A FEATURE-BASED FACE DETECTOR USING WAVELET FRAMES

C. Garcia, G. Simandiris and G. Tziritas

Department of Computer Science, University of Crete
P.O. Box 2208, 71409 Heraklion, Greece
E-mails: {cgarcia,simg,tziritas}@csd.uoc.gr

ABSTRACT

In this paper, we propose a novel scheme for detection and precise segmentation of human faces in color images where the number, the location, the orientation and the size of the faces are unknown, under non-constrained scene conditions such as complex background and uncontrolled illumination. A deformable template is used as a generic model of the face, defined by stable facial features grouped by anthropometric geometric and textural constraints. The different areas of the face template are characterized by extracting simple statistical measures from suitably selected bands of a wavelet decomposition. The candidate face is classified by applying a set of optimally ordered heuristic and probabilistic tests on the extracted statistical feature vectors. Experimental results are provided to demonstrate the robustness of our approach and its capability to precisely detect faces under varying scale, expression and orientation.

1. INTRODUCTION

Human face processing is becoming a very important research topic, due to its wide range of applications, like security access control, model-based video coding or content-based video indexing. Face recognition and expression analysis algorithms have received most of the attention in the academic literature in comparison to face detection. In recent years, considerable progress has been made on the problem of face recognition, especially under stable conditions such as small variations in lightning, facial expression and pose. An interesting survey may be found in [1].

Most automatic face recognition and expression analysis algorithms have either assumed that the face have been cropped from the image or used "mugshot" images with uniform background so that the face is detected in a trivial way. However, the task of face detection is not trivial in complex scenes. Face patterns can present significant variations due to differences in facial appearance, expression and orientation. Some techniques have been developed recently for detecting faces in "non-mugshot" images. These methods can be roughly divided into three broad categories: local facial features detection, template matching and image invariants. In the first case, low level computer vision algorithms are used to detect facial features such as eyes, mouth, nose and chin and statistical models of human face are used like in [3, 5, 10] among others. In the second case, several correlation templates are used to detect local sub-features. These features can be considered as rigid in appearance (view-based eigenspaces [6]) or deformable (deformable templates [9]). The main drawback of these approaches is that either little global constraints are applied on the face template or extracted features are strongly influenced by noise or change in face expression or viewpoint. In the last case, image-invariant schemes assume that there are certain spatial image relationships, like brightness distribution, common and possibly unique to all face patterns, even under different imaging conditions [8]. They proved not to be robust in non-constrained scenes. Instead of detecting faces by following a set of human-designed rules, alternative approaches are based on neural networks like in [7]. Neural networks have the advantage of learning from representative examples but they have the main drawback of being computationally expensive, hard to train and sensitive to changes in face orientation and expression.

In this paper, we propose a novel feature-based approach based on deformable template matching. The proposed template consists in four facial features (eyes, nose and mouth). In addition we incorporate in the template the 'cheek' regions (under the eyes and left or right of the nose) and an area around the mouth. The deformable template is built with these facial areas and geometric and textural constraints between them. This face model has the advantage of using interior areas of the face which are less subject to instability compared to exterior points lying on the face boundary (like in [9]), easily corrupted by background clutter. Moreover, these points are stable, given that the inter-feature ge-

ometry is not easily deformed by different facial expression or different person identity. Facial features are detected and characterized using statistical measures extracted from a set of selected wavelet bands whereas a set of constraints is imposed on the whole template. The template deformation provides adaptation to faces of varying size and orientation.

2. THE PROPOSED APPROACH

2.1. Facial feature detection

The first step of our algorithm consists in locating candidate interior facial features in the image. The face template will be analyzed at locations where facial features have been found while respecting geometric constraints. First, the image is filtered according to admissible skin color values using an estimated skin color subspace in HSV color space that we presented in [4]. By doing so, the search space for facial features is considerably reduced to areas surrounded by a sufficient amount of skin color points. Note that our face detection scheme may be used without applying this skin color filtering step at the cost of searching the facial features all over the frame.

The facial feature detection is based on a wavelet frame decomposition of the image. A discrete wavelet frame analysis is performed using a pair of suitably chosen conjugate quadrature low-pass and high-pass filters. These filters are described in [4], where we proposed a fast approach for face detection by classifying directly rigid rectangular face bounding boxes without performing facial feature detection. The input image is decomposed into a three-level wavelet frame tree. The first level of the tree contains the wavelet coefficient images A (approximation), H (horizontal details), V (vertical details) and D (diagonal details). The next level is obtained by recursively decomposing each node of the previous level producing wavelet images AA, AH, AV, AD, HA, HH, HV, HD, etc.. This approach offers powerful signal analysis possibilities. Statistical measures will be extracted from these wavelet coefficient images in order to characterize the facial feature textures.

Based on a large number of experiments, we have noticed that the AHH band provides a stable signal for eyes and mouths, representing a characteristic 2D waveform. The nose feature proved to be very unstable in all wavelet bands, in most of our test images, especially because it is poorly textured and strongly corrupted by changes in lighting conditions and face orientation. We therefore decided not to detect the nose in the step of facial feature localization. The information related to the nose will be used in a latter stage of the face classification scheme.

The analysis of the wavelet band signals for facial feature detection is performed in a moving window of 40×40 pixels. The choice of this fixed window size has been motivated by the fact that we aim at detecting facial features of different sizes in one pass only. Such a window is large enough to contain eyes up to the limit of an inter-eyes distance of 60 pixels, which corresponds to the biggest size of faces we aim at detecting, the lower limit being 20 pixels. An example of the AHH signal contained in a window centered on an eye is given in Figure 1.

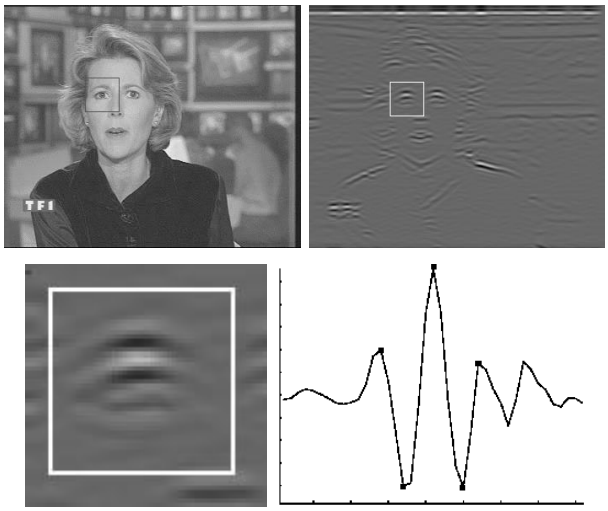

Fig. 1. From top left to bottom right: the original image, the AHH band, the AHH signal in the eye area and the extracted profile.

The vertical waveform characterizes the eyes and permits their detection, while the signal remains roughly invariant in the horizontal direction. For this reason a low-pass horizontal filter is used for extracting a 1-D signal (profile) integrating the vertical characteristics, as shown in Figure 1. As a scale-invariant detector is needed, only the *extrema* of this signal are used for the classification. The data vector on which the decision is based contains three *maxima* and two *minima*. Every profile extracted from the moving window is compared to a reference set of examples, (i.e. profiles corresponding to windows centered on eye points) using the Mahalanobis distance. Every location of the moving window giving rise to a profile with a Mahalanobis distance smaller than an appropriate threshold is classified as an eye point.

The same process is applied for the detection of the mouth.

2.2. Face modelization by a deformable template

A deformable template is used as a geometric model of the face. The template is defined by four elements and their relative positions and sizes: left and right eyes, nose and mouth. The global localization of the template results from the localization of its elements taking into account anthropomorphic geometric constraints. The template is constrained but not rigid. It is elastic in order to allow adaptation to different poses, sizes and shapes.

We perform a set of sequential tests for detecting and localizing the face and its features. During all the process of sequential testing the template is always globally considered. The template is positioned by reference to a detected candidate left-viewed eye, and is reasonably deformed to localize, if verified by facial feature detection, the right-viewed eye and the mouth. If any feature (the two eyes and the mouth) is not detected for a template position, the candidate template is rejected.

In addition to the four facial features, the cheek area, the inter-eye area and the mouth surrounding area are also tested. We have also defined a trapezoidal frame (the face background area) for testing the uniformity of the image intensity inside the template area excluding the facial feature windows. The face template

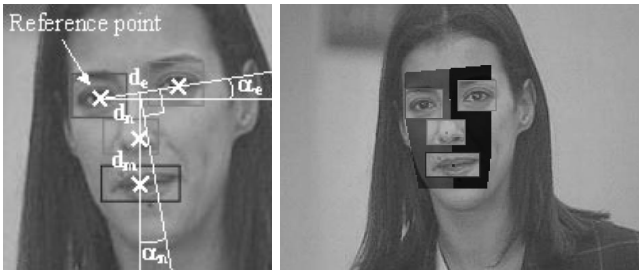

Fig. 2. The parameters of the deformable template and the face background area.

is defined by a set of geometric parameters as described in Figure 2. The face size is determined by the inter-eye distance d_e varying from 20 to 60 pixels (for CIF images, with 288 lines and 360 pixels/line), a common size being 40. The angle α_e between the eye baseline and the horizontal axis is also determined for controlling the face orientation. It is limited to $\pm 20^o$, for permitting the use of horizontal orthogonal frames for the four features. The distance d_n from the eye baseline to the nose is constrained by $0.6d_e \leq d_n \leq 0.8d_e$, while the mouth distance is set to $d_m = 2d_n$. The angle α_n between the nose axis and the normal to the eye baseline has a variation range of $\pm 2\alpha_e$, in order to fit to non-frontal face poses.

The width w and height h of the different facial feature windows are proportional to the inter-eye distance, so that they are automatically adapted during the process of template deformation and scaling: $w = 0.03.w_0.d_e$ and $h = 0.03.h_0.d_e$. The parameters (w_0, h_0) for the eyes, nose, mouth are respectively (25,20), (25,16) and (35,16). The additional windows corresponding to cheeks, inter-eye area and mouth surrounding areas are automatically placed according to the current angles, distances and window sizes (w,h) of the four facial features.

During the search for faces, the face template is moved through the image from top-left to bottom-right and placed according to a candidate left-viewed eye considered as a reference point. Then, the template is deformed according to the admissible range of variations of parameters d_e, d_n, d_m, α_e and α_n and classification of the template content is performed before moving the next position.

2.3. Classification of the template content

As mentioned earlier, the sequence of statistical tests is defined on suitably selected bands of the wavelet frames decomposition. We describe hereafter the tests performed in our detection scheme. They have been ordered to achieve an optimal compromise between discrimination power and computational requirements. All necessary thresholds have been empirically found according to a test set of 100 images.

Local variance analysis Some interior zones of the face are characterized by small variance in certain subbands. More precisely the following tests are performed:

- **Under eye area** This zone is characterized by a small variance in band AH. The choice of this band, which characterizes horizontal details prevents from being influenced by the face boundary when the head is rotated. A threshold, learnt from a set of examples, allows to maintain or reject a candidate eye.
- **Cheek areas** These regions are characterized by small variances in bands AH and AHH. Like in the previous case, threshold values are used.
- **Mouth surrounding areas** The variance of band AH is thresholded in order to reject badly positioned mouth windows. This test improves mouth localization.

Local Energy comparisons A measure invariant to the lighting conditions and to different face poses

is the ratio of the energy in the eye area to the energy in the corresponding cheek area, the latter being higher. A similar relation is expected when the eye area energy is compared to the inter-eye area energy.

Inter-band variance comparisons The eye and mouth areas must have a greater variance in the AH band than in the AV band.

Left/right eye correlation The normalized correlation (on the approximation band A) between the two eye windows should be high, greater than a threshold.

Homogeneity of face skin The area of a trapezoidal frame, defined around the eye and mouth windows and excluding the candidate face features should have a homogeneous image intensity, which is tested by variance thresholding for retaining the candidacy.

A face candidate that passes successfully the above tests is submitted to a global final classification test. In a previous approach for face detection [4], we reported good results obtained by extracting statistical feature vectors from the wavelet coefficients of a global face window and classifying them using the Bhattacharya distance. In the proposed approach, we first extract statistical vectors in each of the four facial feature windows. These vectors provide a textural description of each facial feature. For a given facial feature, a certain number n of wavelet bands are selected and a simple variance σ_i^2 is computed in the facial feature window for each selected band i. Thus, the extracted feature vector contains n components σ_i^2, which describe the texture of the facial feature. We have selected a suitable set of bands for each facial feature. The extracted feature vector is composed from 16 wavelet bands for the eyes (AH, HA, HH, HV, AHA, AHH, AHV, HAA, HAH, HAV, HHA, HHH, HHV, HVA, HVH, HVV), 6 for the nose (AH, HA, HH, AHA, AHH, AHA) and 8 (AH, HH, AAH, AHA, AHH, HHA, HHH, HHV) for the mouth. Under a Gaussian hypothesis, the Bhattacharya distance is defined as follows

$$D_B = \frac{1}{2} \sum_{i=1}^{n} \log \frac{\sigma_i^2 + s_i^2}{2\sigma_i s_i} \qquad (1)$$

where n is the size of the extracted feature vector for the candidate facial feature, s_i^2 are the variances of the corresponding prototype facial feature and σ_i^2 are the measured variances of the candidate facial feature. A prototype feature vector has been computed for each facial feature. A facial feature of the candidate face is accepted if the corresponding distance D_B is less than a threshold value, learnt by experiment on our training set. A face candidate is accepted if all contained facial features are accepted.

2.4. Management of overlapping face candidates

The proposed algorithm usually detects multiple candidates in the area spanned by a face. The most common case corresponds to neighbor templates where the facial features are respectively detected in the same neighbor areas. A connected component analysis groups these

Fig. 3. Examples of overlapping face candidates

neighbor templates into a unique representative template, by computing the centroids of each facial feature window. Other cases of multiple detections exist, where some candidate templates share a common area over the face. Typical examples correspond to cases where the eyes have been correctly detected but the nose or the chin have been wrongly selected as a mouth like in the examples of Figure 3. A case less frequent corresponds to templates overlapping over a face area without sharing eye windows. After connected component grouping, the overlapping candidate template corresponding to the biggest group of initially detected templates is selected as the detected face.

3. EXPERIMENTAL RESULTS

The proposed algorithm has been evaluated using the same test data set as in [4]. This test data set contains images that have been extracted as key-frames from various MPEG videos and especially from the test videos used in the DiVAN project evaluation phase [2]. The video material has been kindly provided by the *Institut National Audiovisuel, France* and by *ERT Radio-television, Greece*. The test data set contains 100 images, most of them being extracted from advertisements, news, movies and external shots. This set of 100 images contains 104 faces (with sizes above the minimal one) and ten images which do not contain faces. They cover most of the cases that the algorithm has to deal with. For this data set, we have obtained a

good detection rate of 91.4% with 9 false alarms, with an average processing time of 19 s per image on a Sun 690-MP station running SunOS 2.4.

In Figure 4, we present some results of the proposed face detection scheme for 12 images of the test data set. These examples include images with multiple faces of different sizes and different poses. False alarms and false dismissals examples are presented as well.

From these examples, it may be observed that our method precisely locate faces of varying size (for an inter-eye distance varying from 20 to 60 pixels). Moreover, tilted or slightly rotated faces are as well successfully detected due to the deformable capability of the template. The number of false alarms is very low because of the application of a considerable number of simple tests in the different area of the template which is not the case for approaches treating the face as a whole like in [4, 7]. False dismissals are mostly due to extreme lighting conditions or face occlusions.

4. CONCLUSION

We have presented a novel scheme for human face detection in complex color images. A deformable template is used as geometric model of the face, defined by facial features and geometric/textural constraints between them. A set of optimally ordered sequential tests is applied on suitably selected bands of a wavelet frame decomposition in the different areas of the elastic face template, resulting in precise detection of face areas. The use of a generic deformable face model allows to detect faces of varying size and orientations. The combined detection of facial features provides a precise segmentation of the detected face and greatly reduces the need for preprocessing in the face or expression recognition stages. As an extension of this work, we believe that the descriptors extracted from the wavelet bands in the different template areas may be used directly in the recognition stage. We are currently working towards the integration of a complete detection-recognition scheme.

5. REFERENCES

[1] R. Chellappa, C.L. Wilson, S. Sirohey. Human and Machine Recognition of faces: A survey. in: *Proceedings of IEEE*, 83(5), 705-740, 1995.

[2] DiVAN: Distributed audio-Visual Archives Network (European Esprit Project EP 24956). http://divan.intranet.gr/info, 1997.

[3] A. Eleftheradis and A. Jacquin Model-assisted coding of video teleconferencing sequences at low

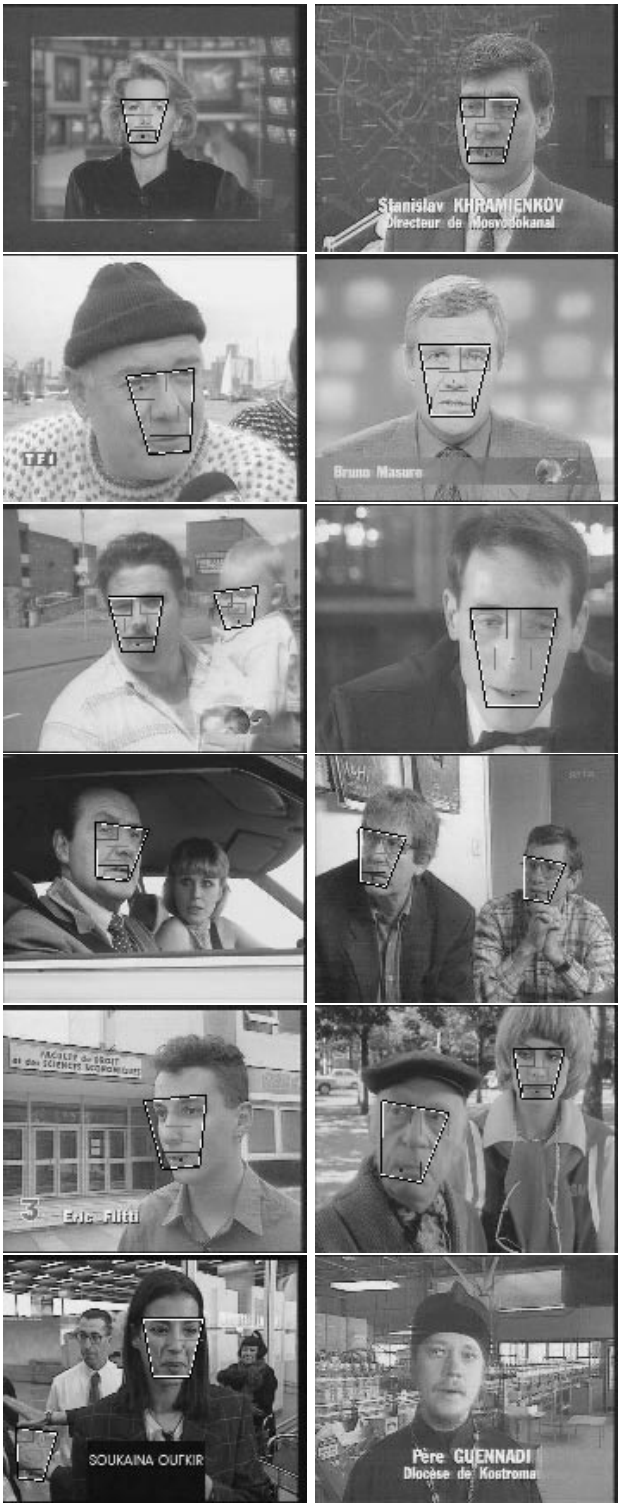

Fig. 4. Some results of the proposed method

bit rates. In *Proceedings of IEEE Int. Symp. Circuits and Systems*, pp. 3.177-3.180, 1994.

[4] C. Garcia and G. Tziritas. Face Detection Using Quantized Skin Color Region Merging and Wavelet Packet Analysis. *IEEE Transactions On Multimedia*,1(3), pp. 264–277, September 1999.

[5] S.-H. Jeng, H. Y. M. Yao, C. C. Han, M. Y. Chern and Y. T. Liu. Facial Feature Detection Using Geometrical Face Model: An Efficient Approach. *Pattern Recognition*, 31(3), pp. 273-282, 1998.

[6] A. Pentland, R.W. Picard, S. Sclaroff. Photobook: Content-Based Manipulation of Image Databases. in: *Proc. of the SPIE Storage and Retrieval and Video Databases II*, 1994.

[7] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20, pp. 23-28, 1998.

[8] P. Sinha. Object Recognition via Image Invariants: A Case Study. *Investigative Ophthalmology and Visual Science*, 35, pp. 1.735-1.740, 1994.

[9] L. Wiskott, JM. Fellous, N. Kruger, C. Von der Malsburg. Face Recognition by Elastic Bunch Graph Matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7), pp. 775-779, 1997.

[10] K. C. Yow, C. Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15, pp. 713-735, 1997.