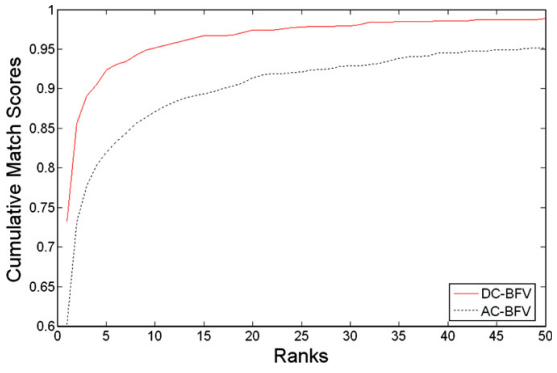


Fig. 4. CMS over FERET database using DC-BFV and AC-BFV. The CMS at certain rank represents the ratio of correct retrievals among all the probe images.

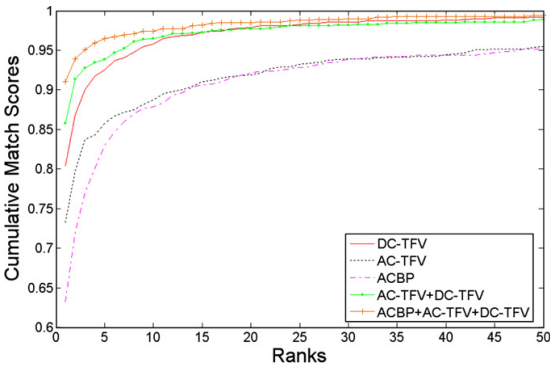


Fig. 5. Cumulative match scores results of combined histograms.

For comparison, results obtained by other methods are shown in Tables 2 and 3. Table 2 lists the results based on the release 2003 of the FERET database (Chunghoon et al., 2005; Shi et al., 2005, 2006; Roure and Faundez, 2005); Table 3 list the results based on earlier releases (Jones and Viola, 2003; Kyungim et al., 2002; Xiangsheng et al., 2005). The results are shown for the Rank-1 CMS. Such comparisons are only informative since there are differences in the preprocessing of images used for testing. Nevertheless, our approach having over 90% correct retrieval is much better than many other methods and it is lower only to some which are much more complicated.

5.6. Performance of combined ACBP and TFV histograms using cross-validation between FERET and ORL database

In order to show the robustness of the proposed method, we also performed testing and validation for the ORL and FERET databases. There is considerable body of research using ORL which justifies the comparison. We used a subset of ORL for training and complementary to it from ORL set for evaluation. We also performed cross-validation between the databases. In this case, the set of parameters obtained from training with FERET is used for retrieval with ORL. Such challenging task should indicate for the robustness of our method with respect to training and retrieval.

ORL database is much smaller than FERET; it has 10 images for each person. Training was performed in this case by randomly selecting five images for the training set and the other five for the evaluation set. The retrieval process is repeated 100 times. First, the training and retrieval are conducted within each database itself. Next, the training and retrieval are cross-exchanged when the optimal parameter set obtained by training over FERET is applied to test with ORL, and vice versa. Since the ORL database is largely different from FERET in many aspects like size, structure, face expression, this experiment can be used to

Table 2
List of the referenced results based on release 2003

References	(Shi et al., 2005)	(Shi et al., 2006)	(Roure and Faundez, 2005)	(Chunghoon et al., 2005)	Proposed
Method	Landmark bidimensional regression	Landmark's geometry	Template matching	Combined subspace	ACBP + AC − TFV + DC-TFV, L_1 norm
Rank-1 CMS (%)	79.4	60.2	73.08	97.9	91

Table 3
List of the referenced results based on different releases

References	(Kyungim et al., 2002)				(Jones and Viola, 2003)	(Xiangsheng et al., 2005)	Proposed
Method	PCA-L_1	PCA-L_2	PCA-Cosine	ICA-Cosine	Boosted local features	JSBoost	ACBP + AC-TFV + DC-TFV, L_1 norm
Rank-1 CMS (%)	80.42	72.80	70.71	78.33	94	98.4	91

Table 4
Two optimal parameter sets obtained respectively from training data of FERET and ORL

	QS of ACBP histogram	Width of ACBP histogram	α	QS of AC-TFV histogram	Width of AC-TFV histogram	β	QS of DC-TFV histogram	Width of DC-TFV histogram
FERET	92	400	1	72	800	3	10	400
ORL	96	400	1	72	800	1	10	400

Table 5
Recognition rates over ORL in comparison to references

Reference	(Issam and Rabih, 2006)	(Kotani et al., 2002)	(Erik, 2000)	(Rahul et al., 2000)	(Travieso et al., 2005)	(Roure and Faundez, 2005)	Proposed	Proposed
Method	IPCA_PCA	VQ	Local Gabor Feature	Arena	SVM + DWT	Template matching	Training ORL	Training FERET
Rank-1 CMS (%)	88.3724	97	85	96.2	90.8	92.5	97.22	96.2

judge the robustness of the system. The training parameter sets from both databases are shown in Table 4.

Results are shown in Table 5 with comparison to other publications (Erik, 2000; Issam and Rabih, 2006; Kotani et al., 2002; Rahul et al., 2000; Roure and Faundez, 2005; Travieso et al., 2005). When training and evaluation is performed using ORL our method gives best results with 97.2% correct retrieval. Using the FERET training set and testing on the ORL set the retrieval level of 96.2% is only slightly diminished. Thus indeed the approach is giving very good and robust performance even if the training is performed on another database set.

6. Conclusions

This paper deals with performance of similarity measures for histograms based on quantized block transforms. We first define several types of feature vectors and based on quantized block transform coefficients. Next histograms of feature vectors and combination histograms are introduced together with city-block similarity measure. Subsequently, we define the image database retrieval system and explain its operation in the training phase when the similarity measure coefficients are optimized. The system performance evaluation is done using face databases FERET and ORL with cumulative match score performance measure. Our final results show over 90% correct first hit recognition rate for FERET and over 97% for ORL. These results are very good when taking into consideration simplicity of our approach which is based only on statistics of selected features and no structural information about locations is used. The system is also shown to be robust for the optimal coefficients selecting during the training phase. Thus, the results indicate that local feature selection and statistical information about them is critical and should be carefully optimized before structural information is included for boosting the correct recognition. The proposed system is going into this direction and in the future research we are planning to incorporate structural information measure into it.

Acknowledgements

The first author would like to thank for the financial grant from Tampere Graduate School in Information Science and Engineering (TISE) and Nokia Foundation Award.

References

Chunghoon, K., Jiyong, O., Chong-Ho, C., 2005. Combined Subspace Method Using Global and Local Features for Face Recognition. In: Proc. of Int. Joint Conf. on Neural Networks.

Delac, K., Grgic, M., Grgic, S., 2006. Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set. Int. J. Imaging Syst. Technol. 15 (5), 252–260.

Erik, H., 2000. Biometric Systems: A Face Recognition Approach. In: Proc. of the Norwegian Conf. on Informatics, pp. 189–197.

FERET Face Database, 2003. Available from: http://www.itl.nist.gov/iad/humanid/feret/. National Institute of Standards and Technology, USA.

He, D.-C., Wang, L., 1990. Texture unit, texture spectrum, and texture analysis. IEEE Trans. Geo Sci. Remote Sens. 28 (1), 509–513.

Issam, D., Rabih, N., 2006. Face Recognition Using IPCA-ICA Algorithm. IEEE Trans. Pattern Anal. Mach. Intell. 28 (6), 2006.

Jones, M., Viola, P., 2003. Face recognition using boosted local features. In: Proc. of Int. Conf. on Computer Vision.

JVT of ITU-T and ISO/IEC JTC 1, 2003. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC). Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050.

Kotani, K., Qiu, C., Ohmi, T., 2002. Face Recognition Using Vector Quantization Histogram Method. In: Proc. of Int. Conf. on Image Process., II-105.

Kyungim, B., Bruce, A.D., J.Ross, B., Kai, S., 2002. PCA vs. ICA: A comparison on the FERET data set. In: Proc. of Int. Conf. on Computer Vision. Pattern Recognition and Image Proceeding, Durham, North Carolina.

Ojala, T., Pietikäinen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on feature distributions. Pattern Recognit. 29 (1), 51–59.

ORL Face Database, 2005. Available from: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. AT&T Laboratories, Cambridge.

Phillips, P.J., Moon, H., Rizvi, S., Rauss, P., 2000. The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 22 (10).

Rahul, S., Matthew, M., Shumeet, B., 2000. Memory-based face recognition for visitor identification. In: Proc. of IEEE Face and Gesture.

Roure, J., Faundez, Z.M., 2005. Face recognition with small and large size databases. In: Proc. of 39th Annual 2005 Int. Carnahan Conf. on Security Technology.

Shi, J., Samala, A., Marx, D., 2005. Face Recognition Using Landmark-Based Bidimensional Regression. In: Proc. of ICDM 2005, pp. 765–768.

Shi, J., Samala, A., Marx, D., 2006. How Effective are Landmarks and Their Geometry for Face Recognition. Comput. Vis. Image Understand. 102 (2), 117–133.

Travieso, C.M., Alonso, J.B., Ferrer, M.A., 2005. Strategy for improving the reliability in the facial identification. In: Proc. of 39th Annual 2005 Int.Carnahan Conf. Security Technology.

Xiangsheng, H., Li, S.Z., Yangsheng, W., 2005. Jensen–Shannon boosting learning for object recognition. In: Proc. of CVPR 2005, 2, pp. 144–149.

Yossi, R., Jan, P., Carlo, T., Joachim, M.B., 2001. Empirical evaluation of dissimilarity measures for color and texture. Comput. Vis. Image Understand. 84 (1), 25–43.

Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A., 2003. Face recognition: A literature survey. ACM Comput. Surveys (CSUR) 35 (4), 399–458.

Zhong, D., Defée, I., 2005. DCT histogram optimization for image database retrieval. Pattern Recognit. Lett. 26 (14), 2272–2281.

Publication VII

DaiDi Zhong, Irek Defée, "Face Recognition In Compressed Domain Using Ternary Feature Vector", in Proceedings of European Signal Processing Conference (EUSIPCO 2007), pp. 1580-1584, Sep. 2007.

IMAGE RETRIEVAL USING EFFICIENT FEATURE VECTORS GENERATED FROM COMPRESSED DOMAIN

Daidi Zhong, Irek Defée

Department of Information Technology, Tampere University of Technology.
P.O. Box 553, FIN-33101 Tampere, Finland
{daidi.zhong, irek.defee}@tut.fi

ABSTRACT

We consider the use of quantized block transform coefficients to the image database retrieval problem. Based on the transform coefficients feature vectors are constructed. These feature vectors are used in histograms and combination of histograms with similarity measure for database retrieval. Experiments on public face image database show good performance of the approach in comparison with other methods.

1. INTRODUCTION

Block transforms are widely used in signal compression due to their ability to preserve of perceptual information even under strong quantization. This property of block transform should be also very useful for pattern recognition and retrieval tasks. However, application of block transforms in these areas has to be done within a suitable framework. This is because patterns are formed by global distribution of features which is described both by detailed geometry (structure) as well as statistics. In this paper formation of features and description of their global distribution based on statistics are investigated. Methods which are developed are compared with other approaches within the standardized framework of image database retrieval evaluation.

Our approach to feature representation based on quantized block transforms is related to the Local Binary Pattern (LBP) [1, 11] and texture spectrum unit methods [2]. We do not apply those methods directly in the image pixel domain but to the quantized block transform coefficients and we use different thresholding schemes and we do not apply them directly in the image pixel domain but to the quantized block transform coefficients. We form features as binary and ternary vectors based on the DC and AC quantized transform coefficients.

Global description of patterns used in this paper is based on histograms of feature vectors. The histograms describe the statistical content of images using defined feature vectors. We use both histograms and combination of histograms for different feature vectors. Such description does not use geometry of feature locations but nevertheless we are able to show that it is quite powerful. Using histograms, standard city-block metrics can be used as similarity measure for patterns. We apply this approach to the image database retrieval problem. Evaluation is done and results are presented in the standardized framework developed for FERET face database with its performance evaluation based on Cumulative Match Score (CMS) [3]. We first tune the free coefficients using limited training set and then evaluation of performance is conducted using large database test set. Results are presented for the FERET database and compared with other methods. It is shown that our method which is based only on global statistics of selected features and defined similarity measure has performance in the range of achieved by other methods which are more complex.

2. BLOCK TRANSFORMS AND QUANTIZATION

The specific block transform used in this paper was introduced in the H.264 standard [4] as particularly effective and simple. The transform matrix of the transform is denoted as $\boldsymbol{B_f}$ and the inverse transform matrix is denoted as $\boldsymbol{B_i}$. They are defined as

$$\boldsymbol{B}_{\mathrm{f}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \qquad \boldsymbol{B}_{\mathrm{i}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix}$$

A 4x4 image pixel block $\boldsymbol{P}$ can be forward transformed to block $\boldsymbol{H}$ using (1), and the quantization process $\mathbf{Q}(\ )$ is used to remove the irrelevant information, which will result in quantized version of $\boldsymbol{H}$, $\mathbf{Q}(\boldsymbol{H})$. During the reconstruction process, the inverse quantization process $\mathbf{Q}^{-1}[]$ is applied to the quantized block $\mathbf{Q}(\boldsymbol{H})$, and block $\boldsymbol{R}$ is subsequently reconstructed from the inverse-quantized block $\mathbf{Q}^{-1}[\mathbf{Q}(\boldsymbol{H})]$, using (2)

$$H = B_f \times P \times B_f^T \tag{1}$$

$$R = B_i^T \times Q^{-1}\left[Q(H)\right] \times B_i \tag{2}$$

with superscript T denoting transposition.

The leading element of the matrix $\boldsymbol{H}$ is called the DC coefficient. All other elements are called AC coefficients. There are thus 15 AC coefficients in the matrix $\boldsymbol{H}$ but many of them will have zero value after the scalar quantization $\mathbf{Q}(\boldsymbol{H})$ is applied. Quantization has the effect of limiting the number of different blocks in an image. The quantized blocks with AC coefficients can be used directly in image retrieval as described in [5] where they were called AC Block Patterns (ACBP).

3. FEATURE VECTORS

In this paper we define features as Feature Vectors (FV) based on quantized block transform coefficients. The block transform (1) can be performed for non-overlapping image blocks or for partially overlapping image blocks. In either case we can group separately DC and AC coefficients for all transform blocks of an image into matrices. From these matrices feature vectors are formed.

3.1 DC binary feature vectors

From the matrix of DC coefficients, 3x3 submatrices of DC coefficients are considered. The eight DC coefficients surrounding the center one can be thresholded to form a binary vector with length eight:

If the DC value < threshold then put 0
If the DC value ≥ threshold then put 1

Value of the threshold can defined in many ways, e.g. as the value of the central coefficient in the 3x3 submatrix or as the mean value of all its nine coefficients. We choose the latter way and the binary feature vector formed in this way is denoted as DC-BFV.

3.2 AC binary feature vectors

Following the procedure described for the formation of the DC-BFV above, the binary feature vectors can be also defined for the AC coefficients. We shall denote such vectors as AC-BFV. There are 15 AC coefficients possible within each block but many coefficients will take zero value when quantization is applied. We select in the 4x4 block matrix coordinates AC coefficients which are (0,1), (1,0) and (3,0) and their positions are shown marked in Figure 1. Such decision is made based on the retrieval tests run over training data. In addition, using smaller number of AC coefficients can reduce the complexity.

	X		
X			
X			

Figure 1 – The AC coefficients used

3.3 Ternary feature vectors

Extension of binary feature vector described above to a ternary feature vector which can be obtained by the following thresholding:

If the DC value < threshold put 0
If the DC value = threshold put 1
If the DC value > threshold put 2

with the same rules for the AC coefficients.

The Ternary Feature Vector (TFV) is based on flexible threshold range instead of a single value of the threshold above. For a 3x3 matrix of DC or AC coefficients, the threshold T is defined as

$$T_{-}^{+} = M \pm (X - N) \times f \qquad (3)$$

where f is real number from the interval (0,1), X and N are maximum and minimum coefficient values in the coefficient matrix, and M is the mean value of the coefficients. The thresholded values are to be either 0, 1 or 2

If the coefficient value ≤ T+ put 0
If the coefficient value ≥ T- put 2
otherwise put 1

The resulting thresholded vectors of length eight are subsequently converted to decimal numbers in the range of [0, 6560].

4. HISTOGRAMS OF FEATURE VECTORS AND SIMILARITY MEASURE

4.1 Histograms of Feature Vectors

Statistical distribution of specific feature vectors in images can be represented in normalized way by histograms. Example of such a histogram is shown in Figure 2. The histograms can be also seen as 1-D vectors and city-block metrics can be used as similarity measure.

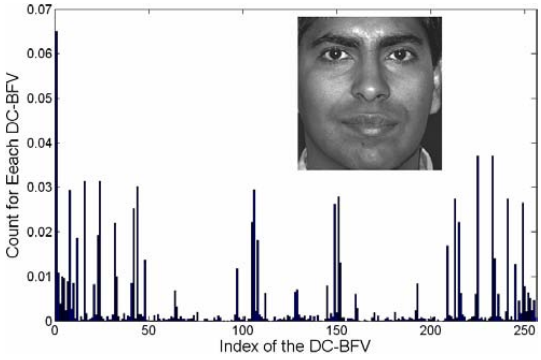


Figure 2. Example histogram of DC Binary Feature Vectors

Several types of feature vectors which were described above will lead to different feature vector histograms. We shall define now combinations of these histograms and similarity measure for them.

4.2 Combination of histograms

We can form three types of feature histograms based on ACBP, DC Feature Vectors (DC-FV) and AC Feature Vectors (AC-FV) respectively; These histograms can be combined in several ways starting with any two of the histograms, as follows

[Combined_2Histogram] =
[Histo_A α x Histo_B] (4)

where Histo_A and Histo_B can be any one of the histograms with ACBP, DC-FV or AC-FV.

The combined histogram is thus a vector formed by concatenating the histograms A and B with coefficient α introduced for controlling the relative weight of A and B in the combination. In the same way we can combine the three histograms, with two weight parameters α and β:

[Combined_3Histogram] =
[Histo_ACBP α x Histo_AC-FV β x Histo_DC-FV] (5)

For AC-FV histograms, we use three AC coefficients as shown in Figure 1. They are combined in the following way:

[Histo_AC] =
[Histo_coef1 Histo_coef2 Histo_coef3] (6)

The similarity measure based on the combined histogram is the same as for the single histogram but obviously the values of coefficients α and β have to be established first. This is done within the framework of database retrieval system training described next.

4.3 Image Database Retrieval

Our image database retrieval problem is formulated as follows. Given an image database set **S**={ I_1, …., I_n} we would like to establish for a certain key image **I** if there are images similar to it in the database. For this we use the feature vector histograms of images and similarity measure defined above to find its minimum values for the image **I** and images from the database **S**.

However, before this can be done the parameters used for the calculation of histograms and similarity measure need to be found using training database set. This set can be selected as small subset of the database $\boldsymbol{S}$ or some key images selected from outside of it. Knowing the correct responses for the training database allows us to tune the parameters to achieve best retrieval results. The optimal parameter set which is going to be found out during the training process includes: the Quantization Scalar for (QS) for the quantization of transform blocks $\mathbf{Q}(\boldsymbol{H})$ in (2), coefficients α and β in (5,6) and the parameters $\boldsymbol{f}$ in (3). The optimal parameter set is identified as the one which is maximizing the retrieval performance over training database. The resulting optimal parameter set is applied to the test database $\boldsymbol{S}$ to evaluate the actual system performance.

5. RESULTS OF RETRIEVAL PERFORMANCE

5.1 Image database

To train and test the retrieval performance of the proposed system we used a famous public database of face images -- FERET [6], which contains about two thousand images from about one thousand subjects. The advantage of using FERET database, apart of its size, is standardized evaluation method based on performance statistics reported as Cumulative Match Scores (CMS), which are plotted on a graph [3]. The horizontal axis of the graph is the retrieval rank and vertical axis is the probability of identification (PI) (or percentage of correct matches). This lets one know how many images have to be examined to get a desired level of performance. For simplicity, many researchers use the CMS at the first rank to represent the recognition rate.

The FERET database provides some tools for preprocessing of the face images. The images can be cropped by the provided geometrical information about eye, nose and mouth. They can be subsequently aligned, and adjusted by illumination normalization. In our evaluation, the images

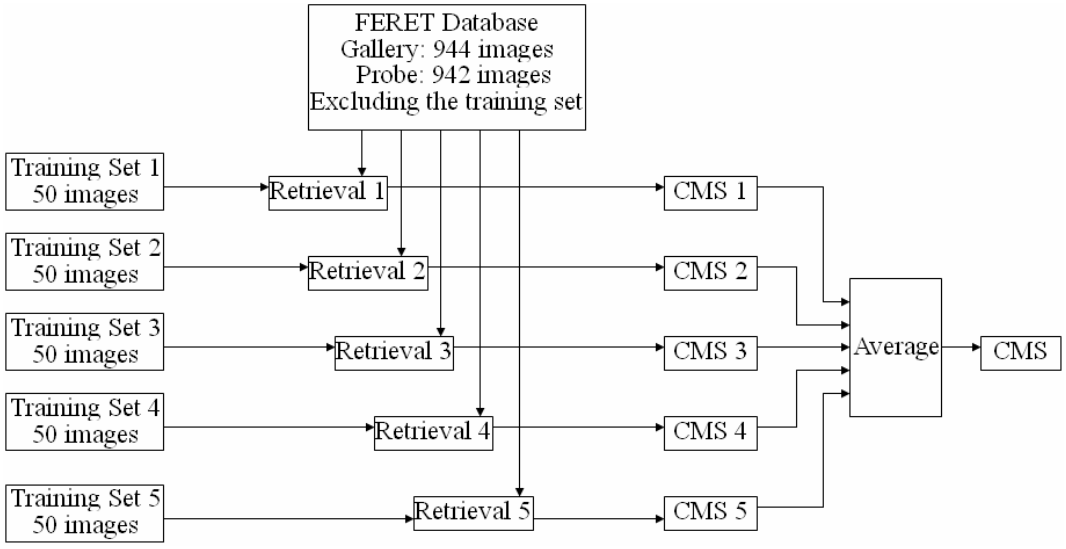


Figure 4. The training and retrieval system. Five training sets are used for testing the robustness and generality of our system.

were simply cropped to sizes, which roughly contain the face area (e.g. Figure 3).

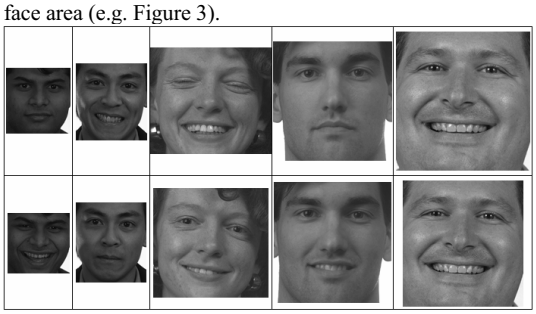


Figure 3. Some exemplar images from FERET database. The upper row is from gallery set, and the lower row is from probe set. No preprocessing is performed over them. They have different sizes.

5.2 Training process

Before the real retrieval task is performed, we first select the training sets from FERET database. Here the problem is selection of the training set which on one hand should be small but on the other hand the results should not depend on particular selection of the set. In our experiment, five different test image sets are used for training, each composed of fifty images selected randomly. The training and retrieval system is illustrated in Figure 4.

The size of one training set is 50 images. Some example pairs of images are shown in Figure 3. For each of the test sets optimal coefficients described in Section 4.3 were found. In result we established that coefficients found for each set are very similar and the system performance is not sensitive to the selection of the test set.

5.3 Performance of BFV histograms using FERET

We first studied the performance of BFV histograms using the FERET database. Figure 5 shows the Cumulative Match Scores (CMS) results for AC-BFV and DC-BFV histograms. From the result, one can find out that DC-BFV performs better than AC-BFV and the Rank-1 CMS is 0.733 and 0.603, meaning that probability for correct first hit retrieval is 73.3% and 60.3% respectively.

5.4 Performance of TFV histograms using FERET

Here we compare the performance of DC-TFV and AC-TFV. The TFV histogram may contain maximum 6561 bins. To reduce the complexity, the bins are sorted and only the first several hundred high rank bins are used for the histogram. The results are shown in Figure 6. Comparing with Figure 5 one can see that the DC-TFV and AC-TFV histograms provide much better results, with 80.4% and 73.3% correct first hit retrieval.

5.5 Performance of combined ACBP and TFV histograms

Since histograms of ternary TFV vectors provide better results than histograms of binary BFV vectors we evaluated the performance of combined histogram of ACBP, AC-TFV and DC-TFV histograms as defined in (6). In Figure 7 the Cumulative Match Scores (CMS) results of combined histograms are shown overlaid with results for single histograms. One can see that combination of ACBP and TFV histograms provides significantly better retrieval performance than using them individually, with 91% Rank-1 CMS correct retrieval.

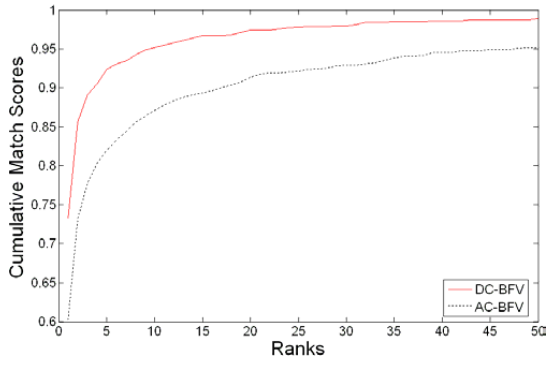


Figure 5. CMS over FERET database using DC-BFV and AC-BFV. The CMS at certain rank represents the ratio of correct retrievals among 992 probe images.

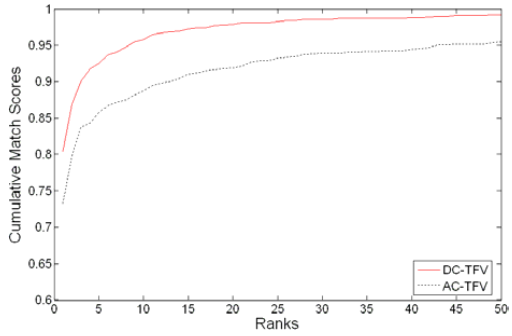


Figure 6. CMS over FERET database using DC-TFV and AC-TFV. The CMS at certain rank represents the ratio of correct retrievals among 992 probe images.

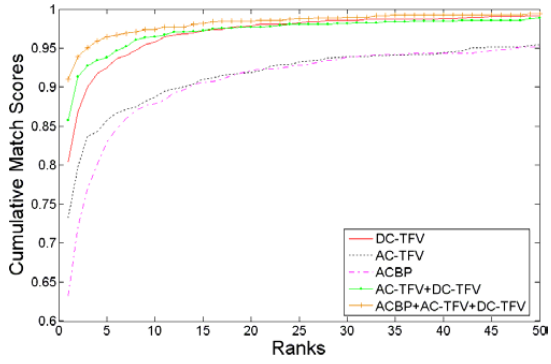


Figure 7. Cumulative match scores results of combined histograms.

And the variation between the performances of different training sets is 1%.

For comparison, results obtained by other methods are shown in Table 1 which lists the results based on the FERET database of the same version (released in 2003); The results are shown for the Rank-1 CMS. Such comparisons are only informative since there are differences in the preprocessing of images used for testing. Nevertheless, our approach having over 90% correct retrieval is much better than many other methods and it is lower only to some which are much more complicated.

Reference	[7]	[8]	[9]	[10]	Proposed
Rank-1 CMS (%)	79.4	60.2	73.08	97.9	91

Table 1. List of the referenced results based on FERET

6. CONCLUSIONS

In this paper we defined several types of feature vectors based on quantized block transform coefficients. Next histograms of feature vectors and combination histograms are introduced together with city-block similarity measure. The system performance evaluation is done using face databases FERET with cumulative match score performance measure. Our final results show over 90% correct recognition rate for FERET. Performance is very good since our approach is based only on the statistics of features, while other more complicated methods may also incorporate some structural information.

REFERENCES

[1] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions", *Pattern Recognition*, Vol. 29, No. 1, pp. 51-59, 1996.

[2] He, D.-C. and Wang, L., "Texture unit, texture spectrum, and texture analysis", *IEEE Trans. Geo Sci. Remote Sens*, 28 (1), 509-513, 1990

[3] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 10, October 2000.

[4] Joint Video Team of ITU-T and ISO/IEC JTC 1, "Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC)", *Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050*, 2003.

[5] D. Zhong and I. Defée, "DCT histogram optimization for image database retrieval", *Pattern Recognition Letters*, 26(14): 2272-2281, 2005.

[6] FERET Face Database, Available at: http://www.itl.nist.gov/iad/humanid/feret/, 2003.

[7] J. Shi, A. Samal, and D. Marx, "Face Recognition Using Landmark-Based Bidimensional Regression", in *Proceeding of ICDM 2005*, 765-768, 2005.

[8] J. Shi, A. Samal, and D. Marx, "How Effective are Landmarks and Their Geometry for Face Recognition", *Computer Vision and Image Understanding*, 102(2):117-133, 2006.

[9] J. Roure, and M. Faundez-Zanuy, "Face recognition with small and large size databases", in Proc. of ICCS 2005.

[10] C. Kim, J. Oh, and C.H. Choi, "Combined Subspace Method Using Global and Local Features for Face Recognition", in *Proceedings of IJCNN 2005*.

[11] Timo Ahonen, Matti Pietikainen, Abdenour Hadid, Topi Maenpaa, "Face Recognition Based on the Appearance of Local Regions", in Proceedings of ICPR 2004.

Publication VIII

DaiDi Zhong, Irek Defée, "Face Image Retrieval System Using TFV and Combination of Subimages", in Proceedings of International Conference series on Visual Information Systems (VISUAL 2007), June 2007.

Face Image Retrieval System Using TFV and Combination of Subimages

Daidi Zhong and Irek Defée

Department of Information Technology, Tampere University of Technology,
P.O. Box 553, FIN-33101 Tampere, Finland
{daidi.zhong,irek.defee}@tut.fi

Abstract. Face image can be seen as a complex visual object, which combines a set of characterizing facial features. These facial features are crucial hints for machine to distinguish different face images. However, the face image also contains certain amount of redundant information which can not contribute to the face image retrieval task. Therefore, in this paper we propose a retrieval system which is aim to eliminate such effect at three different levels. The Ternary Feature Vector (TFV) is generated from quantized block transform coefficients. Histograms based on TFV are formed from certain subimages. Through this way, irrelevant information is gradually removed, and the structural and statistical information are combined. We testified our ideas over the public face database FERET with the Cumulative Match Score evaluation. We show that proper selection of subimage and feature vectors can significantly improve the performance with minimized complexity. Despite of the simplicity, the proposed measures provide results which are on par with best results using other methods.

Keywords: Face image retrieval, FERET, TFV, subimage.

1 Introduction

Face image retrieval is a highly complex pattern recognition problem due to the enormous variability of data. The variability is due to a combination of both statistical and structural factors. Despite this variability, images are very efficiently processed for pattern extraction by biological systems. This means that those systems are robust in extracting both the statistics and structure from the data. The details of how their processing is done are not deciphered yet, and in fact even the formulation what efficiency and robustness means in the biological context is not clear. However, it seems that combination of structure and statistics-based processing is used in this process. In this sense, the face images are mixtures of structure and statistics which makes the description problem hard because its complexity looks like unbounded. In addition, the image quality often suffers from the noise and different light conditions, which make the retrieval tasks more difficult.

Some previous works focused on extracting and processing global statistical information by using the whole image [1], while some other researchers start from

G. Qiu et al. (Eds.): VISUAL 2007, LNCS 4781, pp. 49–60, 2007.

some key pixels [2] to represent the structural information. Based on their achievement, a reasonable way to further improve the retrieval performance is to extract the visual information in a way like a mixture of statistical and structural information.

In this paper, we illustrate our idea by proposing a retrieval system which is based on subimages and combinations of feature histograms. This can also be seen as a pathway to from local information (pixels) to the global information (whole image). The experimental results disclose that the usage of subimage and local feature vectors can lead to the combination of statistical and structural information, as well as minimized impact of noise, which finally improve the performance of the approach.

In order to achieve a comparable result, we tested our method over a public benchmark of face image database. The evaluation method of this database has been standardized, which allow us to see the change of performance clearly. However, using face images as an example here does not mean our method is limited to the application of face image retrieval; it also has the potentiality to be applied to other image retrieval tasks.

2 Transform and Quantization

Visual images are usually in the format of 2-D matrix of pixels. Such representation contains a large amount of redundant information. Several image and video compression standards are intend to minimize such redundancy by utilizing lossy compression methods. Transform and quantization are two common steps of these methods. From the view of compression, they can reduce the redundancy; from the view of retrieval, they help us to find out the most distinguishable features.

Some transforms have been found useful in extracting local visual information from images. Popular transforms include: Gabor Wavelet, Discrete Wavelet Transform, Discrete Cosine Transform (DCT), and Local Steerable Phase. Specially, DCT and Wavelets have already been adopted to the image and video compression standards [3],[4]. The specific block transform we used in this paper was introduced in the H.264 standard [5] as particularly effective and simple. The transform matrix of the transform is denoted as $\mathbf{B_f}$ and the inverse transform matrix is denoted as $\mathbf{B_i}$. They are defined as

$$B_f = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \quad B_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix} \tag{1}$$

The 4x4 pixel block **P** is forward transformed to block **H** using (2), and the block **R** is subsequently reconstructed from **H** using (3). The '**T**' means linear algebraic transpose here.

$$H = B_f \times P \times B_f^T \tag{2}$$

$$R = B_i^T \times H \times B_i \tag{3}$$

A 4x4 block contains 16 coefficients after applying the transform. The main energy is distributed around the DC coefficient, namely, the upper-left coefficient. The rest are called AC coefficients, which are usually quite small. The small coefficients contain small energies, which is generally regarded as irrelevant information or noise. Therefore, quantization can be applied immediately after the transform to remove the redundancy. The reduction of the amount of information greatly facilitates our information retrieval tasks.

Furthermore, it also has the effect of limiting the dynamic range of coefficients. Larger quantization level tends to reduce the range of possible coefficients. In the ultimate case, when the quantization level is large enough, all the coefficients will be thresholded to zero; thus, no distinguishing ability can be achieved from them, which is not the case we wish to see. Therefore, we use certain training process to find out the optimal range of quantization level. The objective is a good compromise between removing non-important visual information and preserving the main perceptually critical visual information.

3 Feature Vectors

Block transform and quantization arranged the local information in a suitable way for retrieval. Based on this merit, we utilize the specific feature vector defined below to further group the local information in the neighboring blocks. The grouping process can be applied separately or jointly over DC and AC coefficients for all transform blocks of an image.

Considering a 3x3 block matrix containing nine neighboring blocks, the DC coefficients from them can form a 3x3 coefficient matrix. The eight DC coefficients surrounding the center one can be thresholded to form a ternary vector with length eight. This vector is called DC Ternary Feature Vectors (DC-TFV), which encode the local information based on those quantized transform coefficients.

The threshold is defined as a flexible value related to the mean value of all the nine DC coefficients.

$$\begin{aligned} Threshold_1 &= M + (X - N) \times f \\ Threshold_2 &= M - (X - N) \times f \end{aligned} \tag{4}$$

where f is real number from the interval (0, 0.5), X and N are maximum and minimum values in the 3x3 coefficient matrix, and M is the mean value of the coefficients. Our initial experiments have shown that performance with changing f has broad plateau for f in the range of 0.2~0.4. From this reason, we use f = 0.3 in this paper. The thresholded values can be either 0, 1 or 2.

If the coefficient value $\leq$ Threshold$_2$ put 0
If the coefficient value $\geq$ Threshold$_1$ put 2
otherwise put 1

The resulting thresholded vectors of length eight are subsequently converted to decimal numbers in the range of [0, 6560], where $6560=3^8$. However, not all of these 6561 TFV are often present in the face images. In fact, only a small part of them often appear. Therefore, we only utilize the most common TFVs for retrieval tasks. The proper selection is done based on training process.

The same process as above can be applied to AC coefficients. The resulted feature vectors are called AC-TFV. Due to the quantization and intrinsic property, not all the AC coefficients are suitable for the retrieval task. Especially, those coefficients representing high frequency energy tends to have little dynamic range of values. Their distinguishing ability is too poor to achieve good performance. Therefore, we only select those coefficients which can show high distinguishing ability. Such selection is done based on training process.

On the other hand, using more coefficients will certainly increase the complexity. The proper selection can be conducted with training set. However, we only present the results with one very capable AC coefficient, which can already show fairly good retrieval accuracy. The information obtained from it will be further combined with the information obtained from DC coefficient to enhance the performance. Since they are representing different information, the combination of them is expected to show more aspects of the visual object.

4 Representation Based on Subimages

Certain facial areas, such as eye, nose and mouth, are generally believed to contain more distinguishing ability than other areas. Here we refer to such areas as subimage. It is not sagacious enough to treat all the subimages of face image equally. Some researchers have already noticed this point. For example, Ahonen tried to manually assign different weights to the features generated from a set of predefined subimages [6]. Alternatively, in [7], the features obtained from subimages are combined with the features obtained from the whole image. Proper decision is made by certain data fusion strategy. In [8], Chunghoon manually selected several subimages and generate PCA features based on them. In this paper, we do it in a different way: we randomly defined some rectangular subimages over the original image. TFV is extracted from each subimage separately. These TFV vectors are represented by special histograms, which may further be combined to serve the retrieval tasks.

Rather than manually select the eye or noise areas, we randomly selected 512 subimages, which can be overlapped to cover the whole image. There is a wide dynamic range of the sizes of these subimages: the smallest one is about 1/150 of the size of whole image, while the largest one is about 1/5. Both the large and small subimages are selected from every region of the face image. Some examples of subimage are shown in Fig.1.

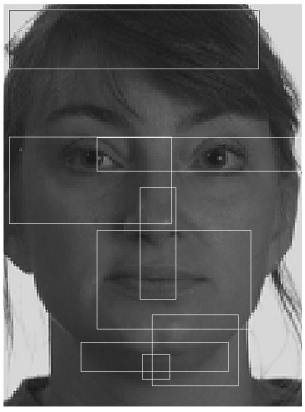


Fig. 1. Examples of subimage (each rectangle is a subimage)

5 Histograms of Feature Vectors and Similarity Measure

The aforementioned block transform, TFV and subimage can be regarded as three levels of collecting local information. Such system gradually moves from local to global aspects of the image. They are linked by using a histogram of TFV features. The generation of histograms is done in the following way:

1. The 4x4 H.264 AC Block Transform is applied to a subimage.
2. Quantization is applied separately to all the AC and DC coefficients.
3. TFV is generated from certain coefficient.
4. Histogram is generated from this subimage by simply counting the number of each occurring TFV.
5. Histogram is normalized according to the size of subimage.

Specifically for AC-TFV histogram, there is one bin which is too dominant comparing to other bins. This is caused by the smooth area in image and quantization. Such areas will generate many all-one vectors, like [1 1 1 1 1 1 1 1]. Our retrieval does not use this bin, since it decreases the discriminate ability.

Histogram based on DC-TFV and AC-TFV can be used separately or collectively. Since they represent different information, the combination of them can leads to better performance, which will be shown in the following experiment. The combination is done by simply concatenating each histogram one by one. For example, two subimages (sub1 and sub2) are used to generate the DC-TFV and AC-TFV histograms, the combined histogram can be shown like:

[Combined Histogram] =

$$[\text{DC-TFV-sub1}\ \ \text{AC-TFV-sub1}\ \ \text{DC-TFV-sub2}\ \ \ \text{AC-TFV-sub2}] \tag{5}$$

There is no specific weight for each histogram. In another word, the information coming from different subimages are treated equally.

The retrieval task is completed by classifying any input image to a known person, according to the distance between their corresponding histograms. There are several

well-known similarity measures which are widely used in pattern recognition society, e.g., L1 norm, L2 norm, Cosine distance and Mahalanobis distance. The *pdist* function in MATLAB [9] has implemented several of them, which are used by us to test over some training data. Finally, we choose to use the L1-norm. It is simple to be calculated and suitable for the TFV histograms. Thus, the distance between two histograms $H_i(b)$ and $H_j(b)$, b= 1, 2, … B are calculated as:

$$\text{Distance}(i, j) = \sum_{b=1}^{B} \left| H_i(b) - H_j(b) \right| \tag{6}$$

6 Experiments with FERET Database

6.1 FERET Database

For testing the performance of the proposed method we use the FERET face mage database [10]. The advantage of using this database is its standardized evaluation method of based on performance statistics reported as Cumulative Match Scores (CMS) which are plotted on a graph [11]. Horizontal axis of the graph is retrieval rank and the vertical axis is the probability of identification (PI) (or percentage of correct matches). On the CMS plot higher curve reflects better performance. This lets one to know how many images have to be examined to get a desired level of performance since the question is not always "is the top match correct?", but "is the correct answer in the top n matches?"

National Institute of Standard and Technology (NIST) have published several releases of FERET database. The release which we are using is the one published at October 2003, called Color FERET Database. It contains overall more than 10,000 images from more than 1000 individuals taken in largely varying circumstances. Among them, the standardized FA and FB sets are used here. FA set contains 994 images from 994 different objects, FB contains 992 images. FA serves as the gallery set, while FB serves as the probe set.

Before the experiments, all the source images are cropped to contain face and a little background. They are normalized to have the same size. The eyes are located in the similar position according to the given information from FERET. Simple histogram normalization is taken to the entire image to tackle the luminance change. However, we did not apply any mask for the face images, which are used by some researchers to eliminate the background.

6.2 Retrieval System with the Training Process

We made experiments over certain number of subimages. During the training and retrieval process, the same set of subimages is applied to all the images. TFV histograms are generated from them, and are subsequently used for retrieval. The training process is described in Fig.2. To ensure the independence of training set, five different groups of images are randomly selected to be the training sets. Each group contains 50 image pairs (from 992 pairs). Five parameter sets are obtained from them, which will be applied for evaluation of performance of the rest 942 images of the

whole database. The resulting five CMS curves are averaged, which is the final performance result.

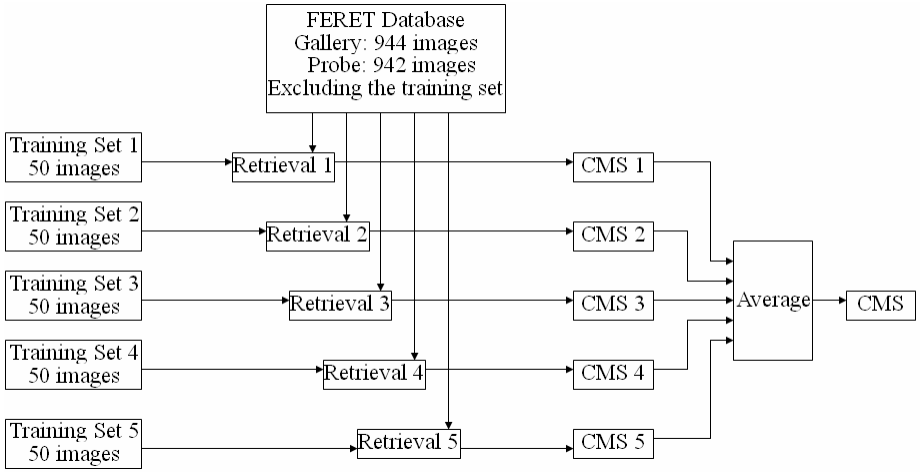


Fig. 2. Training and Retrieval process. The optimal parameter set from five training sets are utilized separately, which give five CMS scores. The overall performance of given subimage will be evaluated as the average of above five CMS scores.

6.3 Performance of TFV Histograms Using the Single Subimage

The first experiment is conducted over the aforementioned randomly selected 512 subimages. Since we have five training sets, the final result is actually a matrix of 5x512 CMS scores. They are further averaged to be a 1x512 CMS vector. The maximum, minimum and average of these 512 CMS scores are shown in Table 1. One can see from it that there is very wide performance variation for different subimages. The selection of subimage is thus critical for the performance which can be achieved. However, this is also reasonable since there are certain subimages which are too small to provide enough information. Fig.3 shows the distribution of rank-1 CMS scores using DC-TFV from single subimage.

When interpreting the results in Table 1, one may also conclude that: the DC-TFV subimage histograms always perform markedly better than DC-TFV histograms but their combination performs still better in the critical high performance range. When comparing them, one may use the average CMS scores as the criterion. It is representing the overall performance, rather than the performance of any specific subimage.

Table 1. The Rank-1 CMS results of using single subimage

Rank-1 CMS (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Maximum	93.77	60.77	95.30
Minimum	9.01	1.69	12.94
Average	56.59	20.99	62.11

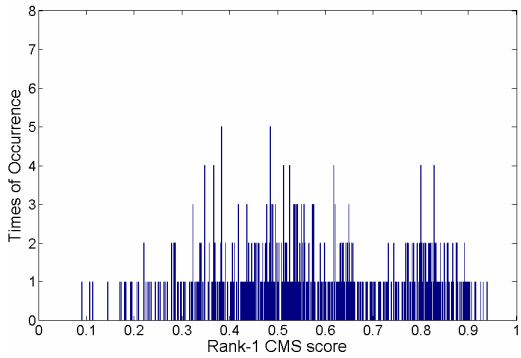


Fig. 3. The distribution of rank-1 CMS scores using DC-TFV generated from single subimage

As we mentioned before, the training process is conducted over five training sets. Therefore, we have five set of CMS scores which are obtained from the same database. It is thus interesting to study their differences. Table 2 shows the maximum, minimum and average difference between five training sets. As one can see, the average difference between different training sets is only about one percent. This is quite good considering the size of training set is about only 5.2% of the probe set.

Table 2. The difference between five training sets, using DC-TFV from single subimage, represented by Rank-1 CMS scores

	Maximum	Minimum	Average
Rank-1 CMS (%)	3.73	0.1	1.26

6.4 Performance of TFV Histograms Using the Single Subimage

Based on above results, a reasonable way to improve the performance is to combine multiple subimages. In the following experiment, we utilized 216 pairs of subimage for retrieval. Most of these selected subimages can provide relatively better performance in the previous experiment. In addition, when any two subimages are making a pair, they must be coming from different region of the face image. For example: one is from eye region, and the other is from mouth region. The retrieval task is performed over them and the best pairs are identified.

Furthermore, one additional subimage from different region is added to above 216 pairs. Therefore, we have another experiment which is using three subimages. Totally, we have 432 combinations of 3-subimage. All of these three subimages are coming from different regions of the face image.

The maximum, minimum and average of the resulted scores for both the 2-subimage and 3-subimage cases are shown in Table 3. Clearly we can see the improvement from 1-subimage, 2-subimage to 3-subimage. In addition, we show in Table 4 the average difference between five training sets when doing these two sets of experiments. The difference is less than one percent, which can be safely regarded that the results are not sensitive to the selection of training sets.

Table 3. The Rank-1 CMS results of using two/three subimage

Rank-1 CMS (%)	DC-TFV		AC-TFV		DC-TFV + AC-TFV	
Subimage	Two	Three	Two	Three	Two	Three
Maximum	97.76	97.50	81.94	89.19	97.70	98.23
Minimum	47.54	76.21	13.47	24.60	52.50	78.99
Average	79.06	92.03	43.89	63.15	82.56	93.59

Again, similar to the Fig.2, we also show the distribution of rank-1 CMS scores in Fig.4 and Fig.5. They are obtained by using DC-TFV histograms generated from 2-subimage and 3-sbuimage respectively.

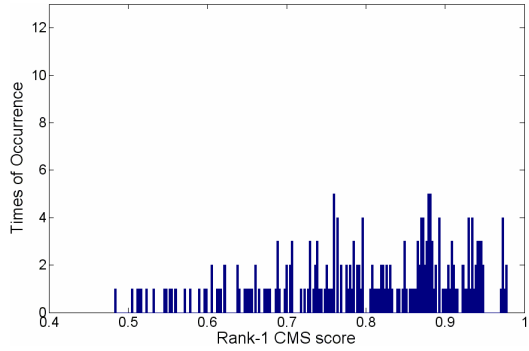


Fig. 4. The distribution of rank-1 CMS scores using DC-TFV generated from two subimages

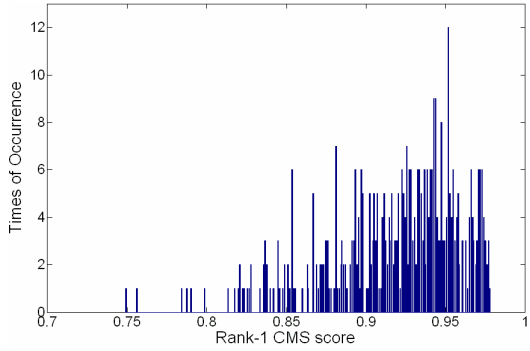


Fig. 5. The distribution of rank-1 CMS scores using DC-TFV generated from three subimages

Table 4. The average difference between five training sets, using DC-TFV from one/two/three subimage, represented by Rank-1 CMS scores

	1-subimage	2-subimage	3-subimage
Average difference of Rank-1 CMS (%)	1.26	0.94	0.62

To give a better understanding of the improvement from 1-subimage to 3-subimage, we would like to compare the result of any combination and its corresponding component subimages. The CMS score of each 3-subimage combination is compared to the corresponding CMS score of 2-subimage combination; and the CMS score of each 2-subimage combination is compared to the corresponding CMS scores of two component 1-subimage. The compared results are shown in Table 5. The notation I_1, I_2 and I_3 means the subimages which constitute the combinations.

Table 5. The average difference between five training sets, using DC-TFV from one/two/three subimage, represented by Rank-1 CMS scores

	$I_1+I_2 < \min(I_1, I_2)$	otherwise	$I_1+I_2 > \max(I_1, I_2)$
2-sub VS. 1-sub (among 216 combinations)	0	37/216	179/216
	$I_1+I_2+I_3 < \min(I_1+I_2, I_3)$	otherwise	$I_1+I_2+I_3 > \max(I_1+I_2, I_3)$
3-sub VS. 2-sub (among 432 combinations)	0	220/432	212/432

We can notice that the possibility to get improvement from 1-subimage to 2-subimage is higher than the possibility from 2-subimage to 3-subimage. In fact, the performance of 2-subimage is already in a saturation range, which is quite difficult to get improvement. Furthermore, using more subimages is not necessarily to achieve better performance, especially when the previous subimages have already represented quite distinguishing contents, adding the more subimages usually do not result in any improvement. This is possible when the added subimage has too small size, or it contains too much noise or useless texture, and so on. However, we can also conclude that adding new subimage at least does not reduce the performance. Based on above results, we also believe that using four or five subimages may further improve the performance, but such improvement will be relatively small. Proper decision has to be made by also taking the complexity into consideration.

6.5 Comparison with Other Methods

In order to compare the performance of our system with other methods, we list below some reference results from other researches for the FERET database. These results are all obtained by using the FA and FB set of the same release of FERET database. In [12], the eigenvalue weighted bidimensional regression method is proposed and applied to biologically meaningful landmarks extracted from face images. Complex principal component analysis is used for computing eigenvalues and removing correlation among landmarks. An extensive work of this method is conducted in [2], which comparatively analyzed the effectiveness of four similarity measures including the typical L1 norm, L2 norm, Mahalanobis distance and eigenvalue-weighted cosine (EWC) distance. The author of [13] employs a simple template matching method to complete a verification task. The input and model faces are expressed as feature vectors and compared using a distance measure between them. Different color

channels are utilized either separately or jointly. A combined subspace method is proposed in [8], using the global and local features obtained by applying the LDA-based method to either the whole or part of a face image respectively. The combined subspace is constructed with the projection vectors corresponding to large eigenvalues of the between-class scatter matrix in each subspace. The combined sub-space is evaluated in view of the Bayes error, which shows how well samples can be classified. Table 6 lists the result of above papers, as well as the result of 3-subimaghe case of our method. The results are expressed by the way of Rank-1 CMS score.

Table 6. Referenced results based on release 2003 of FERET

References	[12]	[2]	[13]	[8]	Proposed Method
Rank-1 CMS (%)	79.4	60.2	73.08	97.9	98.23

7 Conclusions

We proposed a hierarchical retrieval system based on block transform, TFV and subimage for visual image retrieval. Such histogram gradually integrates the structural and statistical information in the face images. The performance is illustrated using a public face image database. This system achieves good retrieval results due to the fact it efficiently combines the statistical and structural information. Future research will try to find out any simpler way to select the optimal subimages.

Acknowledgments. The first author would like to thank for the financial grant from Tampere Graduate School in Information Science and Engineering (TISE).

References

1. Ekenel, H.K., Sankur, B.: Feature selection in the independent component subspace for face recognition. Pattern Recognition Letter 25, 1377–1388 (2004)
2. Shi, J., Samal, A., Marx, D.: How effective are Landmarks and Their Geometry for Face Recognition. Computer Vision and Image Understanding 102(2), 117–133 (2006)
3. Pennebaker, W.B., Mitchell, J.L.: JPEG still image compression standard. Van Nostrand Reinhold, New York (1993)
4. ISO/IEC 14496-2: Information Technology - Coding of Audio-Visual Objects - Part 2: Visual (1999)
5. ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC: Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (2003)
6. Ahonen, T., Hadid, A., Pietikainen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
7. Rajagopalan, A.N., Srinivasa, R.K., Anoop, K.Y.: Face recognition using multiple facial features. Pattern Recognition Letters 28, 335–341 (2007)
8. Chung, H.K., Jiyong, O., Chong-Ho, C.: Combined Subspace Method Using Global and Local Features for Face Recognition. In: IJCNN 2005 (2005)

9. MathWorks, Inc.: Documentation for MathWorks Products (2007), Available at http://www.mathworks.com
10. FERET Face Database: (2003), Available at http://www.itl.nist.gov/iad/humanid/feret/
11. Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S.: The FERET evaluation methodology for face recognition algorithms. IEEE Pattern Analysis and Machine Intelligence 22, 10 (2000)
12. Shi, J., Samal, A., Marx, D.: Face Recognition Using Landmark-Based Bidimensional Regression. In: proceeding of ICDM 2005 (2005)
13. Roure, J., Faundez, Z.M.: Face recognition with small and large size databases. In: proceeding of ICCST 2005 (2005)

Publication IX

DaiDi Zhong, Irek Defée, "A Framework for Combining Statistical and Structural Pattern Retrieval Based on Feature Histograms", in Proceedings of IEEE International Workshop on Machine Learning For Signal Processing (MLSP 2007), August 2007.

A Framework for Combining Statistical and Structural Pattern Retrieval Based on Feature Histograms

Daidi Zhong, *Student Member, IEEE*, Irek Defée

Abstract—**Patterns are characterized by the distribution of features. In general detailed geometry of feature locations and statistics of features distribution are important for the characterization. We call these aspects structural and statistical information of patterns and aim for developing framework for the unified description of them. Statistical information can be simply and conveniently picked by feature histograms, structural information description is much more complex. In order to deal with it we are introducing the concept of hierarchical decomposition of pattern areas. Areas are described using statistical information by feature histograms, size and number of areas reflects structural information. This formulation unifies statistical and structural information and the problem of minimizing structural information is stated as reducing the number and size of the histograms. We illustrate this on an example of retrieval from face image database using features based on quantized block transform coefficients. We can show that very limited structural information is needed for nearly perfect retrieval performance equal to the best available algorithms.**

I. Introduction

THE problem of representation of visual patterns is very complex. Patterns are formed from limited sets of localized features but the number of features and their distributions are virtually unlimited. This makes the description of patterns difficult and as a result of this, complicates the pattern recognition and retrieval tasks. Human face representation illustrates well the issues which lead to two different views of pattern representation: local and global. The local view aims for describing structures from the bottom up, starting from some set of features. Examples of this are methods based on local filters or transforms which include for instance: Gabor Wavelets [1], Discrete Wavelets Transform (DWT) [2] and Discrete Cosine Transform (DCT) [3]. The global view in turn aims for top level whole pattern description, for example methods like PCA [4], ICA [4], and LDA [5] which perform certain global transformation of the data to reduce dimensionality and at the same time rearrange the information in a suitable way for the retrieval task. Both global and local approaches have been intensively studied and shown to produce good results but it seems there is still missing a common unified framework for the local and global approach.

In this paper we consider the problem of pattern representation from a perspective which aims for the unification of the local and global viewpoints. We take from the local approach the concept that features are important and patterns are constructed from them. We take the global view as to how patterns are formed by features. Global description which we use is based on the realization that it should be including both geometry and statistics of pattern distribution in a unified way. We call the part of description based on geometry structural information since it depends on the location of features. Statistical part of the description depends only on the statistics of the features distribution and not on their location, and it is called statistical information. Statistical information is conveniently described by feature histograms of patterns. Structural information description is much more complicated, our aim is to develop a framework unifying it with statistical information. This is done by statistics described by histograms and decomposition of patterns into areas which are described by histograms. Histograms of areas measure their statistical information, their shapes, sizes and locations reflect structural information. Such formulation allows us to consider problem of minimizing structural information needed to accomplish specific task using the database retrieval problem.

II. Statistical And Structural Description Of Visual Patterns

We consider the visual patterns as composed by concatenations (in 2D plane) of elements from some basic feature set $\mathbf{F} = \{F_1 \ldots F_M\}$. Statistical information about patterns can be conveniently described by feature histograms $\mathbf{H}$ using full or part of the set $\mathbf{F}$. Since histograms are normalized with bins corresponding to the frequency of occurrences of elements of $\mathbf{F}$ they can also be treated as normalized vectors and city-block metrics may be then used to measure their similarity.

Assume now that a pattern P defined over some area C is described by a histogram H over feature set F. We shall now define covering of the image area $\mathbf{C}$ by a set of subareas $C_1 \ldots C_S$ which do not have to be disjoint. For each of the subareas C_s ($s = 1, \ldots, S$) covering subpattern $\mathbf{P_s}$, its corresponding histogram $\mathbf{H_s}$ will be calculated. The description of pattern $\mathbf{P}$ is now done by the set of histograms $\{H_1 \ldots H_s\}$ and patterns can be compared using city-block metrics of vector composed

Manuscript received March 29, 2007.

Daidi Zhong is with the Institute of Signal Processing, Tampere University of Technology, P.O.Box 553, FIN-33101, Tampere, Finland (phone: +358331154503; fax: +35833653087; e-mail: daidi.zhong@ieee.org).

Irek Defée is with Institute of Signal Processing, Tampere University of Technology, POBox 553, FIN-33101, Tampere, Finland (e-mail: irek.defee@tut.fi).

of concatenation of the histograms $[H_1 \ldots H_S]$. As the histograms are normalized by the number of features in each subarea, this is not equivalent to the case when one single histogram is generated from the whole pattern. The difference is that the impacts of histograms from different subarea on the metrics are emphasized. An example of subpatterns and covering for S=3 is shown in Fig. 1.

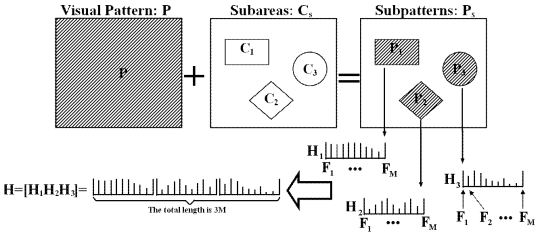


Fig. 1. Example of a pattern P, covering C and histogram formation.

Decomposition of the pattern into covering brings structural information into description. This can be seen by considering decompositions where the pattern area is partitioned into increasing number of disjoint subareas. Each subarea carry then increasing amount of structural information related to the location of features. In the extreme case subarea is just a single feature which is then located with maximal precision. In our framework the subarea is described by statistical information, i.e. the histogram, but the number of subareas, their size and shape reflects structural information. There is no structural information associated with the histogram calculated over the whole pattern area. When the pattern is split into subareas, the covering carries structural information. The content of each subarea is measured only by statistical information but this is not a problem since in case when this is not sufficient the subarea can be split by a cover and/or new subareas can be added. Using subareas for pattern retrieval is natural and has been used in the past in many different contexts [6,7,17]. Our approach using feature histograms allows for unifying statistical information with structural information carried by subareas in a way allowing for the investigation of minimizing the information needed for achieving specific retrieval performance. This problem is formulated in the context of image database retrieval as follows.

Given an image database set $\mathbf{S}=\{I_1, \ldots, I_N\}$ we would like to establish for certain key image I if there are images similar to it in the database. For this we can use histograms of suitably selected features sets images city-block similarity measure to find its minimal values for the image I and images from the database **S**. If the histograms are calculated for whole images, the retrieval will be based on statistical information only. When this would give suitable performance, no structural information is necessary but it often will not be the case. Then structural information has to be used and possibly optimized. For this, we select a representative subset of the database called training set and will search for minimal covering which will provide best retrieval performance. When such covering is found, it will represent minimal structural information for a given performance level. In this way, using statistical information measure for coverings, structural information can be minimized. At the same time the overall computational complexity is not increased since once the optimal covering is found the calculation of histograms for subareas is essentially equivalent to the calculation of a single histogram for the whole image.

III. Feature Selection And Structural Information

The feature set selected, the length of the feature histograms and the covering play an important role in the retrieval process. The goal for using these histograms is to allow efficiently comparison between different patterns. The histograms themselves have to show enough discriminating ability between each other. If the length is too short, we cannot achieve enough discrimination; while if the length is too long, the retrieval will suffer from the heavy computation, the redundant information and the possible noise. Therefore, only a judicious selection of features may lead to a good trade-off between accuracy and efficiency.

In this paper, we fulfill this requirement by statistically evaluating the a priori probability of the appearance of each feature. Those features which are less likely to be present in most of the images are considered as redundant features, thus are removed from the histograms. The feature selection is conducted by a training process which will be described later. Given a certain area, there are numerous ways to define subareas in order to cover it. However, not every distribution of subareas can lead to good retrieval result. The distribution should properly fit the inherent visual property of the target pattern. Therefore, an optimization is required when defining the subareas. In our method, this is done by evaluating the a priori performances of some randomly selected subareas. The histograms from subareas which provide good results can be further combined to provide better performance.

We show this idea by using the face image as an example. One complete face image can be seen as a combination of the following regions: forehead, eyes, nose, mouth, chin and rest area of the face. Some exemplar subareas are shown in Fig. 2. Using these subareas one can encode a certain amount of structural information. Basically, it can be regarded as a description of macro-structure. In order to further exploit the micro-structural in-formation, we propose a hierarchical system to encode such information step by step. This system indirectly collects information from image areas according to the manipulations described in the following sections.

Theoretically, the number of possible distribution of subareas is unlimited. It seems to be that the searching process of optimal subareas is an endless process. However, our experimental results show that even with two or three selected rectangular subareas, we can already achieve an excellent performance.

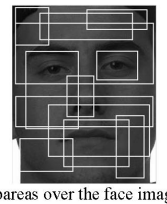


Fig. 2. Some example subareas over the face image.

IV. Features From Pixel Blocks By Block Transform

The Block Transform Coding methods are widely used in image and multimedia compression. The transform coding method compact image data by representing the original signal with a small number of transform coefficients. It exploits the fact that for typical images a large amount of signal energy is concentrated in a small number of coefficients. The goal of transform coding is to minimize the number of retained transform coefficients while keeping distortion at an acceptable level. Transform coding is an integral part of one of the most widely known standards for lossy image compression, the JPEG (Joint Photographic Experts Group) standard [8]. Such kind of ability to re-arrange the information in suitable way can contribute not only to the compression, but also to the recognition. Some researchers have already utilized it in their face image retrieval works.

The DCT Block transform coding divides an image into blocks of equal size and processes each block independently. Block processing allows the coder to adapt to local image statistics, exploit the correlation present among a block of neighboring image pixels, and to reduce computational and storage requirements. For example, JPEG uses 8×8 blocks. Larger blocks do not offer significantly better compression, since correlation between pixels tends to decrease as the block size grows. A fixed block size allows the design of optimized software and hardware.

The transform used in this research is taken from the H.264 standard [9]. This transform is a 4x4 integer transform, which is originally aim to encode the coefficients of inter blocks. Overall, this transform performs in a similar way with the widely-used DCT. The first uppermost coefficient after transform is called DC and it corresponds to average light intensity level of a block, other coefficients are called AC and they correspond to components of different frequencies. The AC coefficients provide us some useful information about the texture detail of this block. The ability of integer calculation allows rapid process. In addition, it makes the information compact, which greatly facilitates the information extraction.

The forward transform matrix of H264 AC Transform is $\mathbf{B_f}$ and the inverse transform matrix is $\mathbf{B_i}$.

$$B_f = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \quad B_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix} \quad (1)$$

The 4x4 pixel block **P** is forward transformed to block **K** using (2), and block **R** is subsequently reconstructed from **K** using (3). The 'T' means linear algebraic transpose here.

$$K = B_f \times P \times B_f^T \quad (2)$$

$$R = B_i^T \times K \times B_i \quad (3)$$

V. Quantization Of Transform Coefficients

The reason why the above block transforms are used stems from their robustness to quantization. Quantization of transform coefficients by special matrices as in the case of JPEG or scalars as in the 4x4 H.264 transform preserves to very high degree important perceptual features. As the energy is mostly distributed around the DC coefficient, quantization can make most of the AC coefficients to zero. The remaining non-zero AC coefficients indicate the existence of major textural information in this block area, while the removed (become zero) AC coefficients are usually perceptually redundant information.

During the process when the applied Quantization Scalar (QS) is being increased, both the useful and redundant information are gradually removed. Fortunately, the speeds of these two reductions are different. There is an optimal area within which the redundant information is largely removed, while the major part of the useful information is still well preserved [10].

VI. Feature Vectors And Their Histograms

A. Ternary Feature Vector

The above block coefficients inherently contain the structural information of the corresponding block. In the following step, the Ternary Feature Vector (TFV) described in [11] which is used to further collect the information from a larger area. In our case here, this larger area is the area covering nigh neighboring blocks. That also means one TFV actually represents a part of information coming from neighboring 12x12 pixel blocks.

Each of the 4x4 transform coefficient block contains sixteen coefficients as shown in Fig. 3. The TFV is formed from the same coefficient all over the subarea. The basic unit for generating a TFV is also a 3x3 coefficient matrix. The formation is done by thresholding the eight neighboring coefficients with two thresholds. Each block can produce one ternary vector with length of eight. This 8-bin vector is subsequently converted to a decimal number in the range of [0, 6560].

0	**1**	**2**	**3**
4	**5**	**6**	**7**
8	**9**	**10**	**11**
12	**13**	**14**	**15**

Fig. 3. Each 4x4 transform block has 16 coefficients ordered as shown.

The thresholds are defined in a flexible way based on the

local information. Within each 3x3 matrix, assuming the max value is MAX, the min value is MIN, the mean value is MEAN, the threshold is calculated by:

$$Threshold_{-}^{+} = MEAN \pm (MAX - MIN) \times f \qquad (4)$$

where f is a real number within the range of (0, 0.5). Our initial experiments have shown that performance with changing f has broad plateau for f in the range of 0.2~0.4. From this reason, the value f =0.3 is fixed in our system. The thresholding of coefficients is done in the following way:

0 - The neighboring pixel value $\leqslant$ Threshold^{+}
1 - The neighboring pixel value otherwise
2 - The neighboring pixel value $\geqslant$ Threshold$_{-}$

B. *Histogram of TFV*

By collecting all the TFVs generated from one subarea, a histogram of TFV can be generated with length of 6561. To assure fair comparison, the histogram is normalized by the size of subarea.

This histogram shows the statistical distribution of TFV within this subarea. As we described in section 2, not all the bins of the histogram is useful for retrieval. The length will be significantly reduced by the training process.

In addition, not all the 16 coefficients will be used for retrieval. We only use two of them in our experiments. They are the 0th and 4th coefficient in Fig. 3. The corresponding TFV are called DC-TFV and AC-TFV respectively.

During the face image retrieval process, the input image is compared to any image stored in the database, in order to find the most similar one. In our method, such similarity is measured by calculating the L1-norm distance (city-block distance) between two histograms. For example, suppose we have two histogram $H_i(b)$ and $H_j(b)$, b= 1, 2, ... L. The distance will be calculated like:

$$\text{Distance}(i, j) = \sum_{b=1}^{L} \left| H_i(b) - H_j(b) \right| \qquad (5)$$

In Fig. 1, we have shown an example of combining histograms from three subareas. In fact, from each subarea, we can generate one DC-TFV histogram, as well as one AC-TFV histogram. They can be used separately, or in a concatenating way. They can be combined like:

[Combined Histogram of certain subarea]
= [Histogram-DC-TFV Histogram-AC-TFV] (6)

VII. Experiments With FERET Database

National Institute of Standard and Technology (NIST) have published several releases of FERET database. The release which we are using is the one published at October 2003, called Color FERET Database [12].

The Color FERET Database contains overall more than 10,000 images from more than 1000 individuals taken in largely varying circumstances. Among them, the standardized FA and FB sets are used here. FA set contains 994 images from 994 different objects, FB contains 992 images. FA serves as the gallery set, while FB serves as the probe set.

The advantage of using this database is the standardized evaluation method of FERET [13] based on performance statistics reported as Cumulative Match Scores (CMS), which are plotted on a graph. This lets one know how many images have to be examined to get a desired level of performance since the question is not always "is the top match correct?", but "is the correct answer in the top n matches?" The author would like to emphasize: although the face image database is used here to present the result, no adjustments specific to face image have been taken. In fact, the only reason to use face image database is its widely usage and standardized testing protocols, so that we can compare with different methods.

Before the real experiment, all the source images are cropped to contain face and a little background. They are normalized to have the same size. The eyes are located in the similar position according to the given information from FERET. Simple histogram normalization is taken to the entire image to tackle the luminance change.

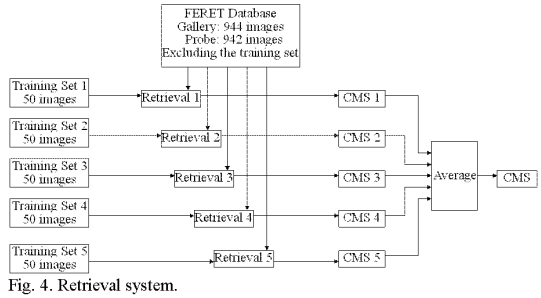


Fig. 4. Retrieval system.

In order to ensure the generalness of the proposed methods, 5 different groups of images are randomly selected to be the training sets. Each group contains 50 image pairs (from 992 pairs). Five optimal parameter sets are obtained from them, which will be applied to the rest 942 images of the whole database. Subsequently, the resulted five CMS curves are averaged, which gives the final results. The retrieval system is illustrated in Fig. 4.

A. *Performance of TFV histograms using the complete image*

We first studied the performance of TFV by using the complete image. Table I shows the CMS results of DC-TFV, AC-TFV histograms and their combination. These results can be the starting point and reference for the following results. We will refer to this experiment as Test-A in the following text.

From the result, one can find out that:

- Overall, DC-TFV performs better than AC-TFV. In another word, DC-TFV is the main contributor.

- Overall, their combination can provide better result. However, since the performance is already near the saturation

area, the improvement is relatively small, only about one percentage.

B. *Performance of TFV histograms using the single subarea*

Secondly, we studied the performances when using single subarea. We will refer to this experiment as Test-B in the following text. Our focus is not to find the best subarea but to investigate the problem of minimizing structural information by selecting few subareas. Evidently subareas with good performance would be of interest but there is also the question how much the performance of subareas differs. To have idea about this we randomly defined 512 subareas (Fig. 2).

The retrieval performance of each subarea is obtained by one retrieval experiment. Since we have five training sets for cross validation, the final result is actually a matrix of 5x512 CMS scores. They are further averaged to be a 1x512 CMS vector. The maximum, minimum and mean of these 512 CMS scores is shown in Table II. As should be expected poor results often come from very small subareas.

Next we consider the case when complete image is divided into two subareas, one is the rectangle subarea we used above, and the other is the rest area of the image when the subarea is cut out. We called the two cases as: Partial Image Decomposition (PID) and Full Image Decomposition (FID). Table III shows the same contents as Table II, for the FID case.

Clearly, the FID case shows better result than the PID case. Reason for this is that the subarea emphasizes the specific part of the image which can contribute to increased discrimination ability. This is better than using the complete image, without any consideration of subarea specificity.

C. *Full image by subareas processing*

Based on above results, a reasonable way to improve the performance is to combine multiple subareas. For this purpose, the Test-C is conducted here by randomly select two subareas from different regions (forehead, eye, mouth, chin, etc.). Based on the above 512 subareas in Test-B, there are totally 216 2-subarea combinations for Test-C. The corresponding CMS results of Test-C are shown in Table IV and V for PID and FID case respectively.

To give a better understanding of the improvement from 1-subarea to 2-subarea, The CMS score of each 2-subarea combination is compared to the corresponding CMS scores of 1-subarea components. The compared results are shown in Table IV. The notation I_1, I_2 and I_3 means the subareas which constitute the combinations.

Based on above results, we conclude that:

1. For PID case, the 2-subarea generally performs better than the 1-subarea case.
2. Using more subareas is not necessarily to achieve better performance.
3. The probability to achieve improvement in the PID case is much higher than in the FID case.

TABLE I
FOR RESULTS OF USING COMPLETE IMAGE

TEST-A

	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Rank-1 CMS score (%)	92.84	64.31	93.65

TABLE II
FOR RESULTS OF USING SINGLE SUBAREA (PID)

TEST-B

Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Maximum	93.77	60.77	95.30
Minimum	9.01	1.69	12.94
Mean	56.59	20.99	62.11

TABLE III
FOR RESULTS OF USING SINGLE SUBAREA (FID)

TEST-B

Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Maximum	97.94	82.82	98.06
Minimum	31.49	7.48	35.04
Mean	84.12	51.42	86.48

TABLE IV
FOR RESULTS OF USING TWO SUBAREA (PID)

TEST-C

Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Maximum	97.76	81.94	97.70
Minimum	47.54	13.47	52.50
Mean	79.06	43.89	82.56

TABLE V
FOR RESULTS OF USING TWO SUBAREA (FID)

TEST-C

Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Maximum	98.43	89.31	98.71
Minimum	76.15	45.28	80.54
Mean	92.87	71.30	94.14

TABLE VI
COMPARE 216 COMBINATIONS OF 2-SUBAREA TO CORRESPONDING TWO SUBAREA COMPONENTS.

2-sub VS. 1-sub (among 216)	$< \min(I_1, I_2)$	otherwise	$> \max(I_1, I_2)$
PID	0	37 / 216	179 / 216
FID	0	379 / 216	53 / 216

This conclusion is reasonable since the added subarea cannot always contribute the retrieval. This may happen when the added subarea has too small size, too much noise or useless texture, etc. However, we can see that with proper selection of subareas the performance improves when the number of subareas is increased.

D. Comparison with other methods

In order to compare the performance of our system with other methods, we list below some results from other researchers. Table VII lists the result (maximum Rank-1 CMS) based on the same release of our FERET database as well as the result of 2-subarea FID (2-FID) case of our method. These results are all obtained by using the FA and FB set of the same release of FERET database, therefore, they are directly comparable with our results. All the results are expressed in the way of Rank-1 CMS score.

TABLE VII
LIST OF THE REFERENCED RESULTS BASED ON RELEASE 2003 OF FERET DATABASE

Reference	[14]	[15]	[16]	[17]	Proposed 2-FID Method
Rank-1 CMS score (%)	79.4	60.2	73.08	97.9	98.71

VIII. CONCLUSIONS

In this paper we investigated a framework for image database retrieval problem which combines statistical and structural information by using feature histograms computed over pattern coverings. Our features are selected as quantized block transform coefficients which guarantees their high perceptual relevancy with low complexity of description. Structural information of patterns in our approach is described by subsets included in the covering and can be minimized by reducing the number of subsets. We illustrate our approach on the example of FERET image database. We show that with using of just two subareas (and the rest of the image as the third one) our approach gives retrieval accuracy close to 99% and better than other methods.

ACKNOWLEDGMENT

The first author would like to thank for the financial grant from the Tampere Graduate School in Information Science and Engineering (TISE).

REFERENCES

[1] F. Annalisa, L. Alessandra, M. Dario, and N. Loris, "An enhanced subspace method for face recognition," *Pattern Recognition Letter*, vol. 27, pp. 76-84, 2006.

[2] H. K. Ekenel and B. Sankur, "Multiresolution face recognition," *Image and Vision Computing*, vol. 23, pp. 469-477, 2005.

[3] D. Ramasubramanian and Y. V. Venkatesh, "Encoding and recognition of faces based on the human visual model and DCT," *Pattern Recognition*, vol. 34, pp. 2447-2458, 2001.

[4] H. K. Ekenel and B. Sankur, "Feature selection in the independent component subspace for face recognition," *Pattern Recognition Letter*, vol. 25, pp. 1377-1388, 2004.

[5] L. Juwei, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol. 26, pp. 181 – 191, 2005.

[6] A. N. Rajagopalan, K. S. Rao, and Y. A. Kumar, "Face recognition using multiple facial features", *Pattern Recognition Letters*, vol. 28, pp. 335–341, 2007.

[7] K. Keun-Chang and W. Pedrycz, "Face recognition: A study in information fusion using fuzzy integral", *Pattern Recognition Letters*, vol. 26, pp. 719–733, 2005.

[8] W. B. Pennebaker and J. L. Mitchell, "JPEG still image compression standard," New York, Van Nostrand Reinhold, 1993.

[9] Joint Video Team, "ITU-T Recommendation H.264: Advanced video coding for generic audiovisual services," *ITU-T Rec. H.264*, May 2003.

[10] Z. Daidi and I. Defée, "DCT histogram optimization for image database retrieval," *Pattern Recognition Letters*, vol. 26, pp. 2272-2281, 2005.

[11] Z. Daidi and I. Defée, "Study of image retrieval based on feature vectors in compressed domain," In *Proceedings of NORSIG 2006*.

[12] FERET Face Database, Available at: http://www.itl.nist.gov/iad/humanid/feret/.

[13] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090-1104, 2000.

[14] J. Shi, A. Samal, and D. Marx, "Face Recognition Using Landmark-Based Bidimensional Regression," In *Proceeding of ICDM 2005*.

[15] J. Shi, A. Samal, and D. Marx, "How Effective are Landmarks and Their Geometry for Face Recognition," *Computer Vision and Image Understanding*, vol. 102, pp. 117-133, 2006.

[16] J. Roure and M. Faundez-Zanuy, "Face recognition with small and large size databases," In *Proceeding of ICCS 2005*.

[17] K. Chunghoon, O. Jiyong, and C. Chong-Ho, "Combined Subspace Method Using Global and Local Features for Face Recognition," In *Proceedings of IJCNN 2005*.

Publication X

DaiDi Zhong, Irek Defée, "Fast Searching For The Optimal Area Of TFV Representation", in Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP 2007), October 2007.

Fast Searching For The Optimal Area Of TFV Representation

Daidi Zhong
Department of Information Technology
Tampere University of Technology
Tampere, Finland
daidi.zhong@ieee.org

Irek Defée
Department of Information Technology
Tampere University of Technology
Tampere, Finland
irek.defee@tut.fi

***Abstract*—Visual images are often characterized by the distribution of certain key features. Taking the face image as an example, the eye, nose and mouth are often regarded as characterizing features for recognizing face image. We call these aspects structural and statistical information of visual images and aim for developing framework for the unified description of them. We extracts certain features from randomly chosen subareas, these features have good capability to represent the local texture information. We show our retrieval results over the public face database. We found that certain subareas can provide quite good retrieval results, but the thorough searching for such subareas are time-consuming. We further developed a simple fast searching method which can large simplifies the searching process, while in the same time preserve the good performance.**

Keywords—image retrieval; subarea; fast searching method; statistical; structural

Topic area—Single- and multimedia analysis.

I. INTRODUCTION

Visual image retrieval has been an active research area of pattern recognition and computer vision for decades. Our previous research [1] in this field is concentrated on efficient features which can properly describe the statistical information of the visual content. We proposed the Ternary Feature Vector (TFV) to represent the local texture, and these vectors are arranged in a way of histogram. We tested the TFV over a public face database – FERET [2], which is often used by researchers as a benchmark. The results proved the good distinguishing ability of TFV.

However, the problem of representation of visual patterns is more complex than a simple global histogram. In fact, the image can contain a number of key features which can essentially describe the most informative content of it. Generally speaking, the detailed geometry of feature locations and statistics of features distribution are important for the characterization. We call these aspects structural and statistical information of visual images and aim for developing framework for the unified description of them. Therefore, generating a global histogram over the whole image is not an optimal way to extract the major visual information there. We would like to concentrate on certain areas which can optimally describe the major information.

Some researchers have already moved further along this direction. For example, Ahonen tried to manually assign different weights to the features generated from a set of predefined subareas [3]. Alternatively, in [4], the features obtained from subareas were combined with the features obtained from the whole image. Proper decision was made by certain data fusion strategy. In [5], Kim manually selected several subareas and generated PCA features based on them. In this paper, we do it in a different way: we randomly defined some rectangular subareas over the original image. TFV is extracted from each subarea separately. These TFV vectors are represented by special histograms, which may further be combined to serve the retrieval tasks.

II. DESCRIPTION OF VISUAL PATTERNS USING SUBAREA

We consider the visual patterns as composed by concatenations (in 2D plane) of elements from some basic feature set $F = \{F_1 \ldots F_M\}$. Statistical information about patterns can be conveniently described by feature histograms H using full or part of the set F. Assume now that a pattern P defined over certain area C is described by a histogram H over feature set F. We shall now define covering of the image area C by a set of subareas $C_1 \ldots C_S$. For each of the subareas C_s (s = 1, …, S) covering subpattern P_s, its corresponding histogram H_s will be calculated. The description of pattern P is now done by the set of histograms $\{H_1 \ldots H_s\}$ and patterns can be compared using city-block metrics of vector composed of concatenation of the histograms $[H_1 \ldots H_s]$. As the histograms are normalized by the number of features in each subarea, this is not equivalent to the case when one single histogram is generated from the entire pattern. The difference is that the impacts of histograms from different subarea on the metrics are emphasized. An example of subpatterns and covering for S=3 is shown in Fig.1.

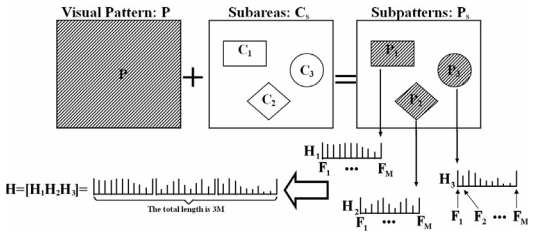


Figure 1. Example of a visual pattern P, covering C and histogram formation.

Decomposition of the pattern into covering brings structural information into description. This can be seen by considering decompositions where the pattern area is partitioned into increasing number of disjoint subareas. Each subarea carry then increasing amount of structural information

related to the location of features. In the extreme case, the subarea is just a single feature which is then located with maximal precision. In our framework the subarea is described by statistical information, i.e. the histogram, but the number of subareas, their size and shape reflects structural information. There is no structural information associated with the histogram calculated over the whole pattern area. This framework allows considering a problem of minimal structure information.

III. Feature Selection And Structural Information

The feature set selected, the length of the feature histograms and the covering plays an important role in the retrieval process. The goal for using these histograms is to allow efficiently comparison between different patterns. The histograms themselves have to show enough discriminating ability between each other. If the length is too short, we cannot achieve enough discrimination; while if the length is too long, the retrieval will suffer from the heavy computation, the redundant information and the possible noise. Therefore, only a judicious selection of features may lead to a good trade-off between accuracy and efficiency.

In this paper, we fulfill this requirement by statistically evaluating the a priori probability of the appearance of each feature. Those features which are less likely to be present in most of the images are considered as redundant features, thus are removed from the histograms. The feature selection is conducted by a training process which will be described later. Given a certain area, there are numerous ways to define subareas in order to cover it. However, not every distribution of subareas can lead to good retrieval result. The distribution should properly fit the inherent visual property of the target pattern. Therefore, an optimization is required when defining the subareas. In our method, this is done by evaluating the a priori performances of some randomly selected subareas. The histograms from subareas which provide good results can be further combined to provide better performance.

We show this idea by using the face image as an example. One complete face image can be seen as a combination of the following regions: forehead, eyes, nose, mouth, chin and rest area of the face. Using these subareas one can encode a certain amount of structural information. Theoretically, the number of possible distribution of subareas is unlimited. It seems to be that the searching process of optimal subareas is practically an endless process. However, our experimental results show that even with two selected rectangular subareas, we can already achieve an excellent performance. Furthermore, we also proposed a method allowing fast searching.

IV. Feature Histogram Generated by Block Transform and Quantization

A. Block Transform

The Block Transform Coding methods are widely used in image and multimedia compression. The transform coding method compact image data by representing the original signal with a small number of transform coefficients. Such kind of ability to re-arrange the information in a suitable way can contribute not only to the compression, but also to the recognition. Some researchers have already utilized it in their face image retrieval works.

The transform used in this research is taken from the H.264 standard [6]. This transform is a 4x4 integer transform, which is originally aim to encode the coefficients of inter blocks. Overall, this transform performs in a similar way with the widely-used Discrete Cosine Transform (DCT). However, the ability of integer calculation allows rapid process.

The forward transform matrix of the 4x4 integer block transform is B_f and the inverse transform matrix is B_i.

$$B_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix} \qquad B_f = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix}$$

B. Quantization

The reason why the above block transforms are used stems from their robustness to quantization. Quantization of transform coefficients by special matrices or scalars as in the 4x4 block transform preserves to very high degree important perceptual features. As the energy is mostly distributed around the DC coefficient, quantization can make most of the AC coefficients to zero. The remaining non-zero AC coefficients indicate the existence of major textural information in this block area. While the removed (become zero) AC coefficients are usually perceptually redundant information.

During the process when the applied Quantization Scalar (QS) is being increased, both the useful and redundant information are gradually removed. Fortunately, the speeds of these two reductions are different. There is an optimal area within which the redundant information is largely removed, while the major part of the useful information is still well preserved.

C. Feature Vector

The above block coefficients inherently contain the structural information of the corresponding block. In the following step, the Ternary Feature Vector (TFV) described in [1] is used to further collect the information from a larger area. In our case here, this larger area is the area covering nine neighboring blocks. That also means one TFV actually represents a part of information coming from neighboring 12x12 pixel block.

The basic unit for generating a TFV is also a 3x3 coefficient matrix. The TFV is formed from the same coefficient all over the subarea. The formation is done by thresholding the eight neighboring coefficients with two thresholds. Each block can produce one ternary vector with length of eight. This 8-bin vector is subsequently converted to a decimal number in the range of [0, 6560]. The thresholds are defined in a flexible way based on the local information. Within each 3x3 matrix, assuming the max value is MAX, the min value is MIN, the mean value is MEAN, the threshold is calculated by:

$$Threshold_{-}^{+} = MEAN \pm (MAX - MIN) \times f \qquad (1)$$

where f is a real number within the range of (0, 0.5). Our initial experiments have shown that performance with changing f has broad plateau for f in the range of 0.2~0.4. From this reason, we use f = 0.3 in this paper. The thresholded values can be either 0, 1 or 2

$$\begin{aligned} 0 &- \text{The neighboring pixel value} \le \text{Threshold}_- \\ 1 &- \text{The neighboring pixel value} \quad \text{otherwise} \\ 2 &- \text{The neighboring pixel value} \ge \text{Threshold}^+ \end{aligned}$$

D. Feature Histograms

Histogram of TFV vectors form one subarea may have in general 6561 bins and is normalized by the overall number of bins in the area. Not all bins are significant for the retrieval and their numbers can be significantly reduced in the training process. All the bins are sorted by their corresponding probabilities of appearance. A set of most common bins are used for retrieval. Based on the result of training process, we selected one AC coefficient to be combined with the DC coefficient, to form a combined TFV histogram. Furthermore, the histograms from different subareas are concatenated together to enhance the retrieval performance.

Since there are multiple well-known similarity measures which are widely used by the pattern recognition society, we did an experiment with the training data. The result shows that the city-block (L1 norm) distance is the most suitable one for our system. It fit well with the intrinsic property of our data, and its computational complexity is low. Therefore, we use it for the following experiments.

V. Experiments With FERET Database

A. Experimental System

For testing the performance of the proposed method we use the FERET face mage database. The advantage of using this database is its standardized evaluation method of based on performance statistics reported as Cumulative Match Scores (CMS) [7]. CMS is curve which shows the ascending variation of correct retrieval as a function of N, where N is the number of images displayed in the first page of returned results. Horizontal axis of the graph is retrieval rank and the vertical axis is the percentage of correct matches. On the CMS plot, higher curve reflects better performance. For simplicity, many researchers use the CMS at the first rank (N=1) to represent the overall performance. We refer to it as "Rank-1 CMS".

Using face images as an example here does not mean our method is limited to the application of face image retrieval; it also has the potentiality to be applied to other image retrieval tasks. There are several releases of FERET database. The release which we are using is the one published at October 2003, called Color FERET Database. The Color FERET Database contains overall more than 10,000 images from more than 1000 individuals taken in largely varying circumstances. Among them, the standardized FA and FB sets are used here. FA set contains 994 images from 994 different objects, FB contains 992 images. FA serves as the gallery set, while FB serves as the probe set.

Before the experiments, all the source images are cropped to contain face and a little background. They are normalized to have the same size. The eyes are located in the similar position according to the given information from FERET. Simple histogram normalization is taken to the entire image to tackle the luminance change.

To ensure the independence of training set, five different groups of images are randomly selected to be the training sets. Each group contains 50 image pairs (from 992 pairs). Five parameter sets are obtained from them, which will be applied for evaluation of performance of the rest 942 images of the whole database. The resulting five CMS curves are averaged, which is the final performance result (Fig.2).

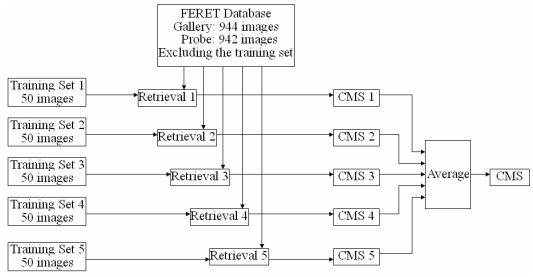


Figure 2. Training and Retrieval process.

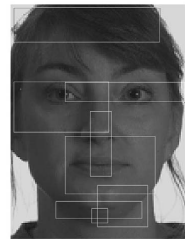


Figure 3. Some examples of subareas in a face image.

B. Experiments

The first experiment is conducted over randomly selected 512 subareas. Some examples are shown in Fig.3. Rather than manually select the eye or noise areas, we randomly selected 512 subareas. If they are overlapped together, the whole image will be covered. There is a wide dynamic range of the sizes of these subareas: the smallest one is about 1/150 of the size of whole image, while the largest one is about 1/5. Both the large and small subareas are selected from every region of the face image. Since we have five training sets, the final result is actually a matrix of 5x512 CMS scores. They are further averaged to be a 1x512 CMS vector. The maximum, minimum and average of these 512 CMS scores are shown in Table I. One can see from it that there is very wide performance variation for different subareas. The selection of subarea is thus critical for the performance which can be achieved. However, this is also reasonable since there are certain subareas which are too small to provide enough information.

Based on above results, a reasonable way to improve the performance is to combine multiple subareas. In the following experiment, we utilized 216 pairs of subarea for retrieval. Most of these selected subareas can provide relatively better performance in the previous experiment. In addition, when any two subareas are forming a pair, they must come from different regions of the face image. For example: if one is from eye region, the other could be from mouth region or chin region. The retrieval task is performed over them and the best pairs are identified. The resulted scores for 2-subarea case are also shown in Table I. Clearly we can see the improvement from 1-subarea to 2-subarea

TABLE I. THE RANK-1 CMS RESULTS OF USING SINGLE/TWO SUBAREA.

Rank-1 CMS (%)	Maximum	Minimum	Average
1-subarea	95.30	12.94	62.11
2-subarea	97.70	52.50	82.56

C. *Fast Searching for Optimal subarea*

From the above results one can see that: by combining subarea histograms it is possible to achieve good retrieval performance when the proper subareas are selected. However, the number of possible subareas is virtually unlimited which makes searching for the optimal ones rather tedious. In order to speed up the search procedure, while at the same time keeping the similar performance, we applied here a three-step searching method over the training sets. This method is matched to the face image structure where it can be expected that horizontal rectangular subareas will have highly informative structure. The searching procedure is thus as follows:

1. Rectangular areas covering the width of images with different height are considered in the first step. For example, in our experiments with images of size 412x556 pixels, the height of areas is ranging from 40 to 160 pixels, with the width fixed at 400 pixels. The rectangular areas are swept over the picture height in steps of 40 pixels, as shown in Fig.4. From here we have 32 subareas, which is a small subset of above 512 subareas. The subarea giving best result is selected as the candidate for the next step.
2. The vertical position of the above candidate is fixed and now its width is changed. A number of widths are tested with the training data set and the one with best performance is selected. Here the number of tested widths is 16. After this, the subarea giving best result is selected as the candidate for the next step.
3. Searching is performed within the small surrounding area of the above candidate. The one giving best result is selected as the final optimal subarea.

Figure 4. Example subareas from the 1st step of searching.

The results from the three-step searching are shown in Table II. The three-step searching method saves a lot of time in searching process, the difference between CMS performance mostly is less than one percent, which is a very good result considering large savings in the computation and the small size of the training set.

TABLE II. THE RANK-1 CMS RESULTS OF USING FAST SEARCHING

	1-subarea	2-subarea
Rank-1 CMS (%)	94.70	96.31

D. *Comparison With Other Methods*

In order to compare the performance of our system with other methods, we list below some results from other researchers. Table III lists the results from the same release of our FERET database. They are all using FA and FB sets. The results are expressed in the way of Rank-1 CMS score.

TABLE III. THE RANK-1 CMS RESULTS OF USING FAST SEARCHING

References	[8]	[9]	[10]	[5]	2-subarea Fast Searching
Rank-1 CMS (%)	79.4	60.2	73.08	97.9	96.31

VI. CONCLUSIONS

In this paper we investigated a framework for image database retrieval problem which combines statistical and structural information by using feature histograms computed over pattern coverings. Our features are selected as quantized block transform coefficients which guarantees their high perceptual relevancy with low complexity of description. Structural information of patterns in our approach is described by subsets included in the covering and can be minimized by reducing the number of subsets. A fast searching method is proposed to find out the optimal subareas. We illustrate our approach on the example of FERET image database. We show that by using just two subareas and the fast searching method, our approach gives good retrieval accuracy which is comparable with other methods.

REFERENCES

[1] Z. Daidi, D. Irek, "Facial Features Detection By Coefficient Distribution Map". *Proceeding of CAIP*, 2005.

[2] FERET Face Database, 2003. Available at: http://www.itl.nist.gov/iad/humanid/feret/

[3] T. Ahonen, A. Hadid, M. Pietikainen, "Face recognition with local binary patterns". *Proceeding of ECCV*, 2004.

[4] A.N. Rajagopalan, R.K. Srinivasa, K.Y. Anoop, "Face recognition using multiple facial features". *Pattern Recognition Letters*, vol. 28, 2007

[5] C.H. Kim, J.Y. Oh, O. Jiyong, C.H. Choi, "Combined Subspace Method Using Global and Local Features for Face Recognition". *Proceeding of IJCNN*, 2005.

[6] ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC: Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (2003).

[7] P.J. Phillips, H. Moon, P.J. Rauss, S. Rizvi, "The FERET evaluation methodology for face recognition algorithms". *IEEE Pattern Analysis and Machine Intelligence*, vol. 22, 10, 2000.

[8] J. Shi, A. Samal, D. Marx, "Face Recognition Using Landmark-Based Bidimensional Regression". *Proceeding of ICDM*, 2005.

[9] J. Shi, A. Samal, D. Marx, "How effective are Landmarks and Their Geometry for Face Recognition". *Computer Vision and Image Understanding*, wol. 102, 2, 117-133 2006

[10] J. Roure, Z.M. Faundez, "Face recognition with small and large size databases". *Proceeding of ICCST*, 2005.

Publication XI

DaiDi Zhong, Irek Defée, "Location Detection of Face Features by DCT Coefficients", in Proceedings of Visualization, Imaging, and Image Processing 2005 (VIIP2005), pp. 99-103, September 2005.

Proceedings of the Fifth IASTED International Conference
VISUALIZATION, IMAGING, AND IMAGE PROCESSING
September 7-9, 2005, Benidorm, Spain

LOCATION DETECTION OF FACE FEATURES BY DCT COEFFICIENTS

Daidi.Zhong, Irek.Defée
Institute of Signal Processing, Tampere University of Technology
P.O.Box 553, FIN-33101 Tampere
Finland
daidi.zhong@tut.fi, irek.defee@tut.fi

ABSTRACT
Images and video are currently predominantly handled in compressed form. Block-based compression standards are by far the most widespread. It is thus important to devise information processing methods operating directly in compressed domain. In this paper we investigate this possibility on the example of simple facial feature extraction method based on the Discrete Cosine Transform (DCT) coefficients. According to our experiments, most horizontal information of face images is mainly distributed over some key features. After applying block transform and quantization to the face images, such significant information become compact and obvious. Therefore, by evaluating the energy of the specific coefficients which are representing the horizontal information, we can locate the key features on the face. The approach is tested on FERET data-base of face images and good results is provided despite its simplicity.

KEY WORDS
DCT Transform, Quantization, Feature extraction, Facial feature, Compressed-domain

1. Introduction

Facial features detection is nowadays a classical area with huge body of knowledge which has been collected over the years. It is defined as the process of locating specific points or contours in a given facial image. Human face and its feature detection is much significant in various applications as human face identification, virtual human face synthesis, and MPEG-4 based human face model coding [1]. Many research works have been conducted over this topic [2], [3]. They extract the key information from the pixel domain, by using some ideal mathematic models, or by using color information.

The features detection is a highly overdimensioned problem which is seen easily if one would try to consider images as matrices in NxN space. Only extremely limited sets of such matrices carry useful information. Therefore, it is advisable to extract the key features by highly effective pre-processing to limit the amount of input information in the first place.

Currently great majority of pictures and video are available in compressed form with compression based on block transform. Compression has a goal of minimizing the amount of information while preserving perceptual properties and this goal is fully compatible with and desirable for pattern recognition and feature extraction. The problem is – how to utilize the efficiency from compression to benefit the feature extraction task, in order to achieve best extraction results? Indeed one could think that elimination of perceptually redundant information should be very beneficial for the efficiency of feature extraction process. Our previous work [4] related to this topic was trying to extract the feature information from DCT domain, by Block Density Matching method.

In this paper, a novel features detection method based on information extracted from DCT coefficients is proposed. First, the 4x4 DCT transform is utilized to make the energy compact. Second, the quantization and luminance normalization are per-formed to further control the precision of the information extraction. One should notice that this step is optional. Third, the most significant coefficients are selected and threshed in specific bin positions. Finally, some detection procedures are performed with some prior geographical knowledge about the features on the human faces. The locations of the eyes, mouth and nose are detected. The example results are shown based on the some face images from the well-known public face recognition database – FERET [6]. The proposed methods can achieve a good result with low computation complexity.

2. 4X4 DCT Transform and Quantization

Discrete Cosine Transform (DCT) is well-known from its excellent perceptual feature preservation, which is widely used in image and video compression, such as JPEG [7], MPEG [1] and H.264 [5]. One can thus expect that the DCT coefficients will have good visual information extraction properties.

The 2-D DCT can be calculated directly by:

$$G(m,n) = a(m)a(n)\sum_{i=0}^{M-1}\sum_{k=0}^{N-1} g(i,k)\cos\left[\frac{\pi(2i+1)m}{2M}\right]\cos\left[\frac{\pi(2k+1)n}{2N}\right]$$
$$0 \le m \le M-1, 0 \le n \le N-1$$
$$a(m) = \begin{cases} 1/\sqrt{M}, m = 0 \\ \sqrt{2/M}, 1 \le m \le M-1 \end{cases}, a(n) = \begin{cases} 1/\sqrt{N}, n = 0 \\ \sqrt{2/N}, 1 \le n \le N-1 \end{cases} \quad (1)$$

Here, g is the source block and the G is the DCT transformed block. N is the dimension of the blocks.

The first uppermost coefficient after transform is called DC and it corresponds to average light intensity level of a block, other coefficients are called AC and they correspond to components of different frequencies. The AC coefficients provide us some useful information about the texture detail of this block. In the standards which are mentioned above, the block coefficients are quantized after DCT transform, in order to remove the redundant coefficients and achieve compression efficiency. However, this process also introduces some difference between the quantized image and the original one. Calculations of the DCT blocks can be also done with block overlap which was used in this paper. As the energy is mostly presented at the low-frequency part, quantization can make most of the high-frequency coefficients to zero. However, from the feature detection point of view, using the whole AC information seems to be redundant. Therefore, a quantization factor (QF) is used to scale down each coefficient during the subsequent quantization process. After the quantization, the remaining high-frequency coefficients, which are non-zero, indicate the existence of a strong edge in this block area.

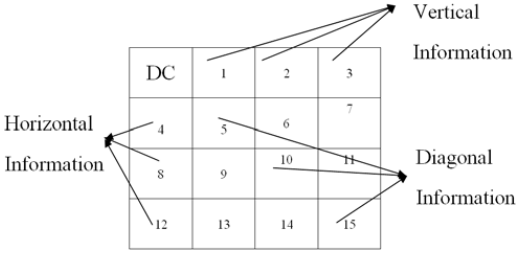


Fig. 1. Directional information represented by different coefficient

Furthermore, coefficients in different bin positions are representing different directional information. Given a 4x4 transformed block:

1. The AC coefficient in first line are corresponding to vertical information
2. The AC coefficient in first column are corresponding to horizontal information
3. The AC coefficient in diagonal direction are corresponding to diagonal information

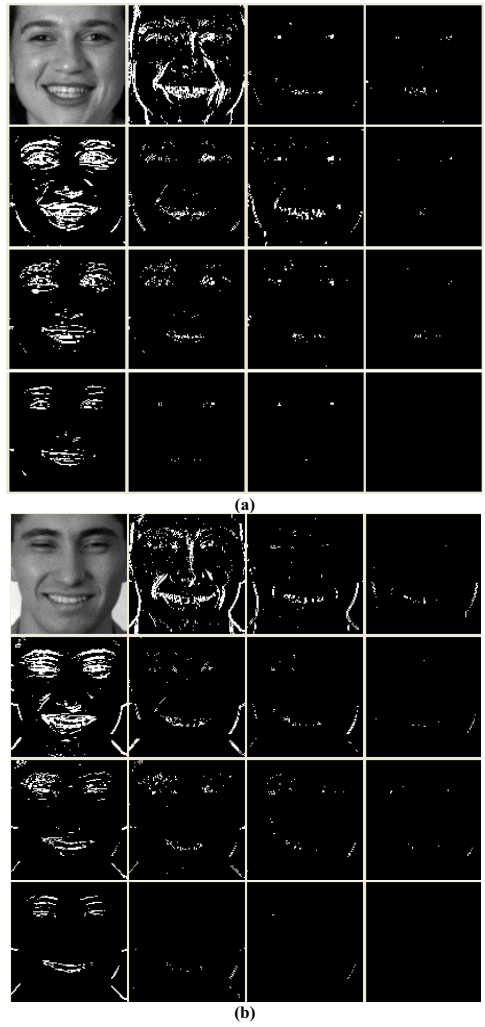


Fig. 2. Coefficient distribution map (QF=20)

This can be known from Fig. 2, which shows the energy distribution of these 15 AC coefficients (when the quantization factor is 20), for two example face images. We call it Coefficient Distribution Map (CDM). All the coefficients are binarized into to zero and non-zero. Non-zero points are the white points in Fig.3. As we can see, after quantization, some coefficients are mostly distributed and compact around key features, such as mouth, eyes and nose. Two good examples are the 8th and 12th coefficient according to the order in Fig. 1. Based on above observation, one may think to detect the facial features according to the distribution of these coefficients.

3. Luminance Normalization

The overall luminance condition has direct effect on the final detection performance. Same quantization will produce different coefficients from a scene taken at low luminance than from the same scene at higher luminance. To eliminate this impact, we normalize the luminance of images by rescaling the coefficients according to the average luminance level. The average luminance level is calculated based on the DC coefficients of the transformed blocks.

Assume there are N transformed blocks in an image j, and the DC value for each block is denoted by $DC_i(j)$, $1 \leqslant i \leqslant N$. From these DC values, we can calculate the mean DC value for this image.

$$DC_{mean}(j) = \frac{1}{N}\sum_{i=1}^{N} DC_i(j) \qquad (2)$$

Next, in similar way the average luminance DC_{all} of all images in a database is calculated based on (3). The ratio of luminance rescaling for image j is calculated through:

$$R = \frac{DC_{all}}{DC_{mean}(j)} \qquad (3)$$

Next the, AC coefficients of a block are rescaled by

$$\overline{DCT_{i,j}} = DCT_{i,j} \times R, \quad 1 \le i \le N, 1 \le j \le M \qquad (4)$$

After normalization, all the coefficients are then quantized by a quantization coefficient QF

$$\overline{\overline{DCT_{i,j}}} = \frac{\overline{DCT_{i,j}}}{QF}, \quad 1 \le i \le N, 1 \le j \le M \qquad (5)$$

We found that system performance is not too sensitive to the exact value of re-scaling so whenever images are of perceptually tolerable quality (not strongly under- or overexposed) the rescaling works well.

4. Feature Detection

In order to detect these key features, a small block is moved on the binarized Coefficient Distribution Map and the sum of non-zero coefficients is calculated and displayed as a histogram. After that, the peaks of histogram are detected which indicate the position of features. In order to keep the most important information, while re-moving the irrelevant information, the coefficients are binarized according to a thresholded. On the other hand, different coefficients can be used to generate the CDM. Through our test, we found that the horizontal information is more robust than vertical information for detection. And the 12th AC coefficient is especially robust than others, when detecting nose and mouth. In our experiments, we use different CDM to detect the eyes, but only use the 12th AC coefficient to detect the mouth and nose.

Fig. 3 is an example of using the 12th AC coefficient to detect the feature.

1. (b) is obtained by applying a larger threshold to the 12th AC coefficients. This threshold is set to 2/3 of the maximum value of 12th AC coefficients. The number of non-zero coefficients (after threshold) are summed, first horizontally, then vertically, as shown in (a) and (b). The rough locations of eyes are detected.
2. We evaluate the small block around these rough locations, using another threshold to keep the blocks with darkest DC values. The black color shows the locations of eyeballs. Finally, the location parameters are obtained from these black points. This process is shown in (c)
3. A rough location between nose and mouth can be obtained from the locations of left and right eye. They are forming an equilateral triangle. We will search the area arounding this point. The width of this searching window is the horizontal distance between the eyes. This area is shown in (d) and (e).
4. (e) is also obtained from the 12th AC coefficient, but the threshold is set to 1. This is because the eye areas usually contain the largest horizontal energy, while the nose and mouth areas contain smaller energy.
5. A similar way to Step1 is performed over (e) and the peaks of histograms indicate the vertical positions of nose and mouth. Presuming that the position of nose and mouth is in the middle of eyes, we can calculate the horizontal positions of them.

Above detection method is tested over 360 images from a public face recognition database – FERET. These images are the first 360 images of the FERET database, without glasses. They have different size, different luminance condition and other properties. The width of these images is ranging from 36 to 124; the height of these images is ranging from 37 to 152. The faces inside have different expressions. Image are quantized at either QF=1 or QF=20. Different combinations are used to form the CDMs. The results with and without step2 are compared. The results with and with-out normalization are also compared. The best correct detection ratio is 95.83%, which is achieved when QF=1, with Step2 and with normalization.

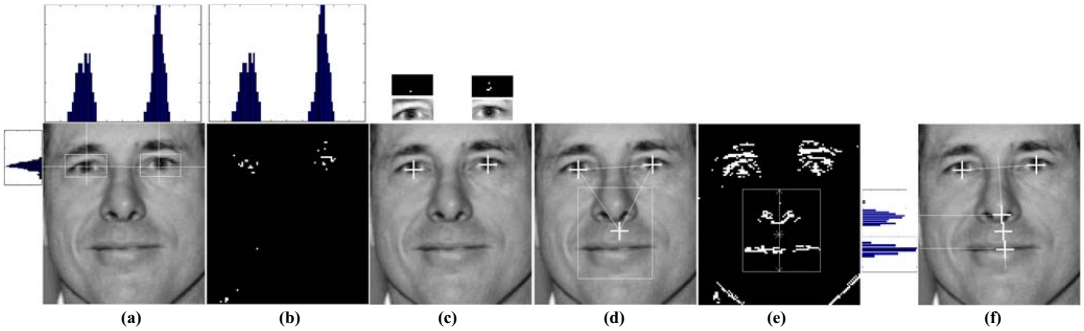


Fig. 3. Feature detection process

False Detections	With Step2				Without Step2			
	Normalization		Non-Normalization		Normalization		Non-Normalization	
AC Coefficents	QP=1	QP=20	QP=1	QP=20	QP=1	QP=20	QP=1	QP=20
12	39	51	33	64	36	33	37	59
12+8	18	33	19	47	31	31	43	61
12+8+10	15	30	18	46	23	31	33	59
12+8+10+2+3	26	42	28	45	28	34	27	50

Table 1. Detection results -- false detections among 360 detections

From the results of Table 1 we can find out:

The results of Normalization is better than Non-Normalization

1. The results with Step2 is better than without Step2
2. The results of QF=1 is better than QF=20
3. The result of (12+8) is pretty much similar to (12+8+10), and they are better than (12) and (12+8+10+2+3).

Since quantization removes some useful information, it is reasonable that the results of QF=1 is better than QF=20. However, in practice, the input images are often in the compressed format, therefore, the results of QF=20 are more meaningful. As we can see, if we use normalization, QF=20 and (12+8)/(12+8+10), the correct detection ratio are always around 91%.

Of course, since such detections are based on blocks, it may be less precise than the detection result from pixel-domain. However, for some application which only requires less precision and need to handle the image in compressed domain, our method is still a good choice. It can also serve as a pre-process step for pixel-based detection. Furthermore, one should also notice that no color information is used here. One may also noticed that some faces with dense beard or exaggerated expression or heavy rotation may are likely to have poor detection results.

Some example results are shown in Fig. 4. (h) and (i) are regarded as false detections.

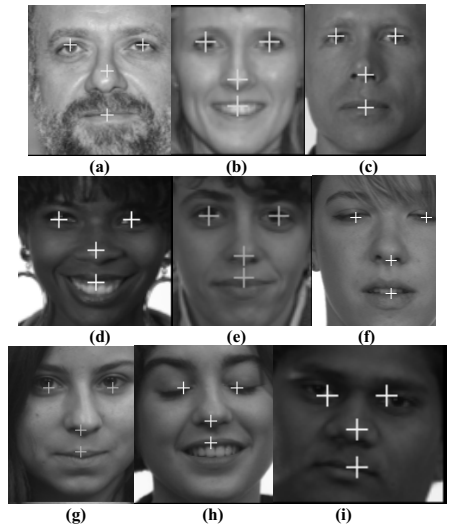


Fig. 4. Some example detection results

5. Conclusions

In this paper, it is shown that facial feature detection using the Coefficients Distribution Map in compressed domain can provide a good performance. The 4x4 DCT block transform is used to extract the energy which is representing the key features. Some prior geographical knowledge about the features on the human faces is used to evaluate the coefficients, in order to detect the positions

of key features. Such method is carried directly in compressed-domain, which requires low computation. Further-more, no color information is used in this process. In the future works, this method is expected to be used. Such structural information, combined with statistical information, is expected to provide good performance in the future works of face image retrieval in compressed-domain.

References:

[1] JTC1/SC29/WG11, MPEG-4, Final Draft of International Standard, Part 2 (Visual), *Doc. No. N2502 of ISO 14496-1*, 1998.

[2] Jun, M., Wen, G., Yiqing C., Jie L., Gravity-Center Template Based Human Face Feature Detection. *Proc. ICMI'2000,* Beijing, 2000, 207-214.

[3] Jörgen A., Facial Feature Extraction using Eigenspaces and Deformable Graphs, *Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging*, 1999.

[4] Daidi. Z., Defée. I., Pattern recognition by grouping areas in DCT compressed images, *Proceedings of the 6th Nordic Signal Processing Symposium*, NORSIG 2004, Finland, 2004.

[5] Joint Video Team of ITU-T and ISO/IEC JTC 1, Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC), *JVT-G050*, 2003.

[6] FERET Face Database. Available at: http://www.itl.nist.gov/iad/humanid/feret/.

[7] Gregory K. W., The JPEG still picture compression standard, *Communications of the ACM, Vol. 34*, 1991, 30 - 44

Publication XII

DaiDi Zhong, Irek Defée, "Pattern Recognition by Grouping Areas in DCT Compressed Images", in Proceedings of 6th Nordic Signal Processing Symposium (NORSIG 2004), pp. 312-315, June 2004.

Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

Pattern Recognition by Grouping Areas in DCT Compressed Images

Daidi Zhong, Irek.Defée

Institute of Signal Processing
Tampere University of Technology
P.O. Box 553,
FIN-33101 Tampere, FINLAND.
E-mail: daidi.zhong@tut.fi irek.defee@tut.fi

ABSTRACT

Images and video are almost exclusively handled in compressed formats based on quantized block DCT transform. Information extraction from images and video has been traditionally studied in the pixel domain. At present methods operating in the DCT domain are more natural and required. There is also argument for DCT based information extraction based on efficiency: compressed images preserve perceptually relevant information at greatly reduced size. This means that all perceptually non-relevant information is eliminated which should facilitate information extraction. While there have been some investigations of pattern recognition in compressed domain in the past, in this paper we analyze the problem from the compression and information reduction perspective. Pattern recognition method based on optimized quantization of DCT blocks and density of blocks in regions is introduced and illustrated on the example of face detection and recognition problem.

1. INTRODUCTION

Images and video are handled nowadays to a great extent in compressed formats based on block DCT transforms. This facilitates storage and transmission. Traditional pattern recognition methods are based mostly on processing in the original picture domain, compression is not taken into account. However, lossy compression methods are highly optimized for generating description of images in which highly relevant perceptual information is preserved and all non-relevant information is eliminated. In result the number of bits for perceptual description is minimized. This is potentially very attractive from the pattern recognition point of view since it reduces redundancy from the processing. In particular, the quantized block DCT is known to minimize the number of bits while preserving perceptually important features [13]. On the other hand the block DCT operates by decomposition of local signal in frequency domain and some analogy with the operation of biological visual processing can be drawn [1].

Information extraction using DCT has been studied in the past [1]-[10] using many different techniques and combinations with other methods. While this previous work has shown that information extraction in the DCT is certainly feasible the advantages of DCT from the perceptual compression point of view were not utilized in our opinion.

In this paper we present an approach to information extraction in the DCT domain taking into account intrinsically the compression properties of the DCT. Our basic idea is that compression reduces the number of different blocks in the picture. These blocks can be grouped and classified according to their number and location. We present our ideas on the example of face detection and recognition problem. We show that quantization of DCT face images leads to specific distributions of DCT blocks. Histograms of block distribution have long tails. Certain block patterns appear often and many block patterns are rare. We show that quantization levels can be selected at which there already well-performing retrieval of face images from database is based on histograms only, without any block location information. [11]

We are also showing that quantized blocks are located in such a way that rare blocks are grouped in areas highly relevant to face information (eyes, nose, lips, etc.). Thus, to facilitate recognition one needs to identify areas with high density of blocks. In this paper we present a method for grouping of blocks and detecting areas of blocks. Measurements performed on those areas provide critical information for face pose evaluation and face recognition.

2. FEATURE DESCRIPTION USING DCT

In compression methods 8x8 DCT block transform is used. However, when higher quantization levels only the frequency components in the 4x4 areas of the DCT are nonzero. This is equivalent to 4x4 blocks DCT performed

on scaled down images. For the information extraction we thus use 4x4 DCT blocks. The 2-D DCT can be calculated directly by:

$$G(m,n) = a(m)a(n)\sum_{i=0}^{N-1}\sum_{k=0}^{N-1} g(i,k)\cos\left[\frac{\pi(2i+1)m}{2N}\right]\cos\left[\frac{\pi(2k+1)n}{2N}\right] \quad (1)$$

$$a(0) = \sqrt{\frac{1}{N}} \quad \text{and} \quad a(m) = \sqrt{\frac{2}{N}},\ 1 \le m \le N$$

Here, ***g*** is the source block and the ***G*** is the DCT transformed block. ***N*** is the dimension of the blocks.

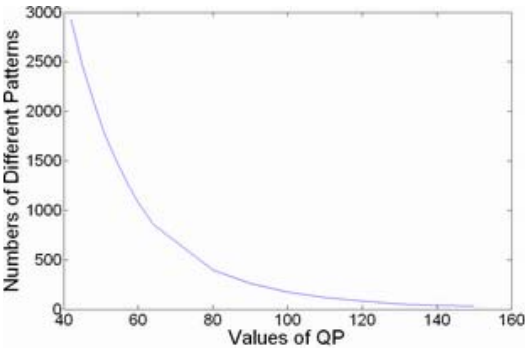


Fig.1 QP vs. Number of Patterns

In compression applications the DCT is quantized. In image compression standards quantization is preformed in very sophisticated way to optimize picture quality, in our approach high quantization levels on the equivalent 4x4 DCT block are used with scalar quantization factor QP similar to the H.264 standard [13]

It can be expected that for certain range of QP values recognition based on the DCT blocks will be facilitated since the number of blocks will be reduced while relevant information will be still preserved. Depending on the QP

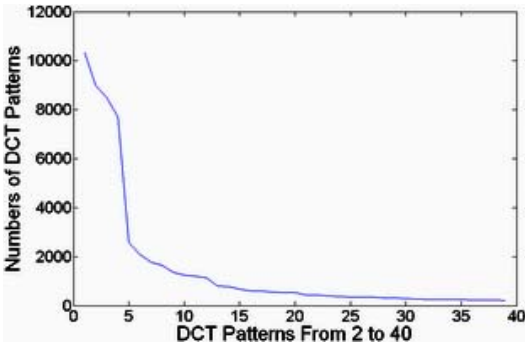


Fig.2 Probability of Patterns (QP=21)

value the number of different blocks is limited as shown in Fig. 1.

Typical block distribution for face image with specific QP factor is shown in Fig. 2. It can be seen that pattern distribution has long tail. There is limited number of patterns which appear in large quantities and significant number of patterns which appear rarely.

Basic observation used in this paper concerns location of the DCT patterns in the images. We split the DCT blocks into two sets: one for that with high quantity and one for those with low quantity. The point of splitting is not very critical. After the splitting, the face images get the following appearance shown in Fig. 3(G). It can be seen that that the set corresponding to rare patterns (black) is distributed over important face features, the set corresponding to common patterns constitute the bulk of the face.

As mentioned before, DCT block patterns can be grouped by evaluating their probabilities of occurrence. Some patterns are common in a particular set of pictures while others are not. Our research shows that, for front face only pictures which have been strongly quantized, the remaining DCT coefficients are most likely to occur at the position near to DC value. Fig. 3 (A) shows two face images. Fig. 3 (B) shows that most of the blocks are DC blocks (The blocks which do not have AC coefficients). These DC blocks are marked as white area in Fig. 3 (B).

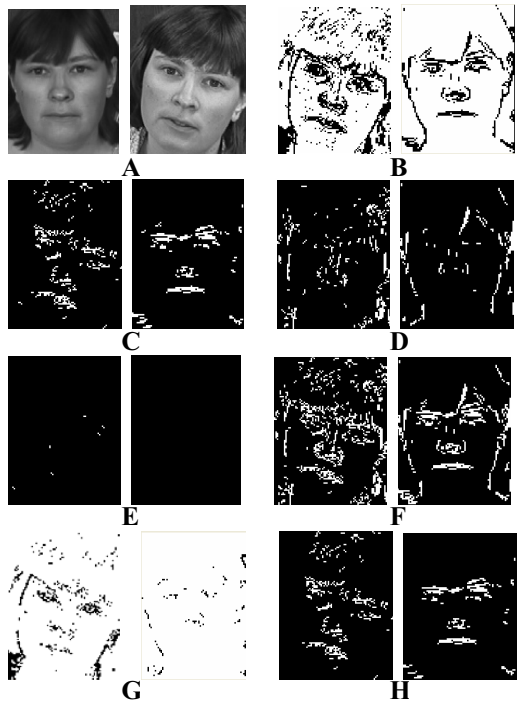


Fig.3 Rare Patterns Depict the Key Features

3. FACE REGIONS

As shown above critical face areas have very specific block distribution. We can classify the blocks further

according to the distribution of their coefficients.

Fig. 3 (C-E) show the positions of the blocks which have AC values only at the position (0, 1), (1, 0) or (1, 1) respectively. Fig. 3 (F) shows the combination of Fig. 3 (C-E). Fig. 3 (G) shows the combination of Fig. 3 (B and F). As one can see, these blocks together make a rough outline of the face. Now it is possible for us to locate the eye, nose and mouth. The areas sets are located in the following way:

1. We form 4x4 DCT transform blocks
2. Perform quantization of the DCT blocks
3. Binarize the blocks coefficients, nonzero coefficients are set to 1

*	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Block Pattern A

*	1	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Block Pattern C

*	0	0	0
1	0	0	0
0	0	0	0
0	0	0	0

Block Pattern B

*	0	0	0
0	1	0	0
0	0	0	0
0	0	0	0

Block Pattern D

Here the notation "*" stands for any DC value.

Fig.4 Most Common Blocks

4. Match each block with the specific block pattern A, B, C or D in Fig. 4, comparing only their AC coefficients. If it is matched, we set a value '1' for this block, namely, the white point; otherwise, we set a value '0' for this block, namely, the black point. Finally, we will get a 1/16 down-scaled binary image.

From these figure, we can conclude that:

1. Most area of the face and hair are DC blocks, which is corresponding to the block Pattern A.
2. Eyes, mouth and bottom line of nose are typically depicted by horizontal AC coefficients, which are corresponding to the block Pattern B. This is shown by Fig. 3(C).
3. Outline of both left and right side of face are typically depicted by vertical AC coefficients, which is corresponding to the block Pattern C. This is shown by Fig. 3(D).
4. For those faces which have been rotated, there will be more count for pattern D. This is shown by Fig. 3(E).
5. A big part of the patterns of eyes, nose and mouth do not belong to the Pattern A, B, C, and D.

4. BLOCK DENSITY MATCHING TO FIND CONNECTIVITY INFORMATION

After the main area of face has been located, locations of particular key features of the face, such as eyes, nose and mouth can be found. We call this Block Connectivity Information. By evaluating the distances and angles between eyes, nose and mouth, one can also find the pose of the face

In order to locate the position of eyes, nose and mouth, here we introduce a "Block Density Matching" method. We start from searching for eyes. As one can see in the Fig. 3(C), the areas of eye, nose and mouth are mainly consisted of white dots, and the density of white point in these areas are higher than the other parts of face. So we can locate them by evaluating the maximum density.

1. We generate a template for the eye, which is a rectangular pixel block. All the pixels are set to 1 (white point).
2. We set the pixels in the four corners to '0' (black points), to make the shape of this polygon similar to the outline of eye. To some degree, this follows the eye oval (Fig. 5).
3. Moving this template as a sliding window in a certain area, matching the area of simple AND operation, we can roughly locate the location of the eyes.
4. The template can be adapted to the size of the face by changing its size. This can slightly improve the accuracy of searching result, while introduce more computation complexity.

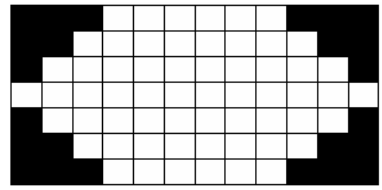


Fig.5 Templates Used for Matching

After we have located the position of eyes, we would next try to locate the position of nose. Here we could use some a priori knowledge to deduce the searching area. For example, for the non-rotated front face, if the center position of left eye, right eye, nose and mouth are respectively at (X_1, Y_1), (X_2, Y_2), (X_3, Y_3) and (X_4, Y_4). Then one can easily draw the conclusions below:

1. $X_2 > X_3, X_4 > X_1; X_3 \approx X_4$
2. $Y_1 \approx Y_2; Y_4 < Y_3 < Y_1, Y_2$
3. X_3 is near the value of $(X_1+X_2)/2$
4. Y_3 is near the value of $(Y_1+Y_2)/2 + (X_2-X_1)x2/3$
5. X_4 is near the value of X_3
6. Y_4 is near the value of $(Y_1+Y_2)/2 + (X_2-X_1)$

After these rough searching areas have been determined, we can perform full-pixel matching, which is similar to previous one. One can finally find out the value of (X_3,

Y_3) and (X_4, Y_4). The template used for searching nose is the same or a little larger width of the one used for eyes; while the template for mouth has 1.5 to 2 times width of the one used for eyes.

Furthermore, if the $X_3 \approx (X_1+X_2)/2$, then we can deduce that the face is at the front position; otherwise, the face maybe horizontally turns to one side. If X_3 is more closer to X_2, then the face turns to left side; if not, the face turns to right side.

This relationship between these key face features, i.e. Block Connectivity Information, is actually what we are looking for. Fig.6 shows an example of the Block Density Matching. The searched positions of eyes and nose notated by two gray lines, which are the top and bottom lines of matching template.

However, it would be more difficult for the case of rotated face. For the rotated face, $Y_1 \neq Y_2$, the two eyes are not in a same horizontal line, but in a diagonal line. The position of nose and mouth need more calculation, but the basic relation remains the same. This is also our future research

From our experiment result we found that, the hair is a major distraction factor. To achieve more accurate result, we should also remove the hair. As we mentioned before, the area of hair is mainly composed of DC block patterns. And the DC value of hair is clearly different from the DC value of face. Therefore, we set a threshold value to filter the compressed image; each DC value which is below the threshold would be set "1". Now the eyes can be easily discriminated. Fig.3 (H) is an example based on Fig.3(C).

5. EXPERIMENTAL SYSTEM AND RESULTS

For experiments we used the Georgia Tech Face (GTF) Database [12]. The experiment result shows that: for most of the front positioned face, we can locate their key features. Figure 6 shows the example of person s26 in the database. The position of eyes, nose and the both side of face are noted by a pair of white lines.

6. CONCLUSION

In this paper, it is shown that quantized DCT and selection of proper pattern set results in very informative description for pattern recognition; the approach of "Block Density Matching" is illustrated based on quantized 4x4 DCT blocks of face database images. By matching the template through pattern indexed images, good results of key features recognition and locating are obtained. The method is computationally efficient and can be used directly for information extraction from compressed video. Further research is needed for dealing with the rotated and turned face images.

Fig.6 Block Density Matching Result

REFERENCES

[1] Yu Zhong and Anil K. Jain, "Object localization using color, texture and shape", Pattern Recognition, Vol.33, No.4, pp.671–684, Apr. 2000

[2] Richard E. Frye, Robert S. Ledley, "Texture discrimination using discrete cosine transformation shift-insensitive (DCTSIS) descriptors", Pattern Recognition, Vol.33, No.10, pp.1585-1598, October 2000

[3] S. Eickeler, S. Muller and G. Rigoll, "Recognition of JPEG Compressed Face Images Based on Statistical Methods", Image and Vision Computing, Vol. 18, No. 4, pp. 279-287, 2000

[4] Ramasubramanian, D. and Y.V. Venkatesh, "Encoding and recognition of faces based on the human visual model and DCT", Pattern Recognition, Vol.34, No.12, pp. 2447-2458, December 2001

[5] Jie Wei, "Image segmentation based on situational DCT descriptors", Pattern Recognition Letters,Vol 23, No.1-3, pp. 295 - 302, January 2002

[6] Jiang J., Armstrong A. and Feng GC "Direct content access and extraction from JPEG compressed images", Pattern Recognition, Vol.35, No.11, pp.2511-2519, November 2002

[7] C. Sanderson and K. K. Paliwal, "Fast Features for Face Authentication Under Illumination Direction Changes", Pattern Recognition Letters, Vol. 24, No.14, pp.2409 - 2419, 2003.

[8] Min-Sub Kim, Daijin Kim and Sang-Youn Lee, "Face recognition using the embedded HMM with second-order block-specific observations, Pattern Recognition, Vol. 36, No. 11, pp.2723-2735, November 2003

[9] GuocanFeng,Jianmin Jiang, "JPEG compressed image retrieval via statistical features", Pattern Recognition, Vol.36, No. 4, pp. 977-985, April 2003

[10] C.Sanderson and K.K.Paliwal, "Features for Robust Face Based Identity Verification". Signal Processing Vol.83, No.5, pp.931-940, May 2003

[11] Koji Kotani, Chen Qiu, and Tadahiro Ohmi, "Face Recognition Using Vector Quantization Histogram Method," in *International Conference on Image Processing*, II-105, Sep. 2002.

[12] Georgia Tech Face Database, Available: ftp://ftp.ee.gatech.edu/pub/users/hayes/facedb/.

[13] Joint Video Team(JVT) of ISO/IEC MPEG&ITU-T VCEG, "Draft ITU-T Recommendation and Final Draft International Standardd Joint Video Specification", Document JVT-G050,March 2003

Publication XIII

DaiDi Zhong, Irek Defée, "A Three-Layer System for Image Retrieval", in Proceedings of International Conference on Signal Processing and Multimedia Applications (SIGMAP 2007), pp. 208-212, July 2007.

A THREE-LAYER SYSTEM FOR IMAGE RETRIEVAL

Daidi Zhong and Irek Defée
Insititute of Signal Processing, Tampere University of Technology, Finland
daidi.zhong@tut.fi, irek.defee@tut.fi

Keywords: Face image, Retrieval, subimage.

Abstract: Visual patterns are composed of basic features forming well-defined structures and/or statistical distributions. Often, they always present simultaneously in visual images. This makes the problem of description and representation of visual patterns complicated. In this paper we proposed a hierarchical retrieval system, which is based on subimages and combinations of feature histograms, to efficiently combine structure and statistical information for retrieval tasks. We illustrate the results on face database retrieval problem. It is shown that proper selection of subimage and feature vectors can significantly improve the performance with minimized complexity.

1 INTRODUCTION

The visual image retrieval is a complex problem since the visual information contains both the statistical and structural information. At one extreme case, the locations of features with respect to each other are critical, this is called structure. At another extreme the statistics of feature distribution is more important than their precise locations. In practice, visual patterns are mixtures of structure and statistics which makes the description problem hard because its complexity looks like unbounded. In addition, the image quality often suffers from the noise and different light conditions, which make the retrieval tasks more difficult.

Some previous works focused on extracting and processing global statistical information by using the whole image (Ekenel and Sankur, 2004), while some other researchers start from some key pixels (Shi et al., 2006) to represent the structural information. Based on their achievement, a reasonable way to further improve the retrieval performance is to extract the visual information in a way like a mixture of statistical and structural information.

In this paper, we illustrate our idea by proposing a retrieval system which is based on subimages and combinations of feature histograms. The experimental results disclose that the usage of subimage and local feature vectors can lead to the combination of statistical and structural information, as well as minimized impact of noise, which finally improve the performance of the approach.

In order to achieve a comparable result, we tested our method over a public benchmark of face image database. The evaluation method of this database has been standardized, which allow us see the change of performance clearly. However, using face images as an example here does not mean our method is limited to the application of face image retrieval; it also has the potentiality to be applied to other image retrieval tasks.

2 TRANSFORM AND QUANTIZATION

Some transforms have been found useful in extracting local visual information from images. Popular transforms include: Gabor Wavelet, Discrete Wavelet Transform, Discrete Cosine Transform (DCT), and Local Steerable Phase. Specially, DCT and Wavelets have already been adopted to the image and video compression standards (ISO/IEC,1999). These transform coefficients inherently contain information about the local area, which cannot be known from individual pixel. We believe that properly applied transforms can improve the performance of retrieval. Block transform strongly eliminate the perceptually non-relevant information and this should be of advantage for the image retrieval tasks. The specific block transform we use was introduced in the H.264 standard (ITU-T, 2003) as particularly effective and simple. The transform matrix of the transform is

denoted as T_f and the inverse transform matrix is denoted as T_i . They are defined as

$$T_f = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix}$$
$$T_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix} \quad (1)$$

A 4x4 image pixel block P can be forward transformed to block C using (2), and the scalar quantization process Q( ) is used to remove the irrelevant information, which will result in quantized version of C, Q(C). For reconstruction purpose, the inverse quantization process $Q^{-1}[\ ]$ is applied to the quantized block Q(C), and the block R is subsequently reconstructed from the inverse-quantized block $Q^{-1}[Q(C)]$, using (3)

$$C = T_f \times P \times T_f^T \quad (2)$$

$$R = T_i^T \times Q^{-1}[Q(C)] \times T_i \quad (3)$$

with superscript T denoting transposition.

The leading element of the matrix C is called the DC coefficient. All other elements are called AC coefficients. There are thus 15 AC coefficients in the matrix H but many of them will have zero value after the quantization Q(C) is applied. The power of the transform stems from the fact that despite of strong quantization, the reconstructed block R will still approximate well the original image block P. Quantization has the effect of limiting the dynamic range of coefficients.

3 FEATURE VECTORS

3.1 DC Ternary Feature Vectors

Block transform and quantization arranged the local information in a suitable way for retrieval. Based on this merit, we utilize the specific feature vector defined below to further group the local information in the neighboring blocks. The grouping process can be applied separately or jointly over DC and AC coefficients for all transform blocks of an image.

Considering a 3x3 block matrix containing nine neighboring blocks, the DC coefficients from them can form a 3x3 coefficient matrix. The eight DC coefficients surrounding the center one can be thresholded to form a ternary vector with length eight. This vector is called DC Ternary Feature Vectors (DC-TFV), which encode the local information based on those quantized transform coefficients.

The threshold is defined as a flexible value related to the mean value of all the nine DC coefficients.

$$Threshold^+ = M + (X - N) \times f$$
$$Threshold\_ = M - (X - N) \times f \quad (4)$$

where f is real number from the interval (0,0.5), X and N are maximum and minimum pixel values in the 3x3 coefficient matrix , and M is the mean value of the coefficients. Our initial experiments have shown that performance with changing f has broad plateau for f in the range of 0.2~0.4. From this reason, we use f = 0.3 in this paper. The thresholded values can be either 0, 1 or 2

If the pixel value $\leq$ Threshold$^+$ put 0
If the pixel value $\geq$ Threshold_ put 2
otherwise put 1

The resulting thresholded vectors of length eight are subsequently converted to decimal numbers in the range of [0, 6560].

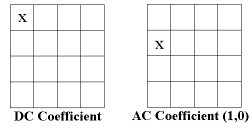


Figure 1: The DC and one AC coefficient are utilized here.

3.2 AC Ternary Feature Vectors

Following the procedure described above, the binary feature vectors are defined for the AC coefficients in the same way by forming 3x3 matrices and thersholding. We denote such vectors as AC Ternary Feature Vectors (AC-TFV). Considering the fact that there are 15 AC coefficients in each 4x4 block, we only use one coefficient here to illustrate our idea in a simple way. The used coefficient is in the position (1,0), which has been shown in Figure 1. Although using more AC coefficients might improve the performance, it also requires more calculation. The proper selection can be conducted with training set. However, we only present the result with one very capable AC coefficient, which already shows good result. For simplicity, the two coefficients shown in Figure 1 can be directly calculated without applying the entire H.264 block transform.

4 REPRESENTATION BASED ON SUBIMAGE

One complete face image can be seen as a combination of different subimages. For example: eyes, nose and mouth, each of these three subimages represents relatively independent key information. Considering them separately may leads to better representation of the image comparing to using the whole image. In our experiments, we divide the original image into several rectangular subimages. Information is extracted from each subimage, and then combined to serve the retrieval tasks. Totally 512 subimages are randomly used in this paper. They can cover almost all the face when overlapped together. Furthermore, the sizes of them vary a lot. The smallest and largest one respectively have 1/150 and 1/5 times of the size of whole image. This is different from the traditional way to select only the mouth and eye areas, since we wish to find out where is the most distinguish area according to training process. Some examples of subimage are shown in Figure 2.

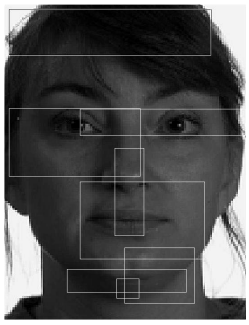


Figure 2: Examples of subimage (each rectangle is a subimage).

5 HISTOGRAMS OF FEATURE VECTORS AND SIMILARITY MEASURE

Our premise is that structural and statistical information should be combined in a graceful way that is allowing smooth and controlled combinations of them. In this paper we consider a step leading into this direction which is done by the histograms of TFV from quantized coefficients. The process of generating histogram is listed below:

1. The 4x4 H.264 AC Block Transform is applied to a subimage.
2. Quantization is applied separately to all the AC and DC coefficients.
3. TFV is generated from certain coefficient.
4. Histogram is generated from this subimage by simply counting the number of each occurring TFV.
5. Histogram is normalized according to the size of subimage.

Specifically for AC-TFV histogram, there is one bin which is too dominant comparing to other bins. This is caused by the smooth area in image and quantization. Such areas will generate a lot of all-one vectors, like [1 1 1 1 1 1 1 1]. Our retrieval does not use this bin, since it decreases the discriminate ability.

Histogram based on DC-TFV and AC-TFV can be used separately or collectively. Since they represent different information, the combination of them can leads to better performance, which will be shown in the following experiment. The combination is done by simply concatenating each histogram one by one. Each histogram may is generated from one subimage, and representing either AC or DC information. Below are three examples of different Combined Histograms (CH) based on two subimages:

[CH1] = [DC-sub1 DC-sub2]
[CH2] = [AC-sub1 AC-sub2]
[CH3] = [DC-sub1 AC-sub1 DC-sub2 AC-sub2]

During the face image retrieval process, the input image is compared to any image stored in the database, in order to find the most similar one. In our method, such similarity is measured by calculating the L1 norm distance (city-block distance) between two histograms. For example, suppose we have two histogram $H_i(b)$ and $H_j(b)$, b= 1, 2, … B. The distance will be calculated as:

$$\text{Distance}(i,j) = \sum_{b=1}^{B} \left| H_i(b) - H_j(b) \right| \qquad (5)$$

6 EXPERIMENTS WITH FERET DATABASE

6.1 FERET Database

The Color FERET Database (FERET, 2003) contains standardized FA and FB sets. FA set contains 994 images from 994 different objects, FB contains 992 images. FA serves as the gallery set, while FB serves as the probe set.

The advantage of using this database is the standardized evaluation method of FERET (Phillips

et.al, 2000) based on performance statistics reported as Cumulative Match Scores (CMS), which are plotted on a graph. The horizontal axis of the graph is retrieval rank and the vertical axis is the probability of identification (PI) (or percentage of correct matches). Simply, a higher curve reflects better performance.

The FERET database provides some tools for preprocessing of the face images. We utilized some of these tools in the preprocessing stage of our evaluation. First, the images were cropped to the same size, which roughly contain the face area. They are subsequently aligned and adjusted by illumination normalization. No mask is applied to the images.

6.2 Training and Retrieval Process

Our image database retrieval problem is formulated as follows. Each probe image from probe set FB has its corresponding image in gallery set FA. We use the feature vector histograms of images and similarity measure defined above to find out the image in FA which gives minimum distance from the probe image. If the found gallery image represents the same person as the probe image, this retrieval will be defined as a correct one.

However, before this can be done the parameters used for the calculation of histograms and similarity measure need to be found using training database set. This set can be selected as a small subset of the database. Knowing the correct responses for the training database allows us to tune the parameters to achieve best retrieval results. The optimal parameter set which will be found out during training process includes: the quantization scalar and length of histogram. The optimal parameter set is identified as the one which is maximizing the retrieval performance over training database. The resulted optimal parameter set is applied to the whole database to evaluate the actual system performance.

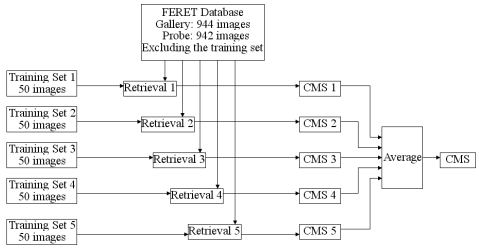


Figure 3: Training process based on five different small sets.

In order to show that the selection of different training set has insignificant impact over final performance, the retrieval process is repeated five times; each time using a different training set containing 50 images, and the remaining 942 images is the testing set. The final CMS curve is the average of the five CMS curves resulted from above five training sets. This process is shown in Figure 3.

6.3 Experiments and Results

We conducted three retrieval tests: A, B and C. They are defined as below. Within each test, performances of histogram based on DC-TFV, AC-TFV and their combinations are evaluated separately.

Test-A: Histograms are generated from the whole image.

Test-B: 512 subimages are randomly defined, covering everywhere of the image. Their sizes are varied a lot. Only one of them is used to generate the histograms.

Test-C: Two of above 512 subimages are used to generate the histograms. The total number of tested combinations is 216. They come from two different areas (eyes, nose and mouth), in another word, they are non-overlapping.

The result of Test-A serves as the reference for the evaluation of the performances of Test-B and Test-C. The corresponding CMS results are shown in Table 1. The Rank-1 CMS is used here to represent the retrieval accuracy (i.e., the CMS at the first rank). On should notice that the performance of DC-TFV has already reached a saturation area, the improvement is relatively small; while significant improvement can be found in the AC-TFV.

Since the subimage is randomly selected and used, we presented the mean of performance of all the subimages or combinations, in order to prevent from any possible bias due to the usage of specific subimage. From here one can see, although the subimages cover less area than the whole image, the performance gets improved. The reason for this is that the division of image emphasizes some key areas containing critical information for retrieval. In addition, based on the block transform, TFV and subimage, the local visual information is efficiently organized by a three-layer hierarchical system. Statistical information is represented by histogram, and involving certain amount of structural information, which finally leads to a good performance.

Table 1: The Rank-1 CMS results of three tests. There are 512 different cases for Test-B, and 216 cases for Test-C. Therefore, to avoid the bias cause by single case, the maximum, minimum and mean of all the 512 cases are shown here. (a) DC-TFV, (b) AC-TFV, (c) Combination of DC- and AC-TFV.

DC-TFV	Max	Min	Mean
Test-A		92.84%	
Test-B	93.77%	9.01%	56.59%
Test-C	97.76%	47.54%	79.06%

(a)

AC-TFV	Max	Min	Mean
Test-A		64.31%	
Test-B	60.77%	1.69%	20.99%
Test-C	81.94%	13.47%	43.89%

(b)

DC-TFV+AC-TFV	Max	Min	Mean
Test-A		93.65%	
Test-B	95.30%	12.94%	62.11%
Test-C	97.70%	52.50%	82.56%

(c)

The achieved result is comparable to others results obtained from exactly the same version (2003) of FERET database, as shown in Table 2. The corresponding references are (Shi et al., 2005), (Shi et al., 2006), (Roure and Faundez, 2005), and (Chung et al., 2005) respectively.

To further justify the robustness of our method, the standard variations of the difference between of the results from five training sets in Test-B are shown in Table 3. The maximum, minimum and mean of 512 cases are listed. They are small enough to be ignored.

Table 2: Referenced results based on release 2003 of FERET.

Reference	[Shi]	[Shi]	[Roure]	[Chung]	Proposed
Rank-1 CMS	79.4%	60.2%	73.08%	97.9%	97.78%

Table 3: Standard variations of difference between five training sets during Test-B. 512 different cases are evaluated here.

Reference	Max	Min	Mean
Standard Variations	2.554%	0.002%	0.323%

7 CONCLUSIONS

We proposed a hierarchical retrieval system based on block transform, TFV and subimage for visual image retrieval. The performance is illustrated using a public face image database. This system achieves good retrieval results due to the fact it efficiently combines the statistical and structural information. Future research will be concentrated on the optimization of the histograms.

ACKNOWLEDGEMENTS

The first author would like to thank for the financial grant from Tampere Graduate School in Information Science and Engineering (TISE).

REFERENCES

Chung, H.K., Jiyong, O., Chong-Ho, C., 2005. Combined Subspace Method Using Global and Local Features for Face Recognition. In *IJCNN 2005*.

Ekenel, H.K., Sankur, B., 2004. Feature selection in the independent component subspace for face recognition. *Pattern Recognition Letter*, 25:1377–1388

FERET Face Database, 2003. Available at: http://www.itl.nist.gov/iad/humanid/feret/.

ISO/IEC 14496-2, 1999. Information Technology - Coding of Audio-Visual Objects - Part 2: Visual.

ITU-T, 2003. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC).

Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S., 2000. The FERET evaluation methodology for face recognition algorithms. *IEEE Pattern Analysis and Machine Intelligence*, Vol. 22, No. 10.

Roure, J., Faundez, Z.M., 2005. Face recognition with small and large size databases. In *ICCST 2005*.

Shi, J., Samal, A., Marx, D., 2005. Face Recognition Using Landmark-Based Bidimensional Regression. In *ICDM 2005*.

Shi, J., Samal, A., Marx, D., 2006. How Effective are Landmarks and Their Geometry for Face Recognition. *Computer Vision and Image Understanding*, 102(2):117-133

Publication XIV

DaiDi Zhong, Irek Defée, "Facial Features Detection by Coefficient Distribution Map", in Proceedings of The 11th International Conference on Computer Analysis of Images and Patterns (CAIP 2005), pp. 822-828, September 2005.

Facial Features Detection by Coefficient Distribution Map

Daidi Zhong and Irek Defée

Institute of Signal Processing, Tampere University of Technology,
P.O.Box 553, FIN-33101 Tampere, Finland
{daidi.zhong, irek.defee}@tut.fi

Abstract. The Images and video are currently predominantly handled in compressed form. Block-based compression standards are by far the most widespread. It is thus important to devise information processing methods operating directly in compressed domain. In this paper we investigate this possibility on the example of simple facial feature extraction method based on the H.264 AC Transformed blocks. According to our experiments, most horizontal information of face images is mainly distributed over some key features. After applying block transform and quantization to the face images, such significant information become compact and obvious. Therefore, by evaluating the energy of the specific coefficients which are representing the horizontal information, we can locate the key features on the face. The approach is tested on FERET database of face images and good results is provided despite its simplicity.

1 Introduction

Facial features detection is nowadays a classical area with a huge amount of knowledge which has been collected over the years. It is defined as the process of locating specific points or contours in a given facial image. Human face and its feature detection is much significant in various applications as human face identification, virtual human face synthesis, and MPEG-4 based human face model coding [1]. Many research works have been conducted over this topic. [2], [3], [4]

The features detection is a highly overdimensioned problem which is seen easily if one would try to consider images as matrices in NxN space. Only extremely limited sets of such matrices carry useful information. Therefore, it is advisable to extract the key features by highly effective preprocessing to limit the amount of input information in the first place.

Currently great majority of pictures and video are available in compressed form with compression based on block transform. Compression has a goal of minimizing the amount of information while preserving perceptual properties. This goal is fully compatible with and desirable for pattern recognition and feature extraction. The problem is – how to utilize the efficiency from compression to benefit the feature extraction task, in order to achieve best extraction results? Indeed one could think that elimination of perceptually redundant information should be very beneficial for the efficiency of feature extraction process. In addition, this topic is also related to our parallel research about extracting the feature information from DCT domain [5].

A. Gagalowicz and W. Philips (Eds.): CAIP 2005, LNCS 3691, pp. 822 – 828, 2005.

In this paper, a novel features detection method based on information extracted from compressed domain is proposed. First, the 4x4 transform from H.264 standard [6] is utilized to remove the redundancy. Second, the quantization and luminance normalization are performed to further control the precision of the information extraction. Third, the most significant coefficients are selected and thresholded in specific bin positions. Finally, some detection procedures are performed with some prior geographical knowledge about the features on the human faces. The example results are shown based on some face images from the well-known public face recognition database – FERET [7]. The proposed methods can achieve a good result with low computation complexity.

2 4X4 H.264 AC Transform and Quantization

The transform we used in this research is introduced from the H.264 standard. This transform is a 4x4 integer transform, which is originally used to encode the coefficients of inter blocks. Overall, this transform performs in a similar way with the widely-used DCT. They can both make the information compact, which greatly facilitates the information extraction. Different from DCT, the integer transform used here allows rapid process.

The first uppermost coefficient after transform is called DC and it corresponds to average light intensity level of a block. Other coefficients are called AC coefficients; they correspond to components of different frequencies. The AC coefficients provide us some useful information about the texture detail of this block. Such information is essential for the following feature detection.

The forward transform matrix of H264 AC Transform is B_f and the inverse transform matrix is B_i.

$$B_f = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix}, \quad B_i = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix}$$

For simplicity, here we removed the '1/2' in the matrix. The 4x4 pixel block P is forward transformed to block H using (1), and block R is subsequently reconstructed from H using (2). The 'T' means linear algebraic transpose here.

$$H = B_f \times P \times B_f^T \tag{1}$$

$$R = B_i^T \times H \times B_i \tag{2}$$

We perform 4x4 H.264 block transforms over more than thousand different blocks, and the results are further averaged. After applying the transform, one could see from Fig. 1(a) that the main energy is distributed around the DC coefficient. Since there are big differences between the values of different coefficients, the natural logarithm is used here to express the data.

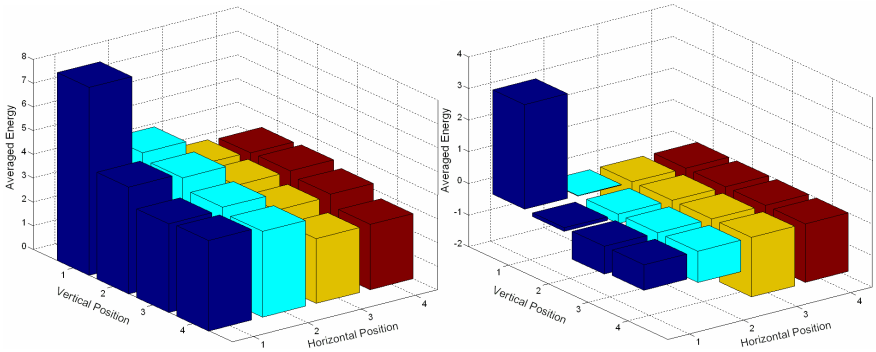


Fig. 1. (a) Natural logarithm of averaged distribution of energy after transform (b) Natural logarithm of averaged distribution of energy after quantization (QF=100)

However, from the feature detection point of view, using the whole AC information seems to be redundant. Therefore, a quantization factor (QF) is used to scale down each coefficient during the subsequent quantization process. As the energy is mostly presented at the upper-left corner, quantization can make most of the high-frequency coefficients to zero. This is shown by Fig. 1(b). After the quantization, the remaining high-frequency coefficients, which are non-zero, indicate the existence of a strong edge in this block area. Through this way, the redundant data is removed and the important data is preserved.

Furthermore, coefficients in different bin positions are representing different directional information. Given a 4x4 transformed block:

1. The AC coefficient in first line are corresponding to vertical information
2. The AC coefficient in first column are corresponding to horizontal information
3. The AC coefficient in diagonal direction are corresponding to diagonal information

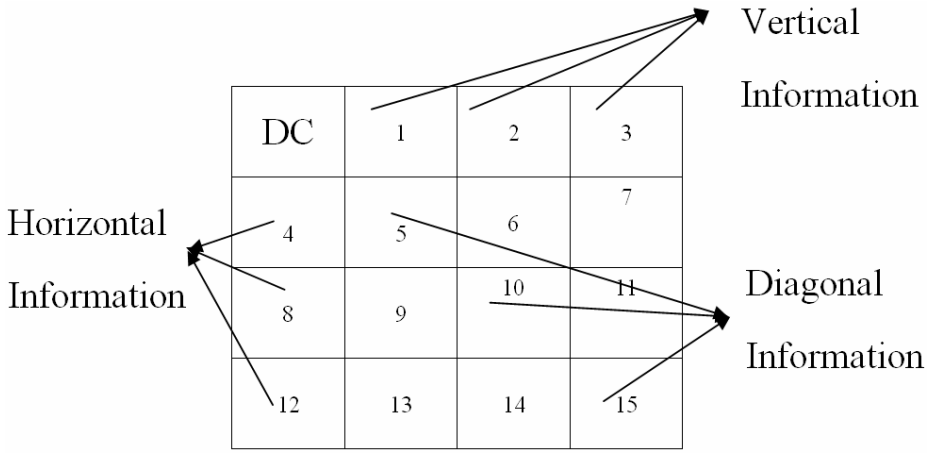


Fig. 2. Directional information represented by different coefficient

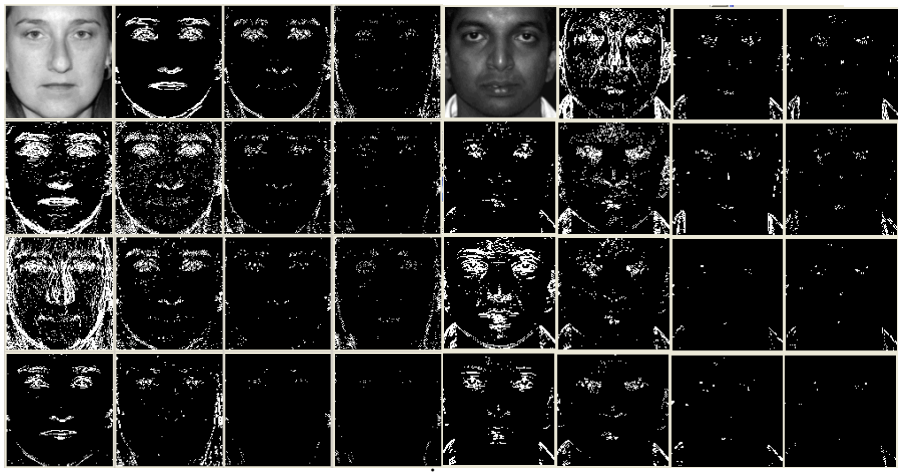


Fig. 3. Coefficient Distribution Map (QF=100)

This can be known from Fig. 3, which shows the energy distribution of these 15 AC coefficients (when the quantization factor is 100), from two example face images. We call it Coefficient Distribution Map (CDM). After quantization, all the coefficients are binarized into zero and non-zero. Non-zero points are the white points in Fig.3. As we can see, after quantization, some coefficients are mostly distributed and compact around key features, such as mouth, eyes and nose. A good example is the 12th coefficient according to the order in Fig. 2. Based on above observation; one may think to detect the facial features according to the distribution of these coefficients.

3 Luminance Normalization

The overall luminance condition has direct effect on the final detection performance. Same quantization will produce different coefficients from a scene taken at low luminance than from the same scene at higher luminance. To eliminate this impact, we normalize the luminance of images by rescaling the coefficients according to the average luminance level. The average luminance level is calculated based on the DC coefficients of the transformed blocks.

Assume there are N transformed blocks in an image j, and the DC value for each block is denoted by $DC_i(j)$, $1 \le i \le N$. From these DC values, we can calculate the mean DC value for this image

$$DC_{mean}(j) = \frac{1}{N}\sum_{i=1}^{N} DC_i(j) \tag{3}$$

Next, in similar way the average luminance $\boldsymbol{DC_{all}}$ of all images in a database is calculated based on (4). The ratio of luminance rescaling for image $\boldsymbol{j}$ is calculated through:

$$R = \frac{DC_{all}}{DC_{mean}(j)} \tag{4}$$

Next the, AC coefficients of a block are rescaled by

$$\overline{AC_{i,j}} = AC_{i,j} \times R, \quad 1 \le i \le N, \quad 1 \le j \le M \tag{5}$$

After normalization, all the coefficients are then quantized by the QF

$$\overline{\overline{AC_{i,j}}} = \frac{\overline{AC_{i,j}}}{QF}, \quad 1 \le i \le N, \quad 1 \le j \le M \tag{6}$$

We found that system performance is not sensitive to the exact value of rescaling so whenever images are of perceptually tolerable quality (not strongly under- or over-exposed) the rescaling works well.

4 Feature Detection

In order to detect these key features, a small block is moved on the binarized images and the sum of non-zero coefficients is calculated and displayed as a histogram. After that, the peak of histograms is detected which indicate the position of features. In order to keep the most important information, while removing the irrelevant information, the coefficients are binarized according to a threshold. On the other hand, different coefficients can be used to generate the CDM. Through our test, we found that the horizontal information is more robust than vertical information for detection, and the 12th AC coefficient is more robust than others.

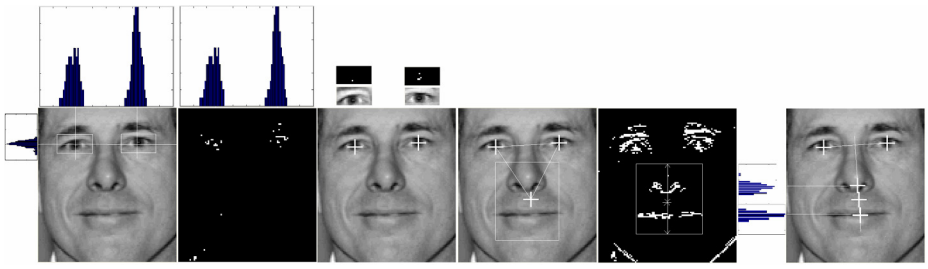


Fig. 4. Feature Detection Process

Fig. 4 is an example of using the 12th AC coefficient to detect the feature.

1. (b) is obtained by applying a larger threshold to the 12th AC coefficients. This threshold is set to 2/3 of the maximum value of 12th AC coefficients. The number of non-zero coefficients (after threshold) are summed, first horizontally, then vertically, as shown in (a) and (b). The rough locations of eyes are detected.

2. We evaluate the small block around these rough locations, using another threshold to keep the blocks with darkest DC values. The black color shows the locations of eyeballs. Finally, the location parameters are obtained from these black points. This process is shown in (c)
3. A rough location between nose and mouth can be obtained from the locations of left and right eye. They are forming an equilateral triangle. We will search the area arounding this point. The width of this searching window is the horizontal distance between the eyes. This area is shown in (d) and (e).
4. (e) is also obtained from the 12^{th} AC coefficient, but the threshold is set to 1. This is because the eye areas usually contain the largest horizontal energy, while the nose and mouth areas contain smaller energy.
5. A similar way to step 1 is performed over (e) and the peaks of histograms indicate the vertical positions of nose and mouth. Presuming that the position of nose and mouth is in the middle of eyes, we can calculate the horizontal positions of them.

Above detection method is tested over 360 images from a public face recognition database – FERET. These images are the first 360 images of the FERET database, without glasses. They have different size, different light condition and other properties. They are quantized at QF=100. The correct detection rate is 91.4%. Some example results are shown as in Fig. 5.

Of course, since such detection is based on blocks, it is less precise than the detection result from pixel-domain. However, for some application which only require less precision, our method is still a good choice. It can also serve as a pre-process step for pixel-based detection. Furthermore, one should also notice that no color information is used here. One may also noticed that some faces with dense beard or exaggerated expression may are likely to have poor detection results, as well as the strongly rotated faces (e.g., Fig.5 (i)).

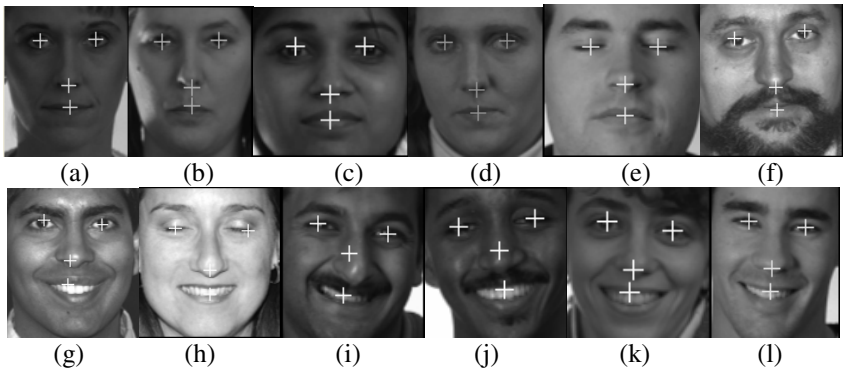


Fig. 5. Some Example Detection Results

5 Conclusions

In this paper, it is shown that facial feature detection using the Coefficients Distribution Map in compressed domain can provide a good performance. The 4x4 H.264 AC

block transform is used to extract the energy which is representing the key features. Some prior geographical knowledge about the features on the human faces is used to evaluate the coefficients, in order to detect the positions of key features. Such method is carried directly in compressed-domain, which requires low computation. Furthermore, no color information is used in this process. In the future works, this method is expected to be used. Such structural information, combined with statistical information, is expected to provide good performance in the future works of face image retrieval in compressed-domain.

References

1. JTC1/SC29/WG11; MPEG-4, Final Draft of International Standard, Part 2 (Visual). Doc. No. N2502 of ISO 14496-1, (1998)
2. Jun, M., Wen, G., Yiqing C., Jie L.: Gravity-Center Template Based Human Face Feature Detection. ICMI'2000. Beijing, (2000) 207-214
3. Saman C., Noel O'C.: Facial Feature Extraction and Principal Component Analysis for Face Detection in Color Images. ICIAR, Lisbon, Portugal, (2004)
4. Jörgen A.: Facial Feature Extraction using Eigenspaces and Deformable Graphs. Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging, (1999)
5. Daidi. Z., Defée. I.; Pattern recognition by grouping areas in DCT compressed images. Proceedings of the 6th Nordic Signal Processing Symposium, NORSIG 2004, Finland (2004)
6. Joint Video Team of ITU-T and ISO/IEC JTC 1; Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC), JVT-G050, (2003)
7. FERET Face Database. Available at: http://www.itl.nist.gov/iad/humanid/feret/.

Publication XV

DaiDi Zhong, Irek Defée, "Face Retrieval Based on Robust Local Features and Statistical-Structural Learning Approach", EURASIP Journal on Advances in Signal Processing Volume 2008, Article ID 631297, 12 pages, doi:10.1155/2008/631297, 2008.

Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 631297, 12 pages
doi:10.1155/2008/631297

Research Article

Face Retrieval Based on Robust Local Features and Statistical-Structural Learning Approach

Daidi Zhong and Irek Defée

Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland

Correspondence should be addressed to Irek Defée, irek.defee@tut.fi

Received 30 September 2007; Revised 15 January 2008; Accepted 17 March 2008

Recommended by Sébastien Lefèvre

A framework for the unification of statistical and structural information for pattern retrieval based on local feature sets is presented. We use local features constructed from coefficients of quantized block transforms borrowed from video compression which robustly preserving perceptual information under quantization. We then describe statistical information of patterns by histograms of the local features treated as vectors and similarity measure. We show how a pattern retrieval system based on the feature histograms can be optimized in a training process for the best performance. Next, we incorporate structural information description for patterns by considering decomposition of patterns into subareas and considering their feature histograms and their combinations by vectors and similarity measure for retrieval. This description of patterns allows flexible varying of the amount of statistical and structural information; it can also be used with training process to optimize the retrieval performance. The novelty of the presented method is in the integration of information contributed by local features, by statistics of feature distribution, and by controlled inclusion of structural information which are combined into a retrieval system whose parameters at all levels can be adjusted by training which selects contribution of each type of information best for the overall retrieval performance. The proposed framework is investigated in experiments using face databases for which standardized test sets and evaluation procedures exist. Results obtained are compared to other methods and shown to be better than for most other approaches.

1. INTRODUCTION

Visual patterns are considered to be composed of local features distributed within the image plane. Complexity of patterns may be virtually unlimited and arises from the size of the local feature set and location of the features. Two aspects of feature locations are worth emphasizing from the description point of view, structural and statistical. The structural aspect is concerned with precise locations of features, reflecting geometry of patterns. Statistical aspect concerns feature distribution statistics. The statistics plays a descriptive role especially for very complex patterns in which there are too many features for explicit description. In real world, the combination of structural and statistical may provide effective description and thus, for example, a leafy tree is described by the structure of a trunk and branches and statistics of features composing leafs. There has been enormous number of studies in the pattern recognition and machine learning areas on how to deal with the complexity of patterns and develop effective methods for handling them, as summarized in a substantial recent monograph [1]. The approach presented in this paper is conceptually different in dealing both with local features and combination with global description within a unified framework of performance optimization via training.

While the statistical description is rather easy to produce by counting the features, the structural one is much more difficult because of potentially unlimited complexity of geometry of feature locations. This creates a conceptual problem of how to produce effective structural description harmoniously combined with the statistics of features. In this paper, relation between structural and statistical aspects of pattern description is studied and a unified framework is proposed. This framework is developed from the database pattern retrieval problem using statistics of local features. Robust local feature set is proposed which is based on quantized block transforms used in the video compression area. Block transforms are well-known for excellent preservation of

perceptual features even under strong quantization [2]. This property allows efficient description of comprehensive set of local features while reducing the information needed for the description. Local feature descriptors are constructed from the coefficients of quantized block transforms in the form of parameterized feature vectors. Statistics of feature vectors describing local feature distributions is easily and conveniently picked up by histograms. The histograms are treated as vectors, and, with suitable metrics, used for comparison of statistical information between the image patterns. This allows us to formulate the problem of maximizing statistical information by considering database pattern retrieval optimization using feature vector parameters as shown in previous paper [3]. Results of this process show that for optimized statistical description, the correct retrieval rate for typical images is high, but obviously the statistical approach alone cannot account for structural properties of patterns. In this paper, we aim to incorporate structural information of patterns extending and generalizing previous results based only on feature statistics. The development is based on a framework in which structural information about patterns is integrated with statistics of features into a unified flexible description.

The framework is based on the decomposition of visual patterns into subareas. The description of pattern subareas by the statistical information is expressed in the form of feature histograms. As a subarea is localized within the pattern area, it contains some structural information about the pattern. Subareas themselves can be decomposed. The smaller the subarea is, the more structural information about location of features it may contain. In an extreme case, a subarea can be limited to single feature and this will correspond to a single feature location. A pattern could be described completely by the single feature subareas, but this would be normally too complex and redundant. Usually, the subareas used for the description will be much larger and will only cover highly informative regions of patterns reflecting important structural information. The decomposition framework with subarea statistics described by vectors of feature histograms allows to search for description with reduced structural information refining the performance achieved purely from the statistical description. This is equivalent to searching for the decomposition with minimal number of subareas. The bigger the subareas are, the less structural information is included, this makes possible for different tradeoffs between the structural and statistical information.

We illustrate our approach on an example of face image database retrieval task. The face database problem is selected because of the existence of standardized datasets and evaluation procedures which allow comparing with results obtained by others. We present the statistical information optimization and structural information reduction process for face databases. Results are compared with other methods. They show that with only the statistical description, the performance is good and the introduction of little structural information by combination of just few subareas is sufficient to achieve near perfect performance on par with best other methods. This indicates that little structural information, combined with statistics of local features, can largely enhance the performance of pattern retrieval.

2. LOCAL FEATURES FOR PATTERN RETRIEVAL

There has been very large number of local feature descriptors proposed in the past [4–9]. Many of them consider edges as most representative, but they do not reflect the richness of the real world. In this paper, we propose to generate a comprehensive local feature set based on perceptual relevancy in describing sets of patterns. Basic requirement for such feature sets is compactness in terms of size and description. Such feature sets can be constructed based on block transforms which are widely used in lossy image compression. Block transforms based on the discrete cosine transform (DCT) block transforms are well known for their preservation of perceptual information even under heavy quantization. This is very desirable for local feature description since it allows for robust elimination of perceptually irrelevant information. The quantized transform represents local features by a small number of transform coefficients which provides efficient description.

The block transform used in this paper is derived from the DCT and has been introduced in the H.264 video compression standard [10]. This transform is a 4×4 integer transform and combines simple implementation with size sufficiently small for describing features. The forward transform matrix of the H.264 transform is denoted by $\mathbf{B}_f$ and the inverse transform matrix by $\mathbf{B}_i$ and has the following form:

$$\mathbf{B}_f = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix}, \qquad \mathbf{B}_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix}. \tag{1}$$

The 4×4 pixel block P is forward transformed to block H as shown in (2), and the transform block R can subsequently reconstructed from H using (3):

$$H = B_f \times P \times B_f^T, \tag{2}$$

$$R = B_i^T \times H \times B_i, \tag{3}$$

where "T" denotes the transposing operation.

The transformed pixel block has 16 coefficients representing block content in a "cosine-like" frequency space (Figure 1). The first uppermost coefficient after the transform is called DC and it corresponds to the average light intensity level of a block, other coefficients are called AC and they correspond to components of different frequencies. These AC coefficients provide information about the texture detail of a block. Typically, only lower-order AC coefficients are perceptually significant, higher-order coefficients can be eliminated by quantization. The distinctive feature of the transform (2) is that even after heavy quantization, the perceptual content is well preserved. On the other hand, such quantization will also reduce the number of different types of blocks. For such purpose, it is sufficient to use

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

Figure 1: 4×4 block transform 16 coefficients order.

scalar quantization with single quantization value Q. The quantization value Q is a parameter used in within our framework to maximize statistical information. A too small value of Q results in producing too many local features; while a too high value will limit the representation ability of the feature set. For each application, a tradeoff must be made when selecting proper value of Q. In our implementation, both the transform calculation and quantization are done by integer processing, which allows for rapid processing and iterations with different values of quantization parameter.

3. FEATURE VECTORS AND HISTOGRAMS

The quantized coefficients of block transforms are used for the construction of local feature descriptions called feature vectors. Feature vectors are formed by collecting information from the coefficients of 3×3 neighboring transform blocks. The ternary feature vector (TFV) described below is a parameterized feature vector; such parameterization provides additional mean for the maximizing statistical information.

3.1. Ternary feature vector

The ternary feature vector, proposed in [11], is constructed from the collected same-order transform coefficients of nine neighboring transform blocks. These nine coefficients form a 3×3 coefficient matrix. The ternary feature vector is formed by thresholding the eight out-of-center coefficients with two thresholds resulting in a ternary vector of length eight. The thresholds are calculated based on the coefficient values and single parameter. Within each 3×3 matrix, assuming the maximum coefficient value is MAX, the minimum value is MIN, and the mean value of the coefficients is MEAN, the thresholds are calculated by

$$\begin{aligned} T^{+} &= \text{MEAN} + f \times (\text{MAX} - \text{MIN}), \\ T_{-} &= \text{MEAN} - f \times (\text{MAX} - \text{MIN}), \end{aligned} \tag{4}$$

where the parameter f is a real number within the range of $(0, 0.5)$. Value of this parameter can be established in the process of statistical information maximization. Our subsequent experiments have shown that the performance with the changing value of f has a broad plateau in the range of $0.2 \sim 0.4$. For this reason, the value $f = 0.3$ is fixed. When the thresholds (4) are calculated, the thresholding of coefficients within the 3×3 block is done in the following way:

$$\begin{aligned} &0 - \text{the pixel value } \leq T_{-}, \\ &1 - \text{the pixel value otherwise}, \\ &2 - \text{the pixel value } \geq T^{+}. \end{aligned} \tag{5}$$

The TFV vector obtained in this way is subsequently converted to a decimal number in the range of $[0, 6560]$. An illustration of the formation of the TFV based on the 0th transform coefficient is shown on example in Figure 2. In the same way, the TFV vectors can be generated for each of the other 15 coefficients from the transform shown in Figure 1. However, many higher-order coefficients values are practically zeroed after quantization. It has also been found that some of the coefficients contribute to the retrieval performance more significantly than others [3]. For this reason, the TFVs generated from the 0th and 4th transform coefficients are used in this paper.

3.2. Histograms of TFV

The global statistics of TFV vectors are described by their histograms. The TFV histogram may have in general 6561 bins. Two examples of such histograms are shown in Figure 3.

Statistical information of patterns can be compared using the TFV histograms. This is done by calculating the $L1$ *norm* distance (city-block distance) between two histograms (other distance measures are computationally more complicated and do not bring clear advantages to the proposed method [3]). Denoting the histograms by $H_i(b)$ and $H_j(b)$, $b = 1, 2, \ldots L$, the $L1$ norm distance is calculated as

$$D(i, j) = \sum_{b=1}^{L} \left| H_i(b) - H_j(b) \right|. \tag{6}$$

It can be seen in Figure 3 that there are large variations in the values of the bins. The bins in the histograms can be ordered according to their size. Small bins will not be contributing significantly to the similarity measure (6) or even harm its performance. Then the size of the histograms can be adjusted and treated as parameter for global statistical information optimization.

As mentioned above, the TFV used in this paper are based on the 0th and 4th transform coefficients which represent different types of information about local features. The histograms for both coefficients can be combined by forming concatenated vector. The length of the combined TFV histogram equals to the sum of lengths of the two subhistograms and the norm distance (6) is still applied as the similarity measure.

Key aspects of the statistical description of patterns based on feature vector histograms of presented are worth to emphasize. The local feature set is derived from perceptually robust description and it is parameterized by quantization and thresholds. The form and size of this feature set can be thus adjusted to from the most relevant set of features. Features are used for the description of statistical information by feature histograms. However, not all features

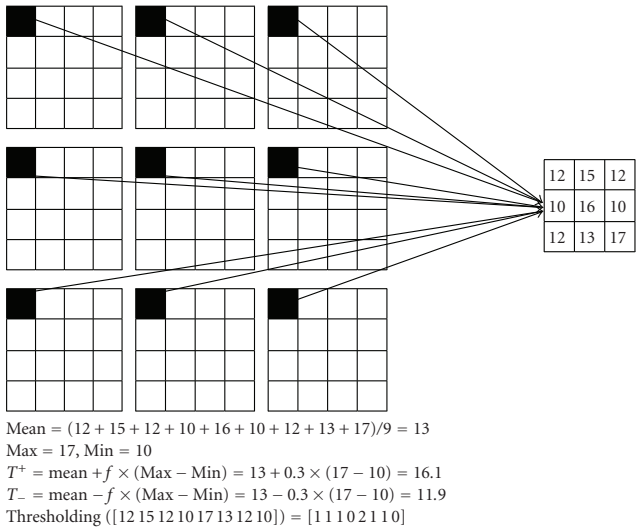


Figure 2: Formation of TFV vector: nine 0th coefficients are extracted from the neighboring 3 × 3 transformed blocks. The corresponding TFV is formed based on this 3 × 3 coefficient matrix.

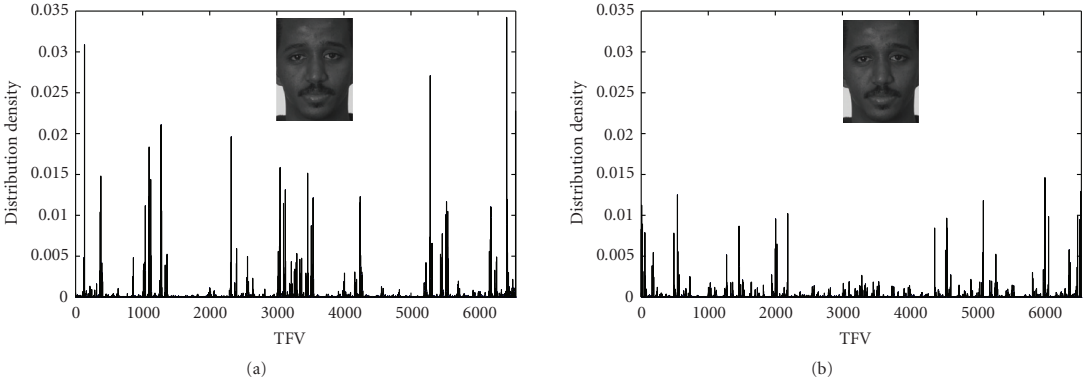


Figure 3: (a) TFV histogram of 0th coefficient; (b) TFV histogram of 4th coefficient. The x-axis shows different TFV vectors. The y-axis shows their corresponding probability distribution.

from the feature set have equal relevance. The feature histogram can be adjusted by including only the features relevant for the performance. There are thus two types of parameters used for maximizing statistical information, those acting locally on features and those acting globally on the feature histograms. The parameters can be adjusted for best performance using training. Performance can be evaluated using the test dataset. Details of this process are explained later in the paper.

4. FRAMEWORK FOR STRUCTURAL DESCRIPTION

The description of patterns by feature histograms does not include information about the structure since locations of local features are not considered. In general, structural information may be very complicated due to the almost unlimited complexity of patterns. The question is how structural information could be described in an effective way and in particular how it could be integrated with the statistical information. Such description requires flexibility in using statistics and/or structure which ever is more appropriate. The framework for such integration of statistical and structural information is described next.

4.1. Structural description of patterns by subarea histograms

Assume that a pattern P is distributed over some area C. Statistical description of the pattern proposed above uses its feature histogram H calculated over a selected local feature

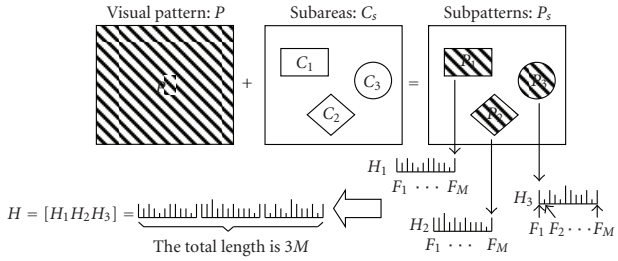


Figure 4: The pattern P is covered by the area C. The C is composed of three subareas: C_1, C_2, and C_3. Single histogram is calculated from each subarea. Each histogram contains M bins, which is corresponding to M features from the feature set F. Finally, the three histograms are concatenated in a form of $[H_1\ H_2\ H_3]$, which is description of pattern P.

set F. This histogram can be used for comparison of patterns based on their statistical content, but it does not provide any structural description since information about the locations of features within the area C is not available. To include such information, we will now define covering of the pattern area C by a set of subareas $C_1, \ldots, C_n$. The subareas do not have to be disjoint and they may have any shape and size. For each subarea C_s, its corresponding subarea feature histogram H_s, $(s = 1, \ldots, n)$ can be computed. The description of pattern P can now be done over the set of subareas using their corresponding histograms $H_1, \ldots, H_n$. This is done by forming a vector with concatenated histograms $H_C = [H_1 \cdots H_n]$. Patterns can now be compared using the city-block metrics of their concatenated vectors as illustrated in Figure 4.

The vector obtained by concatenating histograms of subareas is not equivalent to the vector of the whole pattern histogram even in the case when subareas make a proper partition of the pattern area because the subarea histograms are normalized. Hence the smaller the subarea, the more features belonging to it are weighing in the distance norm of the vector for concatenated histogram. At the same time, subareas describe structural information due to the fact that the in smaller subarea features are more localized. In an extreme case, subareas can cover only a single feature but such precise description of structural would normally be not necessary. By increasing the size of subarea, the structural information about features will be reduced while the role of statistics will be increased. Combining a number of subareas will provide combination of structural and statistical information. Thus the histogram obtained by concatenation of subarea histograms allows for flexible description of global statistical and structural information.

4.2. The database retrieval problem and system architecture

We consider a pattern database $D = \{P_1, \ldots, P_M\}$. The database retrieval problem is formulated as follows. For some key pattern P_i, we would like to establish if there are patterns similar to it in the database under certain similarity criteria. The similar patterns should be ordered according to the degree of their similarity to P_i.

A set of b most similar patterns will be the retrieval result, but sometimes there will be wrong patterns retrieved. The problem is how to find K, which has small amount of wrong patterns when compared with certain ground truth knowledge about them. To solve this problem, the similarity measure of patterns can be based on the feature histograms of suitably selected local features set. One can then take first n patterns for which similarity measure calculated for all the patterns in the database D and the pattern P_i has lowest values, these are patterns matching the P_i best. If the histograms are calculated for the whole patterns, the retrieval will be based on the statistical information only. If this would give required performance level, no structural information about location of features is necessary. This will not always be the case and then structural information of our framework has to be used to refine the performance. For this, one has to decompose the pattern area into subareas and form concatenated histograms. When a proper covering is selected, the retrieval performance will be improved when a covering maximizing the performance measure is selected, such covering can be identified by iterative search over the pattern area. If the covering is found with minimum number of subareas and maximum size, it provides minimal structural description needed to complement the statistical one for a given performance level. In this case, the overall computational complexity is not essentially increased since once the covering is found, the calculation of histograms for subareas is equivalent to the calculation of a single histogram for the whole pattern.

The proposed architecture of retrieval system for visual patterns has several key aspects from the machine learning point of view. First, the set of local features, which is robust from perceptual point of view, is not selected arbitrarily but by adjusting the quantization level of block transforms. Second, the size of feature histograms is selectable. Third, the pattern covering, that is, the scope of structural information matched. The three key parameters: quantization level, size of the histograms, and the pattern covering are optimized by running the system on training pattern sets for best performance under the similarity measure comparing to the ground truth. The overall layered system architecture is shown in Figure 5. As can be seen the system parameter

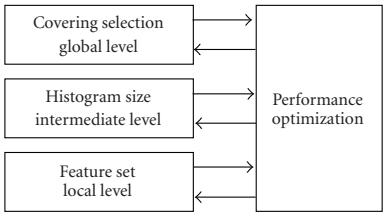


Figure 5: The system architecture layers.

optimization is done on all layers, local (features), intermediate (histogram), and high (covering), under the global performance measure. The parameter space is discrete and finite and thus the best parameters can be found in finite time. The range of quantization values and histogram sizes is very limited making only the search for covering more demanding.

5. RETRIEVAL SYSTEM PERFORMANCE EVALUATION

The proposed system has been extensively tested with retrieval from face databases. Although the method is not limited or specialized to faces, the advantage of using face databases for performance evaluation is the existence of widely used standardized datasets and evaluation procedures which enables comparison with other results. This is especially in the case of FERET face image database maintained by the National Institute of Standard and Technology (NIST) [12]. NIST published several releases of FERET database, the release used in this paper is from October 2003, called color FERET database. The color FERET database contains overall more than 10,000 images from more than 1000 individuals taken in largely varying circumstances. Among them, the standardized FA and FB sets are used here. FA set contains 994 images from 994 different objects, FB contains 992 images. FA serves as the gallery set, while FB serves as the probe set.

For the FERET database, standardized evaluation method based on performance statistics reported as *cumulative match scores* (CMSs) which are plotted on a graph is developed [13, 14]. Horizontal axis of the graph is retrieval rank and the vertical axis is the probability of identification (PI) (or percentage of correct matches). On the CMS plot, higher curve reflects better performance. This lets one to know how many images have to be examined to get a desired level of performance since the question is not always "is the top match correct?", but "is the correct answer in the top n matches?" (These are the first n patterns with the lowest value of similarity measure). However, one should notice that only few publications so far have been made based on release in 2003, many other references are based on other releases. For comparison, we also list the results from publications using both releases. The comparison for different releases can be only approximate due to the different datasets. In addition, the detail setup of experimental data of each method may be different (e.g., preprocessing, training data, version of test data). Before the experiments, all the source images are cropped to a rectangle containing face and a little background (e.g., the face images in Figure 3). They are normalized to have the same size. Eyes are located in the similar position according to the information available in FERET. Such approach is widely used to ensure the same dimensionality of all the images. However, we did not remove the background content at the four image corners (using an elliptical mask), which is believed be able to improve the retrieval performance [15]. Simple histogram normalization is applied to the entire image to tackle the luminance changes.

5.1. The training process for parameter optimization

The training process for parameter optimization for the face database is shown in Figure 6. A set of FERET face images is preprocessed by histogram normalization and next the 4×4 block transform is calculated. Subareas with structural information are selected, and for specific selection of the quantization parameter QP the combined TFV histograms are formed. Based on the histograms, the first $b\,(b = 5)$ database picture best matching to query picture are found and compared to ground truth by calculating the percentage of incorrect matches. Next, the subareas, the QP, and the length of the histograms are changed and the process is repeated until the combination of the parameters is found providing the lowest percentage of errors.

Since there is no standard training process for the color FERET database (release 2003), to minimize the bias introduced by different selection of training data, we repeated our "training + testing" experiment for five times, each time with a different training set. The process is

(1) five different groups of images are randomly selected to be the training sets. Every training set contains 50 pairs of image (all are different from other training sets); the remaining 944 images in FA and 942 images in FB are used together as the testing set;

(2) five parameter sets are obtained from the five training sets, respectively. Each parameter set will be applied to the corresponding testing set (the remaining 942/944 images) for evaluation of retrieval performance. The outcome is five CMS curves;

(3) the resulted five CMS curves are averaged, which is the final performance result.

The conclusions obtained from these five training independent experiments seem to be more robust and effective than other works which use only one training data set [16–18]. The testing system is illustrated in Figure 7.

5.2. Performance of the retrieval system using full image

We first studied the system performance without using subareas, that is, for the full image. Results for different types of TFV vectors are shown in Table 1. The CMS Rank-1 scores results based on the DC-TFV, AC-TFV histograms, and their

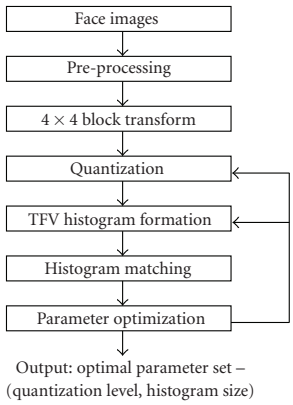


Figure 6: The parameter training process.

Table 1: Results of using complete image.

Test-A (the whole image)			
	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Rank-1 CMS score (%)	92.84	64.31	93.65

combination show that the combined histograms based on the DC and AC coefficients is best and the level of 93% is quite high. This is the starting point and reference for the following results. We will refer to this experiment as Test-A in the following. From the results in Table 1, it can be seen that DC-TFV histograms provide much better results than AC-TFV, reason for this is that feature vectors constructed using DC coefficients pickup essential information about edges. AC TFV vectors play only complementary role, picking up information about changes in high-frequency components.

5.3. Performance of TFV histograms using single subarea

In the next series of experiments, we studied the performance using single subarea of pictures. The goal was to check if the performance can be higher than full picture. We will refer to this experiment as Test-B. Since the numbers of location and size of possible subareas are very large, we generated a sample set of 512 subareas defined randomly and covering the image (Figure 8). The retrieval performance of each subarea is obtained by one retrieval experiment. Since we have five training sets for cross-validation, the final result is actually a matrix of 5×512 CMS scores. They are further averaged to be a 1×512 CMS vector. The maximum, minimum, and mean of these 512 CMS scores is shown in Table 2.

One can see from it that there is very wide performance variation for different subareas. The DC-TFV subarea histograms always perform markedly better than DC-TFV histograms, but their combination performs still better in the critical high-performance range. Comparing to the case of full image histograms before, one can see that performance

Table 2: Results of using single subarea.

Test-B (1-PID)			
Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Maximum	93.77	60.77	95.30
Minimum	9.01	1.69	12.94
Mean	56.59	20.99	62.11

Table 3: Results of using two subareas.

Test-C (2-PID)			
Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Maximum	97.76	81.94	97.70
Minimum	47.54	13.47	52.50
Mean	79.06	43.89	82.56

for best subareas can indeed be better both for DC-TFV and combination of DC-TFV and AC-TFV histograms, but not by high margin. This indicates, however, that even better performance can be achieved by combining subareas.

5.4. Performance of TFV histograms combined from two subareas

Selection of subarea can be seen as adding structural information to the statistical information described by the feature histogram. This reasoning is justified by comparing the performance obtained from the best subarea and full image (Tables 1 and 2). Continuing this line of thinking, a reasonable way to improve the performance is by increasing the structural information combining two subareas. To check for this possibility, an experiment continuing the Test-B was made by randomly selecting two subareas from different image regions. Based on the above 512 subareas in Test-B, 216 combinations of two subareas were used in Test-C for which results of are shown in Table 3. Even from this testing of a very limited set of two subareas, one can see by comparing results from Tables 1, 2, and 3 that for the best subareas, the performance for two subareas is significantly improved than using one subarea or full image. Interpreting this in terms of structural information tells that introducing additional structural information indeed improves the system performance.

5.5. Full image by subareas processing

In the above experiments, only the selected subarea(s) was used, the rest of the image is skipped. It may be argued that this does not use full image information and may result in diminished performance. Due to this reason, we consider here the case when subareas histograms are combined with the histogram of the rest of the image. We call this case the full-image decomposition (FID) case, in distinction to the previous partial-image decomposition (PID) case. The FID

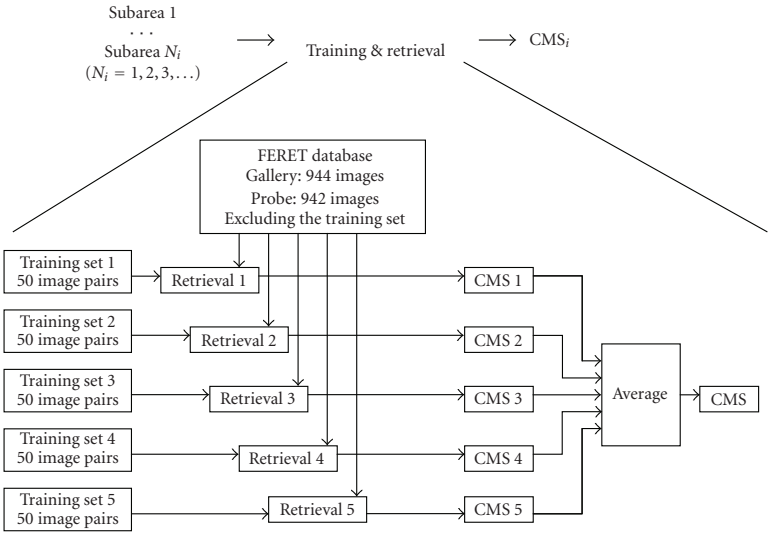


Figure 7: Training process: the optimal parameter set from five training sets is utilized separately, which give five CMS scores. The overall performance of given subarea will be evaluated as the average of above five CMS scores. 50 pairs of images selected from FA and FB are used as the training set. The remaining 944 images in FA and 942 images in FB are used together as the testing set. Such "training + testing" process has been repeated five times. Since the training sets for each time are different from each other; therefore, the testing sets for each time are also different from each other. However, the number of different image pairs between any two tests is 50 out of 942.

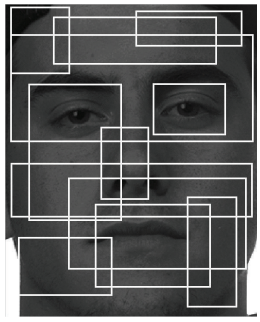


Figure 8: Some example subareas over the face image.

Table 4: Retrieval results of the FID cases.

Test-D (1-FID)			
Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Maximum	97.94	82.82	98.06
Minimum	31.49	7.48	35.04
Mean	84.12	51.42	86.48
Test-E (2-FID)			
Rank-1 CMS score (%)	DC-TFV	AC-TFV	DC-TFV + AC-TFV
Maximum	98.43	89.31	98.71
Minimum	76.15	45.28	80.54
Mean	92.87	71.30	94.14

case can also be compared to retrieval with the full-image histogram. In the full-image histogram, all features have the same impact for similarity measure, while in the FID case, selection of a subarea means increasing the impact of its features in the similarity measure.

The retrieval performance results of the FID case are shown in Table 4, which allows us to compare them with the previous PID cases. In Table 4, Test-D refers to the FID case with single subarea and Test-E refers to the case with two subareas, they are called, respectively, 1-FID (1-subarea FID) and 2-FID (2-subarea FID). One can see that again the results of the FID case are better than the results of PID from Tables 2 and 3. Remembering that in both cases of FID and PID full-image information is taken for retrieval, the reason why the FID provides better performance is that the subarea histograms emphasize information when they are combined comparing to the histogram of full image and this contributes to the retrieval discriminating ability. In other words, subareas in the FID case add structural information to the statistical information obtained from the processing of whole image.

5.6. Searching for the best subareas

As can be seen from the previous results, selection of proper subareas is critical for achieving best retrieval results.

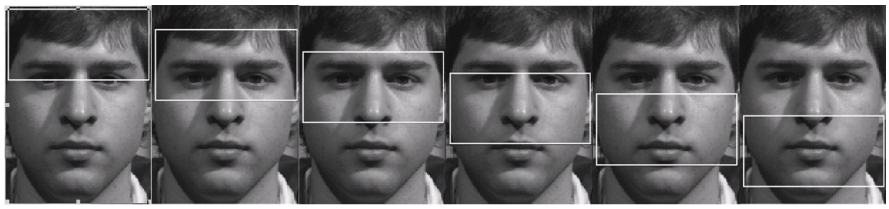


Figure 9: Example subareas from the first step of searching.

Table 5: Comparison between the results of Test-B and Test-F for the single subarea. The difference between the resulting CMS scores is less than one percent.

	Test-B and Test-D, normal searching		Test-F, fast searching	
Rank-1 CMS score (%)	DC-TFV	DC-TFV + AC-TFV	DC-TFV	DC-TFV + AC-TFV
1-PID	93.77	95.30	92.72	94.70
1-FID	97.94	98.06	97.16	97.52

Table 6: Comparison between the results of Test-C and Test-G for two subareas. The difference between the resulting CMS scores is less than one percent.

	Test-C and Test-E, normal searching		Test-G, fast searching	
Rank-1 CMS score (%)	DC-TFV	DC-TFV + AC-TFV	DC-TFV	DC-TFV + AC-TFV
2-PID	97.76	97.70	96.83	96.31
2-FID	98.43	98.71	98.23	98.37

Table 7: List of the referenced results based on release 2003 of FERET database.

Reference	[16]	[17]	[18]	[19]	
Method	Landmark bidimensional regression	Landmark	Combined subspace	Template matching	Proposed 2-FID method, fast searching
Rank-1 CMS (%)	79.4	60.2	97.9	73.08	98.37

Table 8: List of the referenced results based on different releases.

Reference	[20]				[21]	[22]
Method	PCA-L1	PCA-L2	PCA-Cosine	ICA-cosine	Boosted local features	JSBoost
Rank-1 CMS (%)	80.42	72.80	70.71	78.33	94	98.4

Table 9: Comparison of asymptotic behavior between the proposed method against ARENA and PCA-based techniques.

Methods	Training time	Retrieval time	Storage space
PCA-nearest-centroid	$O(N^3 + N^2 d)$	$O(cm + dm)$	$O(cm + dm)$
PCA-nearest-neighbor	$O(N^3 + N^2 d)$	$O(Nm + dm)$	$O(Nm + dm)$
Arena	$O(Nd)$	$O(Nm + d)$	$O(Nm)$
Proposed method	$O(sNa)$	$O(Nm + a)$	$O(Nm + 4r)$

Table 10: Running times of 2 subarea examples.

Running time (sec)	Training time	Retrieval time	Time for retrieving one image
2-PID, one coefficient	0.1908	21.7069	2.304×10^{-2}
2-FID, one coefficient	0.2946	30.5330	3.433×10^{-2}
2-PID, two coefficients	1.7172	54.3845	5.773×10^{-2}
2-FID, two coefficients	3.0340	98.5200	10.459×10^{-2}

Since the number of possible subareas is virtually unlimited, searching for the best ones may be rather tedious. For specific class of images, like faces, this may not even be necessary since searching for subareas defining informative parts of faces can be helped with simple heuristics. We applied heuristics based on the assumption that informative areas of faces can be outlined by rectangles covering the width of images. Search for the best subarea is then limited to sweeping pictures in the training sets with rectangles of different heights and widths. In order to speed up the search procedure, while at the same time keeping the good retrieval performance, we applied here a three-step searching method over the training sets. The searching procedure is thus as follows:

(1) rectangular areas covering the width of images with different heights are considered in the first step. For example, in our experiments with images of size 412×556 pixels, the height of areas is ranging from 40 to 160 pixels, with the width fixed at 400 pixels. The rectangular areas are swept over the picture height in steps of 40 pixels, as shown in Figure 9. From here, we have 32 subareas, which is a small subset of above 512 subareas. The subarea giving best result is selected as the candidate for the next step;

(2) the vertical position of the above candidate is fixed and now its width is changed. A number of widths are tested with the training dataset and the one with best performance is selected. Here, the number of tested widths is 16. After this, the subarea giving best result is selected as the candidate of for the next step;

(3) searching is performed within the small surrounding area of the best candidate rectangle. The one giving best result is selected as the final optimal subarea.

The results from the three-step searching are shown in Test-F and Test-G in Tables 5 and 6 in comparison to Test-B, -C, -D, and -E, respectively. The three-step searching method saves a lot of time in searching process, while the differences between corresponding CMS performances are mostly less than one percent, which is a very good result due to the large savings in the computation and the small size of the training set.

As can be seen from Table 6, the best result of fast searching is 98.37%. It is obtained for two subareas and combination of DC and AC TFV vectors. This result is very close to the overall best result in Test-E in Table 8 which is 98.71% obtained without the fast searching. The results are much better than obtained by other methods and it is in the range of best results obtained to date as shown next.

5.7. Comparison with other methods

In order to compare the performance of our system with other methods, we list below some reference results from other research for the FERET database. These results are all obtained by using the FA and FB set of the same release of FERET database. In [16], the eigenvalue-weighted bidimensional regression method is proposed and applied to biologically meaningful landmarks extracted from face images. Complex principal component analysis is used for computing eigenvalues and removing correlation among landmarks. An extensive work of this method is conducted in [17], which comparatively analyzed the effectiveness of four similarity measures including the typical $L1$ norm, $L2$ norm, Mahalanobis distance, and eigenvalue-weighted cosine (EWC) distance. A combined subspace method is proposed in [18], using the global and local features obtained by applying the LDA-based method to either the whole or part of a face image, respectively. The combined subspace is constructed with the projection vectors corresponding to large eigenvalues of the between-class scatter matrix in each subspace. The combined subspace is evaluated in view of the Bayes error, which shows how well samples can be classified. The author of [19] employs a simple template matching method to complete a verification task. The input and model faces are expressed as feature vectors and compared using a distance measure between them. Different color channels are utilized either separately or jointly. Table 7 lists the result of above papers, as well as the result of 2-subarea FID (2-FID) case of our method. The results are expressed by the way of Rank-1 CMS score.

In addition, we also list in Table 8 some results based on earlier releases of FERET database. They are cited from publications [20–22] which are using popular methods like: PCA, ICA, and Boosting. Although they are not strictly comparable with our results due to the different release used, they illustrate that our method is among the best to date.

The proposed method has also low complexity and it is based only on simple calculations without the need for advanced mathematical operations. In order to compare the computational complexity and storage requirements of different approaches, we use the evaluation method from [23]. The following notations have been defined:

c: number of persons in the training set;

n: number of training images per person;

N: total number of training images: $N = cn$;

d: each image is represented as a point in R^d, where d is the dimensionality of the image;

m: dimension of the reduced representation: number of stored weights, number of pixels (s^2), or number of bins of histogram. Normally, $d \geq m$;

s: number of different subarea rectangles applied to the image during the training process. For the fast-searching case, $s = 64 \sim 70$;

a: number of pixels within (i.e., size of) the applied subarea(s) $a < d$;

r: number of subareas utilized. For this paper, $r \in \{0, 1, 2\}$.

The asymptotic behavior of the various algorithms is summarized in Table 9. The proposed method is compared to the results for ARENA [24], PCA-Nearest-Centroid [25], and PCA-Nearest-Neighbor [26], which is cited from [23]. As one can see, the proposed method is simpler than

listed PCA-based methods, but is more complicated than ARENA, especially for the training process. However, one should also notice that ARENA is an alternative way of using 0th coefficient here. This is because the 0th coefficient here actually represents the average of local pixel block. In addition, the training in [23] requires multiple images per subject, while in our case we need only two images per subject.

We also evaluated the running times for the 2-subarea case using PC with Intel 1.86 GHz CPU and 2GB RAM is used for testing. Both the 2-FID and 2-PID are tested with either one coefficient or two coefficients in the TFV. The comparison between histograms of two images is the basic unit of the whole training and retrieval process. The whole training process of one training set contains 20000 interimage comparisons; the whole retrieval process (942 probe images and 944 gallery images) contains 889248 interimage comparisons. The corresponding running times are shown in Table 10.

6. CONCLUSIONS

In this paper, a framework for combining statistical and structural information of patterns for database retrieval is proposed. The framework is based on combining statistical and structural aspects of feature distributions. Feature histograms of full images represent purely statistical information. Decomposition of images into subareas adds structural information which is described by combined concatenated histograms. The number of the subareas as well as their size, shape, and locations is reflecting the complex nature of structural information.

In our approach, we reduce information needed for retrieval on several levels. First, features which are used are based on the coefficients of quantized block transforms. The ternary feature vectors are constructed from the coefficients by thresholding which further reduces feature information. Next, the information in feature histograms is decreased by reducing their length during the retrieval training process. Finally, image subareas are selected and combined to provide best performance. We present image database retrieval system in which parameters at all levels are adjusted by learning to provide best correct retrieval rate. To illustrate the retrieval capabilities, experiments are performed using standard face databases and evaluation methods. Performance evaluation shows that very good results are obtained with little structural information which is obtained by combining feature histograms from two face image subareas and the rest of the image. The resulting performance obtained is compared to and shown to be better than for other methods using the same evaluation methodology with FERET database. The presented framework is general and allows for flexible incorporation of structural information by decomposition into more subareas, resulting in even better performance. Our results illustrate what can be achieved when structural information combined into the statistical framework is minimized, which is equivalent to the reduction of the number of subareas used in the decomposition. It turns out that surprisingly little structural information is needed to achieve better performance than in other existing methods when statistical and structural information are properly combined.

ACKNOWLEDGMENTS

Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office. The authors would like to thank NIST for providing the FERET data. Support of first author by TISE scholarship is gratefully acknowledged.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA, 2006.

[2] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Compression Standard*, Van Nostrand Reinhold, New York, NY, USA, 1993.

[3] D. Zhong and I. Defée, "Performance of similarity measures based on histograms of local image feature vectors," *Pattern Recognition Letters*, vol. 28, no. 15, pp. 2003–2010, 2007.

[4] A. Franco, A. Lumini, D. Maio, and L. Nanni, "An enhanced subspace method for face recognition," *Pattern Recognition Letters*, vol. 27, no. 1, pp. 76–84, 2006.

[5] H. K. Ekenel and B. Sankur, "Multiresolution face recognition," *Image and Vision Computing*, vol. 23, no. 5, pp. 469–477, 2005.

[6] D. Ramasubramanian and Y. V. Venkatesh, "Encoding and recognition of faces based on the human visual model and DCT," *Pattern Recognition*, vol. 34, no. 12, pp. 2447–2458, 2001.

[7] X. Zhang and Y. Jia, "Face recognition with local steerable phase feature," *Pattern Recognition Letters*, vol. 27, no. 16, pp. 1927–1933, 2006.

[8] H. K. Ekenel and B. Sankur, "Feature selection in the independent component subspace for face recognition," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1377–1388, 2004.

[9] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181–191, 2005.

[10] Joint Video Team of ITU-T and ISO/IEC JTC 1, "Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC)," March 2003, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050.

[11] D. Zhong and I. Defée, "Study of image retrieval based on feature vectors in compressed domain," in *Proceedings of the 7th Nordic Signal Processing Symposium (NORSIG '06)*, pp. 202–205, Reykjavik, Iceland, June 2006.

[12] "FERET Face Database," http://www.itl.nist.gov/iad/humanid/feret/.

[13] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[14] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[15] D. Bolme, J. R. Beveridge, M. Teixeira, and B. Draper, "The CSU Face identification evaluation system: its purpose, features and structure," in *Proceedings of the 3rd International Conference on Vision Systems (ICVS '03)*, pp. 304–313, Graz, Austria, April 2003.

[16] J. Shi, A. Samal, and D. Marx, "Face recognition using landmark-based bidimensional regression," in *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM '05)*, pp. 765–768, Houston, Tex, USA, November 2005.

[17] J. Shi, A. Samal, and D. Marx, "How effective are landmarks and their geometry for face recognition?" *Computer Vision and Image Understanding*, vol. 102, no. 2, pp. 117–133, 2006.

[18] C. Kim, J. Y. Oh, and C.-H. Choi, "Combined subspace method using global and local features for face recognition," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '05)*, vol. 4, pp. 2030–2035, Montreal, Canada, July-August 2005.

[19] J. Roure and M. Faundez-Zanuy, "Face recognition with small and large size databases," in *Proceedings of the 39th Annual International Carnahan Conference on Security Technology (CCST '05)*, pp. 153–156, Las Palmas, Spain, October 2005.

[20] K. Baek, B. A. Draper, J. R. Beveridge, and K. She, "PCA vs. ICA: a comparison on the FERET data set," in *Proceedings of the 6th Joint Conference on Information Sciences (JCIS '02)*, vol. 6, pp. 824–827, Durham, NC, USA, March 2002.

[21] M. Jones and P. Viola, "Face recognition using boosted local features," Tech. Rep. TR2003-25, Mitsubishi Electric Research Laboratories, Cambridge, Mass, USA, 2003.

[22] X. Huang, S. Z. Li, and Y. Wang, "Jensen-shannon boosting learning for object recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 144–149, San Diego, Calif, USA, June 2005.

[23] T. Sim, R. Sukthankar, M. Mullin, and S. Baluja, "Memory-based face recognition for visitor identification," in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG '00)*, pp. 214–220, Grenoble, France, March 2000.

[24] C. G. Atkeson, A. W. Moore, and S. Schaal, "Locally weighted learning for control," *Artificial Intelligence Review*, vol. 11, no. 1–5, pp. 75–113, 1997.

[25] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[26] S. Lawrence, C. Giles, A. Tsoi, and A. Back, "Face recognition: a hybrid neural network approach," Tech. Rep. UMIACS-TR-96-16, University of Maryland, College Park, Md, USA, 1996.