Learning the human face concept in black and white images

Nicolae Duta and Anil K. Jain
Michigan State University
Department of Computer Science
East Lansing, MI 48824-1226, USA
dutanico@cps.msu.edu, http://www.cps.msu.edu/∼dutanico

Abstract

This study presents a learning approach for the face detection problem. The problem can be stated as follows: given an arbitrary black and white, still image, find the location and size of every human face it contains. Numerous applications of automatic face detection have attracted considerable interest in this problem [8, 7, 1, 5, 3, 4], but no present face detection system is completely satisfactory from the point of view of detection rate, false alarm rate and detection time. We describe an inductive learning-based detection method that produces a maximally specific hypothesis consistent with the training data. Three different sets of features were considered for defining the concept of a human face. The performance achieved is as follows: 85% *detection rate, a false alarm rate of* 0.04% *of the number of windows analyzed and 1 minute detection time for a* 320×240 *image on a Sun Ultrasparc 1.*

1. Introduction

This paper explores new ways of learning and retrieving the appearance of human faces in black and white images. The retrieval problem, known as face detection, can be defined as follows: given an arbitrary black and white, still image, find the location and size of every human face it contains. There are many applications in which human face detection plays a very important role: it represents the first step in a fully automatic face recognition system, it can be used in image database indexing/searching by content, in surveillance systems and in human-computer interfaces.

There have been several attempts to automatically detect human faces [8, 7, 1, 5, 4, 3]. Some of them rely on locating facial features [3] or additional cues like color or motion [5]. While these approaches tend to work well in a specified, restricted environment, their ability to handle noisy, complex background images where faces can appear at multiple unknown scales is rather poor. There are few truly general face detection systems and all of them are based on some kind of learning [8, 7, 4, 1]. The general structure of a learning-based face detection system is depicted in Figure 1. Several windows are placed at different positions and scales in the test image and a set of raw features are computed from each window and fed into a classifier. Typically, the features used to describe a face are the "normalized" gray-level values in the window. This may generate a large number of features (of the order of a couple of hundred), whose classification is time consuming and requires a large number of training samples to overcome the "curse of dimensionality". The main difference among these systems is the classification method; both parametric (Markovian models - [4, 1]), nonparametric (neural networks - [7]) and a mixture of the two (Gaussian clusters in a feature space reduced by PCA - [8]) have been used.

2. Defining the human face concept

In order to have a well defined concept to learn, one should first specify the instance (raw feature) space from which the concept examples are drawn. One of the goals of this study was to test feature sets used for discriminating textured or slightly textured objects. At the same time, emphasis was put on having a small dimensional feature space. The first feature set we tested is designed as follows: a hypothesis window (whose aspect ratio is 4/3 - see Fig. 1) is histogram equalized to 32 gray levels and then sub-sampled to an 8×6 window. The resulting instance space of 5-bit 8×6 images has 32^{48} elements of which a very small fraction actually represent faces.

The second and third set of features came from texture analysis. According to Gagalowicz [2], grey-tone textures are well modeled by the histogram of the image together with a set of second-order spatial averages (c_Δ) for $|\Delta| \leq 9^\circ$ of solid angle, where (see Fig. 1 (b)):

$$c_\Delta = \frac{1}{K}\sum_{i=1}^{K} \frac{(X_i - \bar{X}_i) \times (X_{i+\Delta} - \bar{X}_{i+\Delta})}{var(X_i) \times var(X_{i+\Delta})} \qquad (1)$$

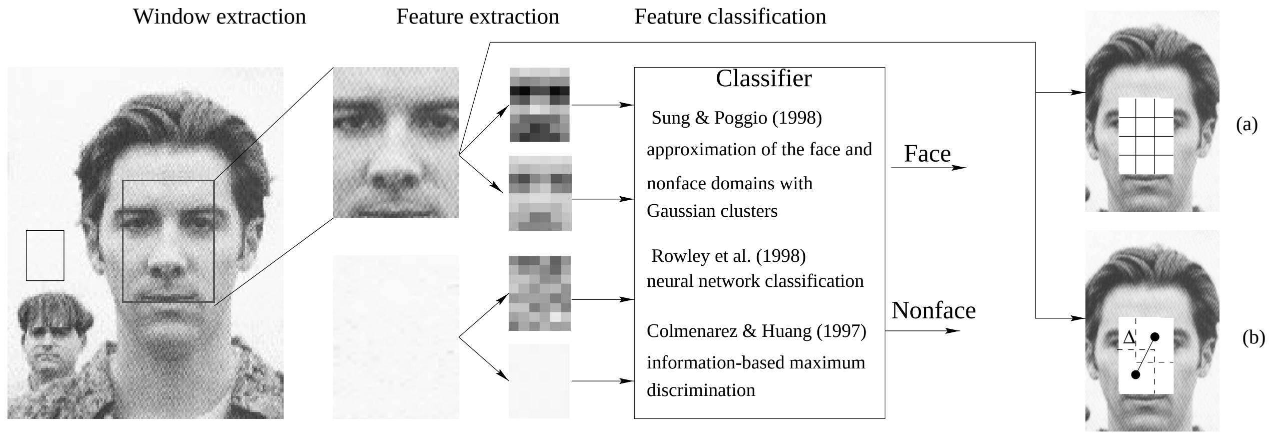


Figure 1. The structure of a learning-based face detector.

- i is a single subscript indicating the location (ix, iy) of a pixel in the image plane,
- X_i and $X_{i+\Delta}$ are the gray-levels at pixels i and $i + \Delta$,
- Δ is a translation $(\Delta x, \Delta y)$ of the plane,
- K is the number of point pairs $(i, i + \Delta)$ in the image.

Since a test image may contain many different textures, we compute the texture features only locally, that is, inside a window W that hopefully contains only one texture. In case of human faces, such a window can be located between the forehead and the mouth (see Fig. 1 (a)). If the histogram of the hypothesis window is equalized, it cannot be used for classification, so we replaced it by the set of 8 gray-level histograms of the 12 equal sized square regions.

To compute the correlation coefficients c_Δ, we considered 40 translations with norms up to 18 pixels (instead of norms of less than 12 pixels which correspond to 9° of solid angle) in order to obtain a better discrimination power.

3. Learning algorithm

Sung and Poggio [8] proposed the hypothesis according to which the face domain can be approximated by the union of a small number of ellipsoids in the instance space. In order to verify this hypothesis, we decided to cluster the examples in the training set and compute the centroid of each cluster and the Mahalanobis distance from each member of the cluster to the centroid. If the clusters are ellipsoidal and the training examples are representatives of the target concept, then the (Mahalanobis) distance from a new face instance to one of the cluster centroids should be smaller than a fraction of the maximum distance from the points of that cluster to its centroid. Analogously, the distance from a non-face instance to each cluster centroid should be larger than most of the distances from the face points of the cluster to the centroid. This strategy is equivalent in a continuous instance space to the Find-S ([6]) learning algorithm, and produces a maximally specific hypothesis consistent with the (positive) training examples. Like Find-S, it doesn't use any negative examples and its bias is a restricted hypothesis space: only unions of ellipsoids are considered as valid hypotheses. A high-level description of our face detection algorithm is given below:

Learning Phase

Given: A set $\mathcal{F} = (F_i)_{1 \leq i \leq n}$ of face feature vectors computed as described in Section 2.

1. Partition $\mathcal{F}$ into a set of m clusters $(\mathcal{C}_j)_{1 \leq j \leq m}$.

2. For each cluster $\mathcal{C}_j$, compute its centroid M_j, inverse of the covariance matrix Σ_j, and the histogram of the (Mahalanobis) distances between the elements of $\mathcal{C}_j$ and its centroid M_j. Define D_j as the "radius" of the smallest ellipsoid centered at M_j that contains a given percentage p of the population in cluster $\mathcal{C}_j$.

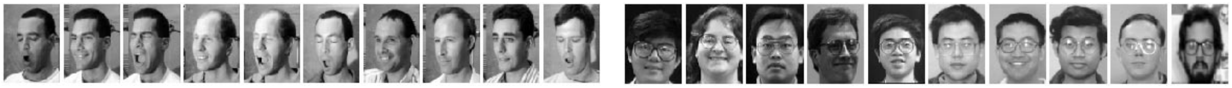
Detection Phase

Given:

1. A set $(M_j, \Sigma_j, D_j)_{1 \leq j \leq m}$ of cluster centroids, inverses of the covariance matrices and radii.

2. A test image with A rows and B columns.

```
For $Sc = MinSc$; $Sc \leq MaxSc$; $Sc = Sc + Step$
  For $i = 1$; $i \leq A$; $i = i + tStep$
    For $j = 1$; $j \leq B$; $j = j + tStep$ do steps (i)-(iv):
      (i). Extract a window $W$ of width = Sc and height = 
           $4 * Sc/3$ centered at pixel $(i, j)$ if this window
           is contained in the image.
      (ii). Extract a feature vector $F$ from this window.
      (iii). Find the centroid $M_c$ closest (using Mahala-
           nobis distance) to $F$.
      (iv). If the distance between F and $M_c$ is less than $D_c$
           then classify W as a face, otherwise classify W as
           a non-face.
```



(a) (b)

Figure 2. Face clusters. (a) Cluster described as left semi-profile with frontal lighting. (b) Cluster described as frontal views of people wearing glasses.

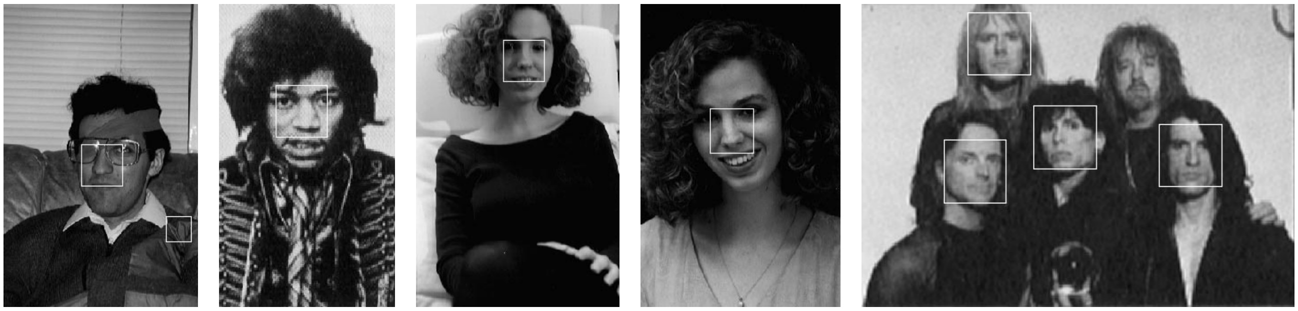


Figure 3. Results of the face detection procedure on images from Carnegie Mellon face database.

4. Experimental method

The training set consisted of 1200 images from the MIT, Weizmann and MSU face databases. The central part of each face (see Fig. 1) was manually extracted and the three feature sets described in Section 2 were computed. A hierarchical clustering of the training patterns in each feature set was performed using Ward's method. Ward's hierarchical clustering algorithm seems to be superior to the K-means algorithm (used by Sung and Poggio [8]) which produces clusters that are too dependent on the positions of the initial cluster centers. An interesting point to make here is that if the dendrogram is cut at a level around 25, the resulting clusters can be easily described semantically as a function of the head position and illumination (the facial expression is not important here since only a portion of the face has been considered). Further cluster unifications are done according to these criteria, and at a level of 5, mostly the illumination information and not the head position is dominant. A couple of typical examples of clusters at level 15 are shown in Fig. 2.

For each feature set, a minimum Mahalanobis distance classification assigns the test pattern to the nearest face cluster. If this distance is smaller than 95% of the maximum distance for that cluster computed using the training samples, then that pattern is classified as being a face according to the given feature set.

5. Results

None of the three classifiers was able to produce by itself a satisfactory detection/false alarm rate. For a detection rate of about 90% of the faces present in the test set, the false alarm rate was up to 5% of the number of windows analyzed. On the other hand, if we combine the outputs of the three classifiers such that a window is classified as face only if it is accepted by all the three classifiers, then the false alarm rate drops to 0.04% for a detection rate of about 85%. The detection time is about 1 minute on a Sun Ultrasparc 1 for a 320×240 image. The results on several images from the Carnegie Mellon face testing set are presented in Fig. 3.

References

[1] A. Colmenarez and T. Huang. Face detection with information-based maximum discrimination. In *Proceedings of CVPR-'97*, San Juan, Puerto Rico, 1997.

[2] A. Gagalowicz. Texture modeling applications. *Visual Computer*, (3):186–200, 1987.

[3] T. K. Leung, M. C. Burl, and P. Perona. Finding faces in cluttered scenes using random labelled graph matching. In *Proceedings of ICCV-'95*, Cambridge, MA, 1995.

[4] M. Lew and N. Huijsmans. Information theory and face detection. In *Proceedings of ICPR-'96*, pages 601–610, Vienna, Austria, 1996.

[5] S. H. Lin, S. Y. King, and L. J. Lin. Face recognition/detection by probabilistic decision-based neural network. *IEEE Trans. Neural networks*, 8(1):114–131, 1997.

[6] T. Mitchell. *Machine Learning*. McGraw Hill, New York, 1997.

[7] H. Rowley, S. Baluja, and T. Kanade. Neural network - based face detection. *IEEE Trans. Pattern Anal. and Machine Intelligence*, 20(1):23–38, 1998.

[8] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. Pattern Anal. and Machine Intelligence*, 20(1):39–52, 1998.