FRONTAL FACE DETECTION ALGORITHMS BASED ON SUPPORT VECTOR MACHINES AND MAXIMUM LIKELIHOOD

E. Loutas, C. Kotropoulos, N. Bassiou, and I. Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 540 06, Greece
Email: {costas,pitas}@zeus.csd.auth.gr

ABSTRACT

Face detection is a key problem in building automated systems that perform face recognition/ verification, model-based image coding, face tracking, and surveillance. Two algorithms for face detection based on either support vector machines or maximum likelihood estimation are described and their performance is tested on a collection of single images from the M2VTS database that depict one frontal face in front of a uniform background using the false acceptance and false rejection rates as quantitative figures of merit. Moreover, we demonstrate how the maximum likelihood face detection performs, when single images that depict multiple frontal faces in front of an nonuniform background are processed.

1. INTRODUCTION

Face detection is a considerably difficult task because it involves locating faces with no prior knowledge about their scales, locations, orientations, with or without occlusions, and with different poses (frontal, profile) [2]. A powerful face detection algorithm facilitates the design and robustness of the aforementioned systems. Many approaches for face detection have already been proposed. Detailed surveys can be found in [1, 2]. In this paper we are interested in appearance based techniques, and particularly in those built on support vector machines and eigenvector decomposition.

The application of support vector machines (SVM) in frontal face detection in images was first proposed in [3]. Besides the SVMs, eigenvalue decomposition methods constitute a popular class of appearance-based algorithms for face detection. A probabilistic method based on density estimation in a high dimensional space using an eigenvalue decomposition is proposed in [4]. An example-based approach for locating vertically oriented and unconcluded frontal face views at different scales by using a number of Gaussian clusters to model the distributions of face and non-face patterns is described in [5]. We are interested in applying such techniques to patterns derived by some optimization procedure and not the raw pixel intensities.

In this paper, two face detection methods for single images of "head and shoulder" type that contain a uniform background are developed. The methods discussed are based on SVMs and maximum likelihood detection, respectively. Although the just mentioned task is considered to be more simple than face detection in scenes with multiple faces in a complex background, we argue that such a study is still useful, because it reveals a sort of "upper bound" on the performance of face detection algorithms. We train and test the performance of the face detection methods described on sets of single images from M2VTS database [8]. The contribution of this paper is two-fold. First, we propose a feature selection criterion in maximum likelihood detection methods. Second, we attempt an objective evaluation of the performance of the methods discussed. More specifically, throughout the paper, the false acceptance and the false detection rates are considered as quantitative figures of merit. To measure these rates, we address what constitutes a "successful" detection. For this purpose, we have recorded the ground truth bounding box for the faces using a combination of the method described in [7] and human intervention. The criterion for a successful face detection is the center of the detected bounding box must be within the ground truth bounding box and the area of intersection of the ground truth bounding box and the detected one exceeds the 70% of the area of the former. For a comparative study on the performance of several face detection algorithms in scenes with multiple faces in a complex background, the interested reader may refer to [2]. However, for the maximum likelihood face detection method we demonstrate its performance on single images with multiple faces and complex recording conditions, such as occlusion and nonuniform background.

The outline of the paper is as follows. The SVM face detection algorithm is described in Section 2. A probabilistic face detection algorithm based on a feature extraction procedure, like the Kanade-Lucas-Tomasi algorithm [14, 15] is presented in Section 3. Experimental results are reported in Section 4 and conclusions are drawn in Section 5.

2. SUPPORT VECTOR MACHINE APPROACH

The face detection approach that is based on SVMs is applied on running windows defined on the *quartet* image [6]. The quartet image is a mosaic image of reduced resolution, where the macroscopic features of a human face can easily be captured. A two-dimensional (2-D) rectangular window is defined that consists of 5 cells in the horizontal and 6 cells in the vertical direction. The window scans the quartet image whose cell intensities have been normalized to the interval $[0, 1]$. Between two successive movements, the windows are half overlapping. By moving the window over the quartet image, several 30-dimensional patterns are obtained that enable the description of faces appearing at different locations in the image. By varying the cell size, we enable the description of faces at different scales.

Let $\mathbf{x}_i$, $i = 1, 2, \ldots, l$, denote the ith training pattern and $t_i = \pm 1$ the class label assigned to it. Let $\mathbf{t} = (t_1, t_2, \ldots, t_l)^T$. In the general case, $\mathbf{x}_i$ are not linearly separable. That is, there will be no pair $(\mathbf{w}, b)$ such that

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T \mathbf{x} + b) \tag{1}$$

is satisfied throughout the training set, i.e., $t_i = f(\mathbf{x}_i)$, $i = 1, 2, \ldots, l$, where b is an offset parameter and $\mathbf{w} = (w_1, w_2, ..., w_l)^T$ is the normal vector to the separating hyperplane $\mathbf{w}^T \mathbf{x} + b = 0$. We may relax this constraint by introducing slack variables $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_l)^T$ and solving the following quadratic optimization problem subject to inequality constraints:

$$\text{minimize} \qquad \Phi(\mathbf{w}, b, \boldsymbol{\xi}) = \frac{1}{2}||\mathbf{w}||^2 + C\,(\sum_{i=1}^{l} \xi_i)^k \tag{2}$$

$$\text{subject to} \qquad t_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2 \ldots, l$$
$$\xi_i \geq 0, \quad i = 1, 2, \ldots, l \tag{3}$$

where k and C are control parameters that penalize the violations of the linearly separable constraints [9]. The solution of the optimization problem (2) and (3), $(\mathbf{w}^*, b^*, \boldsymbol{\xi}^*)$, must be a stationary point of the objective function [9]. To solve this optimization problem Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_l)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_l)^T$ should be introduced for the sets of constraint functions (2) and (3), respectively. According to Theorem 9.5.1 [9, p. 219] if $(\mathbf{w}^*, b^*, \xi^*)$ solves a convex primal problem and the objective and constraint functions are differentiable, then $\boldsymbol{\lambda}^*$ and $\boldsymbol{\gamma}^*$ solve the dual Wolfe problem which yields the so-called *soft margin hyperplane* [3]. For $k = 1$, the Lagrange multipliers that maximize the dual Wolfe problem subject to $0 < \lambda_i^* \leq C$ define the *support vectors* $\mathbf{x}_i$ that are employed to yield the optimal weight vector

$$\mathbf{w}^* = \sum_{i=1}^{l} \lambda_i^*\, t_i\, \mathbf{x}_i. \tag{4}$$

The decision function implemented by the SVM is then

$$f(\mathbf{x}) = \text{sign}\left[\sum_{i=1}^{l} t_i\, \lambda_i^* (\mathbf{x}^T \mathbf{x}_i) + b^*\right] \tag{5}$$

where $b^* = t_i - (\mathbf{w}^*)^T\, \mathbf{x}_i$, for any support vector $\mathbf{x}_i$.

If the input patterns are mapped to a higher dimensional feature space through some non-linear mapping, the inner products in the feature space can be computed by a positive definite kernel function $K(\mathbf{x},\, \mathbf{x}_i)$ [10]. To implement the above described algorithm, the SVM^{light} $Toolbox$ [11] has been used. To model efficiently the non-face class in the training phase, we have used bootstrapping, as is proposed in [5].

3. PROBABILISTIC APPROACH

A matrix that frequently appears in many problems of computer vision, such as optical flow estimation [12], corner detection [13] is the following:

$$\widehat{\mathbf{Z}} = \begin{bmatrix} \sum_W I_u^2 & \sum_W I_u I_v \\ \sum_W I_u I_v & \sum_W I_v^2 \end{bmatrix} \tag{6}$$

where $I(u, v)$ denotes the image intensity at the pixel (u, v) (i.e., grayscale value), I_u and I_v are the partial derivatives of the image intensity in the horizontal -u - and the vertical direction -v-, respectively, and W is a 7 $\times$ 7 window centered on the candidate pixel. A similar matrix to $\widehat{\mathbf{Z}}$ defined in (6) appears also in feature tracking [14] with the difference that the gradients are applied to the sum of a pair of images. Let λ_1 and λ_2 denote the eigenvalues of matrix Z. In all the aforementioned applications, a least squares problem is solved and the sum of the eigenvalues, $\lambda_1 + \lambda_2$, provides a direct measure of goodness of the data. If the sum is small, a high amount of regularization should be performed. In order for $\widehat{\mathbf{Z}}$ to be well-conditioned, both eigenvalues must be large, and their ratio (i.e., the condition number of $\widehat{\mathbf{Z}}$) cannot be large [15]. In practice, when

$$\min(\lambda_1, \lambda_2) > \tau \tag{7}$$

where τ is a predefined threshold, we accept the image pixel under consideration as a feature. Since the algorithm selects as features those pixels having two large eigenvalues, most of the features represent corners [16]. Accordingly, the feature extraction method can be conceived as a first dimensionality reduction step that aims at facilitating the derivation of an eigenvalue representation of the face patterns, as is proposed in [4]. The number of features then depends on the threshold τ in (7). A method for the calculation of τ based on the calculation of the histogram of the smallest eigenvalue is proposed in [16]. Large neighborhoods lead to less features, whereas small

neighborhoods yield more features that tend to gather in certain areas, producing poor detection results.

The training procedure starts with the feature extraction in the set of training images. The number of feature points extracted usually differs between the training images. In order to choose a unique number of features to be used in any training image, we recorded the number of the extracted features per image and we selected the minimum number of features.

A feature set is created for each of the training images. Each feature set is sorted in decreasing order of the smallest eigenvalue and a correspondence is assumed between the features appearing at the same place after sorting in the different training images. Each image is thus represented by the set of grayscaling values at the pixel coordinates of the features that have been ordered using the just described eigenvalue criterion.

The feature extraction can be seen as a transform. Each image in the training set is mapped to a feature space. The representation of each image in the latter space is reduced. Let N be the number of features extracted that form the training patterns. Obviously, $N \ll N_r \times N_c$, where $N_r \times N_c$ is the dimensionality of the training face images. Let N_T be total number of training images. In practice, it is convenient to select $N < N_T$ so that the covariance matrix has full rank. However, since we will apply subsequently principal component analysis and we will be restricted to the first $M < N_T$ eigenvalues, the constraint $N < N_T$ is not crucial. Having decomposed the covariance matrix of the training patterns to its principal components, we can model the distribution of face patterns by a multidimensional Gaussian, $P(\mathbf{x}|\Omega)$, as is described in [4].

Given a test image, the multistage extension of the procedure described in [4] can be applied to yield a maximum likelihood (ML) estimate of position and scale for a face appearing in a test image. Let us assume that we would like to estimate the density $\widehat{P}(\mathbf{x}|\Omega)$ over a subimage $\mathcal{K}$ of the test image. To do so, first feature extraction in $\mathcal{K}$ should be performed. Then we project the pattern vector formed by the features to the subspace defined by the M principal components derived during training and we evaluate the density $\widehat{P}(\mathbf{x}|\Omega)$. We detect a face if

$$\frac{\widehat{P}(\mathbf{x}|\Omega)}{\widehat{P}(\mathbf{x}|\Omega)_{\max}} > \theta \tag{8}$$

where $\widehat{P}(\mathbf{x}|\Omega)_{\max}$ is the maximum value the density attains over the test image and θ is a threshold. The aforementioned algorithm can be generalized in order to handle multiple face detection. Moreover, a tracking algorithm, like the Kanade-Lucas-Tomato algorithm [14, 15], can assist to the face detection scheme by yielding a set of features that can be tracked reliably.

The proposed detection scheme requires the automatic feature extraction on a training or test image. This is an additional computational effort not existing in [4]. However, this additional step makes the entire face detection algorithm faster than the method in [4], because the dimensionality of the patterns on which principal component analysis is applied is much smaller.

4. RESULTS

The proposed algorithms have been tested on single images from the European ACTS project M2VTS database [8]. The database includes the video-sequences of 37 different persons in four different shots. A training set is built from the one frontal face per person for the 37 persons in three shots. The algorithms are trained on this set. One frontal face image per person for the 37 persons from a fourth shot are used as test images. Rotations between the four available shots by leaving one shot out are also tested.

Two quantitative figures of merit have been used in the assessment of the performance of each algorithm, namely the *false acceptance rate* (FAR) and the *false rejection rate* (FRR) during the test phase. The false acceptance rate is the ratio of non-face examples that have been classified wrongly as faces, while the false rejection rate is the ratio of face examples that have been failed to be detected, i.e., they have been rejected as non-faces. *Receiver operating characteristic* (ROC) curves (i.e., plots of FRR versus FAR) for both detection algorithms are provided.

Let us first assess the performance of the SVM-based face detection algorithm. The pattern extraction algorithm described in Section 2 yields roughly $1-10$ face patterns when each frontal face image is processed at several quartet cell resolutions. Accordingly, on average 200 face patterns result for each shot. When three shots are considered, a training set of 600 face patterns is formed. The following kernels have been employed during the training phase: (a) Linear with $C = 1000$; (b) Polynomial $K(\boldsymbol{\chi}, \boldsymbol{\psi}) = (s\,\boldsymbol{\chi}^T\boldsymbol{\psi} + c)^d$ with $s = c = 1$, $d = 3, 4, 5$ and 10; (c) Radial Basis Function (RBF)$K(\boldsymbol{\chi}, \boldsymbol{\psi}) = \exp(-\gamma||\boldsymbol{\chi} - \boldsymbol{\psi}||^2)$ with $\gamma = 1$ and 5; (d) Sigmoidal $K(\boldsymbol{\chi}, \boldsymbol{\psi}) = \tanh(s\boldsymbol{\chi}^T\boldsymbol{\psi} + c)$ with $c = 1$ and $s = 0.005$.

The ROC curves for face detection algorithms based on the aforementioned kernels are plotted in Figure 1. The ROC curves have been computed on four combinations of test and training sets produced by leaving one shot out and rotating between the available shots. It is seen that the sigmoidal kernel yields the lowest equal error rate (EER) that is approximately 4.5%.

We proceed next to the evaluation of the maximum likelihood face detection algorithm. By varying the threshold θ in (8) an ROC curve can be obtained. Such a ROC curve for $N = 100$ features is plotted also in Figure 1. As can be seen, a comparable EER to that of the SVM-

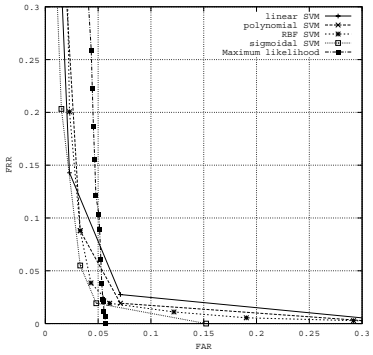

Figure 1: Receiver operating characteristic curves for SVM-based and maximum likelihood face detection algorithms.

based face detection algorithm has been obtained. Moreover, the maximum likelihood face detection algorithm has been proven robust under varying illumination conditions, when multiple faces appear in a scene, and when faces are partially occluded, as can be seen in Figure 2.

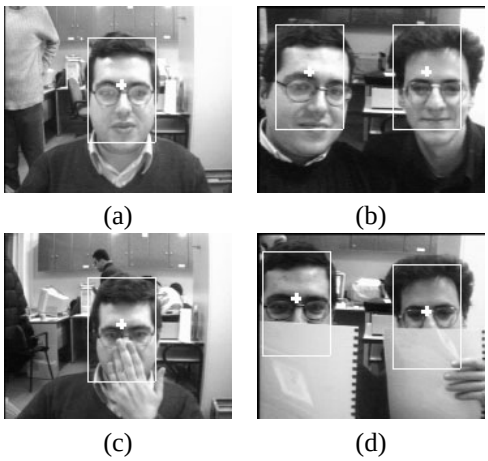

Figure 2: Face detection results (a) when the illumination is not uniform; (b) when multiple faces appear in a scene; (c) when a face is partially occluded; (d) when multiple occluded faces appear in a scene.

5. CONCLUSIONS

In this paper, two methods for face detection in frontal views have been described and their performance has been assessed with respect to the false acceptance and false rejection rates. Both techniques are example-based and offer more flexibility in contrast to the knowledge-based approaches. It has been demonstrated that the attain approximately the same EER. Moreover, the maximum likelihood face detection is shown to perform satisfactorily when the illumination is not uniform and more than one faces that could be partially occluded appear in a scene.

6. REFERENCES

[1] E. Hjelmas and B.-K. Low, "Face Detection: A Survey," *Computer Vision and Image Understanding*, vol. 83, pp. 236–274, 2001.

[2] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, January 2002.

[3] E. Osuma, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection", in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR97)*, pp. 130–136, 1997.

[4] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, 1997.

[5] 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, January 1998.

[6] G. Yang and T.S Huang, "Human face detection in acomplex background," *Pattern Recognition*, vol. 27, no. 1, pp. 53–63, 1994.

[7] C. Kotropoulos and I. Pitas, "Rule-based face detection in frontal views," in *Proc. 1997 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 2537–2540, 1997.

[8] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," in *Lecture Notes in Computer Science: Audio- and Video-Based Biometric Person Authentication* (J. Bigün, G. Chollet, and G. Borgefors, Eds.), vol. 1206, pp. 403–409, 1997.

[9] R. Fletcher, *Practical Methods of Optimization*, 2/e. Chichester, U.K.: J. Wiley & Sons, 1987.

[10] V.N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.

[11] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods: Support Vector Learning* (B. Schölkopf, C.J.C. Burges, and A.J. Smola, Eds.), pp. 41–56, Cambridge, MA: The MIT Press, 1998.

[12] H.H. Nagel, "On the estimation of optical flow: Relations between different approaches and some new results," *Artificial Intelligence*, vol. 33, pp. 299–324, 1987.

[13] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Jth Alvey Vision Conf.*, pp. 147–151, 1988.

[14] C. Tomasi and T. Kanade, Shape and Motion from Image Streams: A Factorization Method - Part 3: Detection and Tracking of Point Features, Tech. Report CMU-CS-91-132, Computer Science Department, Carnegie Mellon University, April 1991.

[15] J. Shi and C. Tomasi, "Good features to track", in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR94)*, pp. 593–600, Seattle, June 1994.

[16] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*. Englewood Cliffs, N.J.: *Prentice Hall*, 1998.