FACE EXTRACTION FROM NON-UNIFORM BACKGROUND AND RECOGNITION IN COMPRESSED DOMAIN

Nicolas Tsapatsoulis, Nikolaos Doulamis, Anastasios Doulamis and Stefanos Kollias

Electrical & Computer Engineering Department, NTUA
Heroon Polytechniou 9, 15773 Zografou, Greece
E-mail: ntsap@image.ntua.gr

ABSTRACT

A complete face recognition system is proposed in this paper by introducing the concepts of foreground objects, which are currently used in the MPEG-4 standardization phase, to human identification. The system automatically detects and extracts the human face from the background, even if is not uniform, based on a combination of a retrainable neural network structure and the morphological size distribution technique. In order to combine face images of high quality and low computational complexity, the recognition stage is performed in compressed domain. Thus, in contrast to existing recognition schemes, the face images are available in their original quality and not only in their transformed representation.

1. INTRODUCTION

Computer recognition of personal identity is an old but one of the most fundamental problems in the field of pattern analysis. Besides its significant role in office automation or human-machine interaction (e.g., we would prefer machines to acquire human skills, like recognizing faces, gestures or speech rather than humans try to acquire machine skills) it is also very useful in security systems (e.g., for admission access of persons to buildings) and in application such as crime prevention or criminal identification. Today humans prove their identity by handling cards, keeping serial numbers or using passwords. Nonetheless, all these methods point out only the persons that carry these keys and not the genuine owners themselves.

The large variations of a human face, like changes of facial expressions or viewpoints, age, illumination conditions, disguise and noise make face recognition a difficult task since a face recognition system should be able to identify persons regardless of the previous "distortions". Many works have been proposed in the literature for face recognition problem, which have generally been focused on highly constrained environments, such as clean background, frontal or profile views, faces in the center of the images and so on [1]. These methods deal with extraction of features, elements of which either include geometry characteristics or linear transformations [3,7,10]. One of the most popular approaches is the eigenvector decomposition which has been shown to be an effective technique for representing human faces in a low dimensional spaces [10]. The Karhunen-Loeve Transform and Principal Component Analysis are in the heart of eigenvector representations and are used both for face detection as well as for the recognition task [7,10].

Nevertheless, the previous techniques demand the face databases to be in a proprietary format and retain only a coarse transformed representation of them. In other words the face databases are used only for face recognition and are not available for other applications like multimedia or indexing [2]. However, to achieve high recognition accuracy, images in face databases should be in "head format". This means that background and cloth information should be discarded, faces should have similar scaling, viewpoint and illumination, and protuberant features on them (eyes, mouth) should be aligned.

In this paper we propose a complete recognition system by introducing the concepts of foreground objects, which are currently used in the MPEG-4 standardization phase [2,5], to human identification. Faces are initially extracted from the background, regardless of its type (uniform or not), by using a combination of on line retrainable neural networks (meaning that they are able to be automatically retrained each time their performance is inadequate) described in [4] and the morphological size distribution transform [6]. In the subsequent phase, the extracted face is appropriately rescaled and normalized to meet given scale and illumination demands of "head format". Face feature alignment is achieved by using a detection procedure to other dominant points of a human face, like eyes, mouth and nose. Face images in "head format" are compressed using conventional techniques, like JPEG, to build the face database. Each test face image is put in "head format" and the recognition procedure is executed in the compressed domain (e.g., block DCT coefficients for JPEG formatted databases) using the nearest neighbor rule.

2. AN OVERVIEW OF THE SYSTEM

A block diagram of the proposed recognition system is presented in Fig. 1. The system consists of two basic stages, the face extraction and normalization stage and the feature extraction and recognition stage.

The incoming image is first fed as input to a neural network structure, the output of which includes a compact human object, and perhaps some background objects which cannot be isolated by the network structure in the form of a binary mask. This step is described in subsection 3.1. Although in video coding applications extraction of a human (or in general of a foreground) object is required so that the algorithm permits high flexibility, different frame rate and object manipulation, in face recognition systems only the human face affects the identification. Once the

binary mask is extracted, the human face is isolated by applying a technique based on the morphological size distribution transform (subsection 3.2). The final subsystem of the first stage is the normalization one where the masked face image is transformed to a given scale, size, and illumination and the prominent facial features are aligned in order to fit "head format" (subsection 3.3).

After the normalization step, the recognition phase takes place relied on features extracted from the "head format". The extracted features lie in the compressed domain so that the same database is used for face recognition and multimedia applications (section 4).

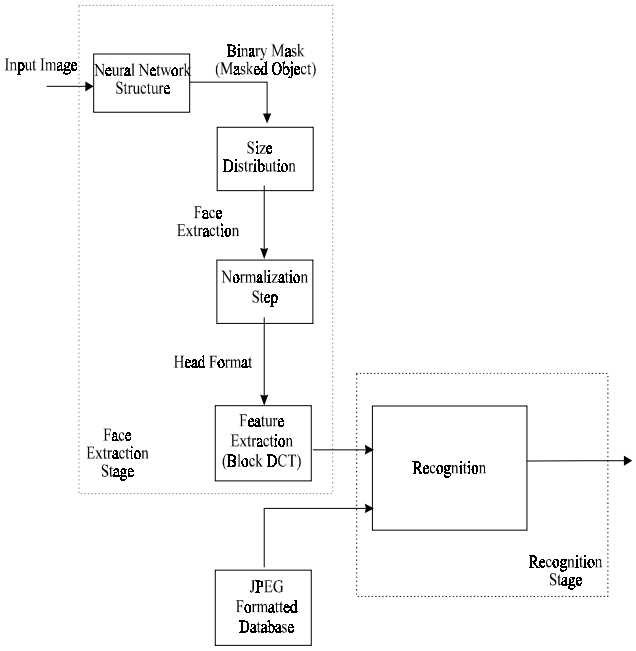

Figure 1. The proposed face recognition architecture.

3. FACE EXTRACTION STAGE

3.1 On Line Retrainable Neural Network Structure

The first step through to the face extraction is the masking of face and other objects of the foreground. Conventional segmentation based on spatial homogeneity criteria fail due to the fact that a physical object, such as a human being, generally contains regions with completely different colors and texture which are classified to different segments according to spatial homogeneity criteria [8]. On the other hand dedicated neural networks, although capable of segmentation, make the assumption of stationarity of training data which is not valid in face recognition systems since dynamic changes of the environment usually occur. Therefore to mask the foreground objects an intelligent network structure should be applied to the input face image, being able to adapt its behavior to the current condition.

To implement the adaptation behavior, a retraining procedure should take place to improve the network performance whenever the current condition is different from the initially trained ones. This procedure should include three parts:

- a decision mechanism which will determine when and what type of retraining should be applied
- a training set composed of input and desired output data automatically computed from the current environment and
- a training algorithm which will efficiently adapt the network weights.

Thus when the decision mechanism detects change of the current environment, it activates the retraining procedure which selects the new training set so as to improve the network performance. After the selection of the new training set, a training algorithm is applied to calculate the network weights. The selection of the new training set is based on a Markov Random Field, (MRF), model which satisfies the local connectivity criteria of the foreground/background object.

However, to achieve the new training set selection we assume that a coarse approximation of the final classification is available, in any operational environment using only a part of the total features required for the classification. This means that we have a weak classifier that provides us an indication of the final classification. The optimal selection of the new training set is given by maximizing the following equation

$$\{\hat{\underline{z}}, \hat{\underline{w}}, \hat{\underline{s}}\} = \arg\max_{\underline{z},\underline{w},\underline{s}} L(\underline{z}, \underline{w}, \underline{s} / \underline{y}, \underline{x}) \quad (1)$$

where L denotes the likelihood function, $\underline{z}$, $\underline{y}$ the final and the approximate network output respectively and $\underline{s}$, $\underline{w}$ the new training set and the network weights. Taking into account that the applied training algorithm finds the optimal weights for a given training set, and that the network output depends on the network weights and the incoming image, we can conclude that

$$\hat{\underline{s}} = \arg\max_{\underline{s}} \{\log \Pr(\underline{s} / \underline{y})\} =$$

$$= \arg\min_{\underline{s}} \{-\log \Pr(\underline{y} / \underline{s}) - \log \Pr(\underline{s})\} \quad (2)$$

where the Bayesian formula has been applied to Eq. (2).

The probability $\Pr(\underline{s})$ is modeled as an MRF, as follows

$$\Pr(\underline{s}) = A \exp\left(-\frac{\sum_{c \in C} V_c(\underline{s})}{\beta}\right) \quad (3)$$

where A and β are constants and $V_c(\underline{s})$ any function of a local group of points c called clique and C is the set of all such local groups.

An interest choice for the function is

$$\sum_{c \in C} V_c(\underline{s}) = \sum_{i=0}^{L-1} \sum_{l \in \partial_i} \rho(s_i - s_l) \quad (4)$$

where cost function $\rho(\cdot)$ should award image blocks that satisfy the smoothness property and discourage the rest:

$$\rho(s_i - s_l) = (s_i - s_l)^2 + \xi(s_i - y_i)^2 \quad (5)$$

while the parameter ξ controls the contribution of each of the two terms.

In a difference to the technique proposed in [4], the decision mechanism which activates the retraining procedure should be put in use in every frame in recognition problem where the images are still and coming at random.

Therefore by selecting the new training set and then applying a conventional training algorithm to estimate the network weights, such as an LVQ in our case, the proposed network structure is able to adapt its behavior to the current operational condition and perform good results in all environments.

3.2 Morphological Size Distribution

The following step is the isolation of the human face based on the extracted foreground objects since this part is the one which affects the recognition stage. Assuming that the human face is the largest part of the extracted human object, we are able to apply the size distribution technique (granulometries), as it is described in morphological theory, for detecting the human face. This technique also discards background objects which have been misclassified by the network structure.

If we denote by a_r the algebraic opening where r is a positive number, then the size distribution of a binary set *X* is defined as

$$a_r(X) = X \circ rB \tag{6}$$

where the symbol $\circ$ characterizes the opening operator of the set *X* by the structure element *rB*.

The structure element *B* is supposed to be compact and convex while *r* is a number such that $rB = \{rb: b \in B\}$. As a result *B* is the shape pattern and *r* its size. In general the following inequalities are held

$$mes(X) \geq mes(a_r(X)) \geq mes(a_s(X)) \text{ when } 0 \leq r \leq s \tag{7}$$

while the symbol *mes* characterizes a measure of the output set, such as length, area or volume. In our case where the binary mask lies in the two dimensional space the area measure is used for the *mes*.

Eq. (7) means that as the size of the structure element *B* increases, the area of the opening of the set *X*, $a_r(X)$, reaches the area of *X*.

The detection of critical scales by the size distribution is implemented through the shape-size histogram given as follows:

$$PS_X(+n, B) = area[X \circ nB - X \circ (n+1)B] \; n \geq 0 \tag{8}$$

Thus, if a pattern of size *nB* exists in the set *X*, a peak appears in the shape-size histogram and the specific pattern as well as its size is detected. Otherwise the previous difference will be small denoting that the given pattern does not exist in the image at size *n*.

Approximating a face as a polygon of eight unequal vertices and applying the previously described method using as set *X* the binary mask extracted by the network output we achieve to isolate the human face (under the assumption of being the largest object) from the rest of the image.

3.3 Normalization Phase

Once the face, modeled as polygon, is isolated the next step is to fit it into the "head format". This is accomplished in three different steps:

Manipulation and rescaling: The isolated face is put to an object-center reference frame and its scaling is converted using either an interpolating or downsampling filter. The proportion of rescaling is calculated based upon the estimated face size and the given scale of the "head format".

Illumination normalization: In order to normalize luminance scale, histogram equalization is performed. The histogram function *H(u)* of the isolated face is partitioned into n_k segments which have equal histogram density G_k:

$$G_z = \frac{1}{n_k} \sum_{u=1}^{255} H(u) \tag{9}$$

Facial features alignment: Protuberant facial features should be in specific locations so as to meet "head format". To perform the alignment, these features should be accurately located. Moreover, in some cases further local rescaling is needed. Facial feature detection is a generally difficult task, demanding multiscale searching [9]. However, due to the earlier manipulation and rescaling step, using simple template matching in limited face areas, the eyes, mouth and nose are exactly located.

After the normalization phase, large variations due to different scaling, translation, or illumination conditions are minimized and the recognition error dramatically drops. However, small variations arising from inaccurate performance of the previous steps exist as well as variations emerging from face expressions. The recognition stage should not be affected from the above variations and therefore the recognition features should be appropriately selected.

4. FACE RECOGNITION STAGE

The recognition stage is performed directly to compressed domain so that only minimal decoding of the compressed images is required. Consequently the algorithm is able to run in real time on regular workstations, which is very important especially in security systems where the admission of access should be executed immediately. We assume JPEG compressed images since this is the most popular and commonly-used compression format. The algorithm concentrates on block level since it is the lowest level of the JPEG format.

As recognition criterion we use the distance of the block DCT coefficients which are available directly from the bit stream. One advantage of this criterion is its ability to increase the recognition accuracy, increasing also the computational complexity, by allocating more DCT coefficients of each block for the recognition task. Furthermore due to the database format (JPEG) and recognition scheme the whole system is also appropriate for multimedia and indexing applications and content-based retrieval.

Although the main concept of the proposed scheme is to accelerate the required execution procedure rather than achieving

a high accuracy, the combination of the automatic extraction of human faces and their normalization, which are performed in real time by the previous stage, increases the rate of recognition accuracy to levels comparable with conventional techniques such as the eigenface decomposition [10]. However, in applications where the execution time is not so crucial, other techniques can be applied to the normalized face images effectively increasing the recognition accuracy.

5. EXPERIMENTAL RESULTS

Face segmentation and extraction results were obtained for 20 face images, of non-uniform background, with excellent performance. An example is presented in Fig. 2 where the masked images provided by the neural network structure and the morphological size distribution technique are shown.

The experimental results for the recognition stage presented here, were obtained using the ORL face database. Comparisons were accomplished for all methods using face images being in "head format". To compare our recognition stage (BDCT coefficients) with eigenface (KL) one we restrict the spatial resolution of "head format" to 64x72. All recognition stages were evaluated under variations resulting from inaccurate face extraction (small scaling and illumination variations) as well as variation in facial expression and slope. We also compare our recognition stage with one based on the DCT coefficients of whole images.

It is necessary to mention that the minimum number of retained coefficients in BDCT recognition stage is equal to 72 (number of 8x8 blocks in "head format" images) and therefore its superior performance over the other recognition stages is rather misleading. However, BDCT recognition stage has the great advantages of working in the compressed domain and being capable for multimedia and content-based retrieval applications.

(a) (b)

Figure 2. (a) Objects masked by the neural network structure (b) Isolated face surrounded by a polygon of unequal vertices.

Performance (%) (Retrained Coefficients)				
KLT	44 (16)	54 (25)	65 (36)	77 (49)
DCT	52 (16)	59 (25)	73 (36)	84 (49)
BDCT	81 (72)	81 (144)	85 (216)	88 (288)

Table 1. Recognition rate for three different recognition stages.

6. CONCLUSION

A face recognition system which exploits information in the compressed domain have been proposed in this paper. The system is accompanied by a head extraction and a normalized mechanism which substantially supports the recognition task. Furthermore due to the database format (JPEG) and recognition scheme the whole system is appropriate for multimedia and indexing applications and content-based retrieval.

Acknowledgment The authors would like to express the gratitude to Yannis Avrithis for his fruitful remarks and to the Oliveti Reasearch Laboratory in Cambridge for ORL face database

7. REFERENCES

[1] Chellapa P., Wilson C., and Sirohey S. "Human and Machine Recognition of Faces: A Survey". *Proc. IEEE*, vol. 83, no. 5, pp. 705-740, 1995.

[2] Chiariglione L. "MPEG and Multimedia Communications," *IEEE Trans. on Circuits and Systems for Video Techn*. vol. 7, no. 1, pp. 5-18, Feb. 1997.

[3] Doulamis A., Tsapatsoulis N., Doulamis N., and Kollias S. "Innovative Techniques for the Recognition of Faces Based on Multiresolution Analysis and Morphological Filtering" *Proc. of IWISP*, Manchester, November 1996.

[4] Doulamis A., Doulamis N., and Kollias S. "Retranable Neural Networks for Image Analysis and Classification" *Proc. of IEEE Int Conf. on Syst. Man & Cybern*., Orlando, October 1997.

[5] Kollias S., Doulamis N., and Doulamis N. "Improving the Performance of MPEG Compatible Encoders Using on Line Retrainable Neural Networks," *Proc. of ICIP'97*, Santa Barbara, October 1997.

[6] Maragos P. "Differential Morphology and Image Processing". *IEEE Trans. Image Processing*, vol. 5, pp. 922-937, June 1996.

[7] Moghaddam B., and Pentland A. "Probabilistic Visual Learning for Object Representation". *IEEE Trans.on PAMI*, vol. 19, pp. 696-710, July 1997.

[8] Reusens E., Ebrahimi T., Le Buhan C., Castagno R., Vaerman V., de Sola Fabregas C., Bhattacharjee S., Bossen F., and Kunt M. "Dynamic Approach to Visual Data Compression". *IEEE Trans. on Cir. & Syst. Fo Video Techn.*, vol. 7, pp. 197-211, February. 1997.

[9] Tsapatsoulis N., Karpouzis K., Votsis G., and Kollias S. "Analysis by Synthesis of Facial Images Based on Frontal and Profile Views". *Proc. of IWSNHC3DI Conference*, Rhodes, September 1997.

[10] Turk M., and Pentland A. "Eigenfaces for Recognition," *Journ. Cognitive Neuroscience*, vol. 3, no. 1, 1991.