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.

Copyright © 2008 D. Zhong and I. Defée. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

### 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  $4 \times 4$  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:

$$\mathbf{B}_{f} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix}, \qquad \mathbf{B}_{i} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0.5 & -0.5 & -1 \\ 1 & -1 & -1 & 1 \\ 0.5 & -1 & 1 & -0.5 \end{bmatrix}.$$
(1)

The  $4 \times 4$  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:  $4 \times 4$  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  $3 \times 3$  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  $3 \times 3$  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  $3 \times 3$  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

$$T^{+} = \text{MEAN} + f \times (\text{MAX} - \text{MIN}),$$
  
 $T_{-} = \text{MEAN} - f \times (\text{MAX} - \text{MIN}),$  (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  $3 \times 3$  block is done in the following way:

0 – the pixel value 
$$\leq T_{-}$$
,  
1 – the pixel value otherwise, (5)  
2 – the pixel value  $\geq T^{+}$ .

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, ... L, the L1 norm distance is calculated as

$$D(i,j) = \sum_{b=1}^{L} |H_i(b) - H_j(b)|.$$
 (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

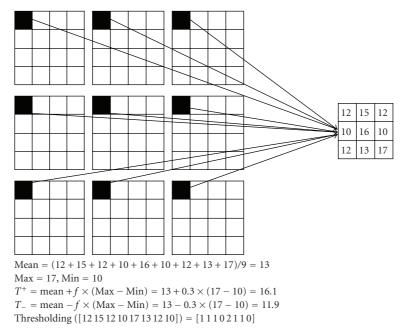


FIGURE 2: Formation of TFV vector: nine 0th coefficients are extracted from the neighboring  $3 \times 3$  transformed blocks. The corresponding TFV is formed based on this  $3 \times 3$  coefficient matrix.

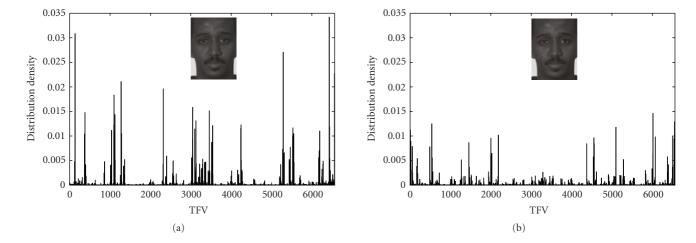


FIGURE 3: (a) TFV histogram of 0th coefficient; (b) TFV histogram of 4th coefficient. The *x*-axis shows different TFV vectors. The *y*-axis shows their corresponding probability distribution.

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

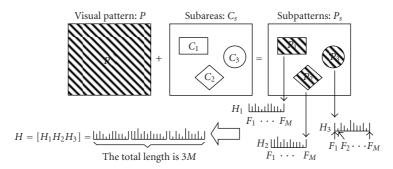


FIGURE 4: The pattern P is covered by the area C. The C is composed of three subareas:  $C_1$ ,  $C_2$ , and  $C_3$ . Single histogram is calculated from each subarea. Each histogram contains M bins, which is corresponding to M features from the feature set F. Finally, the three histograms are concatenated 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  $D = \{P_1, \dots, P_M\}$ . 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

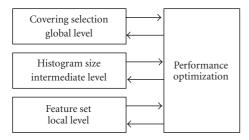


FIGURE 5: The system architecture layers.

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  $4 \times 4$  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

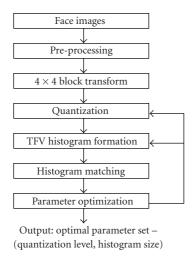


FIGURE 6: The parameter training process.

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  $5 \times 512$  CMS scores. They are further averaged to be a  $1 \times 512$  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

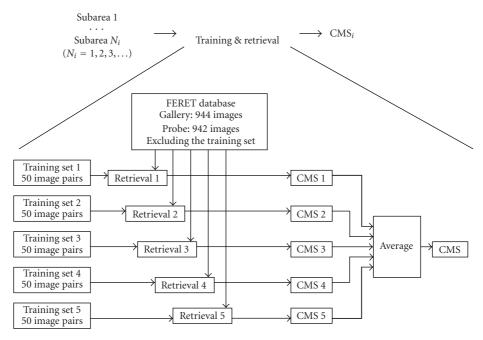


FIGURE 7: Training process: the optimal parameter set from five training sets is utilized separately, which give five CMS scores. The overall performance of given subarea will be evaluated as the average of above five CMS scores. 50 pairs of images selected from FA and FB are used as the training set. 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.

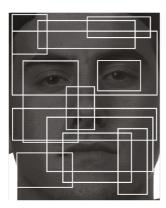


FIGURE 8: Some example subareas over the face image.

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

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			
	Te	st-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			

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	l 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	l 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	Landmark	Combined subspace	Template matching	Proposed 2-FID method,
regression	regression	Landinark	Comonica suospace	remplate matering	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	$2.304 \times 10^{-2}$
2-FID, one coefficient	0.2946	30.5330	$3.433 \times 10^{-2}$
2-PID, two coefficients	1.7172	54.3845	$5.773 \times 10^{-2}$
2-FID, two coefficients	3.0340	98.5200	$10.459 \times 10^{-2}$

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,  $d \ge m$ ;
- 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.