

to appear in Proc. of AVBPA'97, Crans-Montana, Switzerland, March 1997

Generalized Likelihood Ratio-based Face Detection and Extraction of Mouth Features *

C. Kervrann¹, F. Davoine², P. Pérez¹, H. Li², R. Forchheimer² and C. Labit¹

¹ IRISA/INRIA, Campus Universitaire de Beaulieu, 35042 Rennes cedex, France
{kervrann,perez,labit}@irisa.fr

² Linköping University, Department of Electrical Engineering, Image Coding Group,
S-581 83 Linköping, Sweden
{davoine,haibo,robert}@isy.liu.se

Abstract. In this paper we describe a system to reliably localize the position of the speaker's face and mouth in videophone sequences. A statistical scheme based on a subspace method is presented for detecting human faces under varying poses. We propose a new matching criterion based on the Generalized Likelihood Ratio. The criterion is optimized efficiently with respect to similarity, affine or perspective transform parameters using a coarse-to-fine search strategy combined with a simulated annealing algorithm. Moreover we propose to extract a vector of geometrical features (four points) on the outline of the mouth. The extraction consists in analyzing amplitude projections in the regions of the mouth. All the computations are performed on H263-coded frames, with a QCIF spatial resolution. To this end, we propose algorithms adapted to the poor quality of the images and suited to a further real-time application.

1 Introduction

With the rapid development of computer networks, it is now possible to propose automatic systems dedicated to the authentication of persons. Audio and video signals have been confirmed to be relevant elements for this challenging task. In this paper, we propose a method to localize the head of a person in a video sequence, and to extract pixel location of feature points of the mouth. These reference points can be used to fit the wire frame of a mouth model to the original mouth of a face, and to analyze its motion. We perform our tests on H263-coded QCIF (172×144 pixels) video sequences, with a frame rate of 5 Hz and a bitrate of 20 kbit/s. With such constraints, we fulfill the requirements for PSTN videophony. We recall that H263 is based on a hybrid DPCM/DCT video coding method [8], and thus implies DCT and motion compensation on square blocks, as well as variable length coding and scalar quantization. Such a scheme introduces visible 8×8 block artefacts at low bitrates, mainly on the moving regions.

View-based techniques have been proved to be efficient in object detection [11, 10, 6, 5], recognition [3, 4, 6, 5] and tracking [2]. Moreover, the general problem of detecting a human face from a general point of view remains partially unsolved [11, 10, 5, 9]. In this paper, we present a statistical approach for face detection and person

* This work is supported by the European Commission via the ACTS project VIDAS.

identification which we investigate as an alternative approach to methods reported in [10, 9, 11, 5]. We explore how the distribution-based face model of Moghaddam *et al.* [5] can be extended to general viewing conditions. The starting point of our approach is also inspired from the face detection system based on clustering techniques proposed by Sung *et al.* [10] as well as the neural network-based system described in [9]. In these works, two training sets containing *face* images and *non-face* images are collected. We demonstrate that our system allows to detect a face under variable pose and requires a low computational cost.

A lot of algorithms have been proposed until today in order to localize the characteristic features of the mouth of a talking head, for use by face recognition or automatic lipreading systems. Lots of them make use of deformable templates [12] in order to find the boundary of the mouth. They usually need a rough estimation of the location of the mouth, used as a starting point for the detection, and need a precise extraction of textural properties like edges, peaks and valleys. Our mouth features extraction algorithm focuses on an effective approach based on amplitude projections on straight lines of pixels [3, 7].

2 Statistical Detection of Faces

In this section, we present an unsupervised statistical-based algorithm to detect faces under variable pose on H263-coded QCIF grey-scale videophone images. We exploit a subspace method and an *eigenspace* decomposition to approximate the face appearance using a reduced number of eigenvectors in a Karhunen-Loeve (KL) transform [11, 5, 4, 6]. It is now well known that the subspace method allows to cope with considerable variation of the object appearance. In this section, we start by presenting the automatic visual learning based on density estimation in high-dimensional spaces on two learning image sets showing *face* and *non-face* views. A new matching criterion based on a Generalized Likelihood Ratio (GLR) is then proposed. The detection approach combines the advantages of a compact statistical description of images and an efficient optimization scheme for pose estimation: A multiresolution stochastic search technique is used to locate the best match to the *a priori* model.

2.1 A Distribution-based Face Model

In our approach, we assume that the majority of the face can be modeled by a plane. More complex model (*e.g. 3D models*) can be used but the planar model is simple.

In a face detection task, two adverse hypothesis have to be compared : “presence of one face” (H_1) vs. “presence of no face” (H_0). Characterizing the two classes is challenging because, whereas it is easy to get a training set of faces, it is much harder to collect a representative population of images containing no face [10, 9]. Our system avoids the problem of using a huge training set of *non-face* images and by only using a *non-face* training set of face-like images.

“Face” Model Building. The visual learning of faces consists in building a distribution-based model of frontal view photographs of faces at fixed scale to capture the full range

of permissible variation in patterns. The training procedure relies on the KL transform which allows to identify the degrees of freedom of the statistical variability observed on a training set of representative images, on a low dimensional eigenspace. A particular pattern belonging to the training population of N_T images is represented by the N -dimensional vector \mathbf{x}_k made up from the lexicographic reading of image k , where one element of the vector is a pixel intensity value. The Principal Component Analysis (PCA) is efficient to derive a tractable estimate of the probability distribution $P_{H_1}(\mathbf{x})$ of a particular pattern \mathbf{x} based on the first M principal components [5] ($M \ll N_T \ll N$). This decomposition divides the complete vector space \mathcal{R}^N into a *principal M -dimensional subspace* spanned by the first M principal components and an orthogonal *complementary subspace* spanned by the first $N - M$ other eigenvectors. In the following, we assume that $P_{H_1}(\mathbf{x})$ may be modeled by a multivariate Gaussian density for which the mean vector $\bar{\mathbf{x}}$ and covariance matrix \mathbf{C} are already estimated³. The distribution $P_{H_1}(\mathbf{x})$ may be written from the N projections y_i obtained by the change of coordinates in a KL transform. The distribution estimate $\hat{P}_{H_1}(\mathbf{x})$ is given by a product of two independent Gaussian densities computed from the M principal projections [5]:

$$\hat{P}_{H_1}(\mathbf{x}) = \left[\frac{\exp\left(-\frac{1}{2} \sum_{i=1}^M \frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{M/2} \prod_{i=1}^M \lambda_i^{1/2}} \right] \cdot \left[\frac{\exp\left(-\frac{\varepsilon^2(\mathbf{x})}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}} \right] \quad (1)$$

where λ_i are the eigenvalues associated to the eigenvectors derived from the diagonalization of the covariance matrix \mathbf{C} , ρ is the average of eigenvalues in the *complementary subspace* and ε^2 is the residual reconstruction error defined as:

$$\varepsilon^2(\mathbf{x}) = \sum_{i=M+1}^N y_i^2 = \|\mathbf{x} - \bar{\mathbf{x}}\|^2 - \sum_{i=1}^M y_i^2. \quad (2)$$

The distribution estimate was computed based on 3-dimensional *principal subspace* ($M = 3$) with a *face* training set composed of 40 frontal views of different people⁴.

“Non-face” Model Building. There are many naturally occurring non-face patterns in the real world that look like faces when viewed in isolation. Here, we propose to get a reduced number of significant negative examples which look like faces (*w.r.t.* $\hat{P}_{H_1}(\mathbf{x})$) collected in a “bootstrapping” manner [10, 9]. A distribution-based *non-face* model is built according to the visual learning procedure described previously. If the *non-face* training set contains face-like images, it seem reasonable to model the distribution by a multivariate unimodal Gaussian density similar to $\hat{P}_{H_1}(\mathbf{x})$:

$$\hat{P}_{H_0}(\mathbf{x}) = \left[\frac{\exp\left(-\frac{1}{2} \sum_{i=1}^{M'} \frac{y_i'^2}{\lambda_i'}\right)}{(2\pi)^{M'/2} \prod_{i=1}^{M'} \lambda_i'^{1/2}} \right] \cdot \left[\frac{\exp\left(-\frac{\varepsilon'^2(\mathbf{x})}{2\rho'}\right)}{(2\pi\rho')^{(N-M')/2}} \right]. \quad (3)$$

³ Let's note that \mathbf{x}_k is normalized by its mean and standard deviation to cope with global illumination changes.

⁴ The training procedure has been performed on a database of 56×46 pixel images created at Olivetti Research Laboratory.

A 3-dimensional *principal subspace* ($M' = 3$) allows to derive a satisfactory distribution estimate when the training set is composed of 36 examples (56×46 pixel images).

2.2 Matching Criterion

The detection problem is formulated as comparing the two hypothesis “presence of one face” (H_1) and “presence of no face” (H_0). A central problem in object recognition is to determine the transformation that relates the model to the target object from its appearance in the image. To recognize objects, we frequently seek to eliminate the effects of view points. Our method applies to one face assumed to be approximated by a plane under similarity, affine or perspective transform associated to a parameter vector θ . Our matching criterion aims to estimate any rigid transform θ by maximizing the following Generalized Likelihood Ratio:

$$\begin{aligned} \theta^* &= \arg \max_{\theta} \frac{\hat{P}_{H_1}(\mathbf{x}|\theta)}{\hat{P}_{H_0}(\mathbf{x}|\theta)} \\ \theta^* &= \arg \min_{\theta} \left[\hat{d}_{H_1}(\mathbf{x}(\theta)) - \hat{d}_{H_0}(\mathbf{x}(\theta)) \right] \\ \theta^* &= \arg \min_{\theta} \left[\left(\sum_{i=1}^M \frac{y_i^2(\theta)}{\lambda_i} + \frac{\varepsilon^2(\mathbf{x}(\theta))}{\rho} \right) - \left(\sum_{i=1}^{M'} \frac{y_i'^2(\theta)}{\lambda_i'} + \frac{\varepsilon'^2(\mathbf{x}(\theta))}{\rho'} \right) \right] \end{aligned} \quad (4)$$

A drawback in this approach is that the possible inconsistent measure given by $\hat{d}_{H_0}(\mathbf{x}(\theta))$ can affect grossly the estimator of θ and disturbs the matching process when the input pattern does not contain a face-like view. We design a statistical test under the assumption that $\hat{d}_{H_0}(\mathbf{x}(\theta))$ is distributed as a chi-square distribution with $(M' + 1)$ degrees of freedom, scaled by a factor $N/(M' + 1)$. We propose the following thresholding which will be integrated in the matching criterion :

$$\hat{d}_{H_0}(\mathbf{x}(\theta)) = \zeta \quad \text{if} \quad \hat{d}_{H_0}(\mathbf{x}(\theta)) > \zeta \quad (5)$$

where ζ is inferred from the scaled chi-square distribution law with $(M' + 1)$ degrees of freedom. In our experiments, the threshold ζ corresponds to a 95% confidence. Finally, the face detection is validated if $[\hat{d}_{H_1}(\mathbf{x}(\theta^*)) - \hat{d}_{H_0}(\mathbf{x}(\theta^*))]$ does not exceed a statistical threshold ξ inferred from the scaled chi-square distribution laws of random variables $\hat{d}_{H_0}(\mathbf{x}(\theta))$ and $\hat{d}_{H_1}(\mathbf{x}(\theta))$.

2.3 Computational Issues

In this section, we see that for similarity, affine and perspective transforms, the estimation of θ requires an efficient coarse-to-fine strategy [9]. For every image in the two training sets we construct a pyramid of images by spatial filtering and subsampling. The images at each level in the pyramids form distinct training sets and at each level an PCA is performed to construct the *eigenspace* description of that level. The input QCIF image is similarly smoothed and subsampled. At the coarsest level in the pyramid, only the spatial position and scale parameters are estimated. An exhaustive search



Fig. 1. GLR face detection on 4 intermediate frames of a QCIF videophone sequence (affine transform).



Fig. 2. GLR face detection on 4 intermediate frames of a QCIF videophone sequence (perspective transform).

of this space would take too long time to find a good match if no particular architecture is used [5]. Our search algorithm, based on a fast version of simulated annealing, estimates automatically the 4 parameters and avoids a suboptimal search strategy [10, 9]. It uses a Metropolis dynamic and the temperature cooling is inverse linear [1]. Using deterministic refinement techniques, the affine or perspective transforms are estimated at each level and the matching procedure stops when the algorithm converges at the finest resolution. The total cpu time, on SUN SPARC20 workstation, is respectively 1.5s and 3s to estimate an affine and a perspective transform on the first frame when a three-level Gaussian pyramid is created. The main limitation of previous approaches is that systems detect upright faces looking at the camera [10, 5, 9] and may require high computational complexity [10, 9].

2.4 Experimental Results

In our experiments, the H263-coded QCIF image sequences show a speaker moving against an uniform background (Fig. 1) and a more complex textured background (Fig. 2). The two sequences are originated in the audio-video corpus of the ACTS VIDAS project. The initialization of θ is random on the first highly degraded image of the sequence. The final results of face detection on two different sequences when an affine and a perspective transform are respectively considered, are presented on Fig. 1 and Fig. 2. Let's note that a temporal tracking can be introduced by projecting the final estimate obtained on the current frame as an initial estimate in the next frame. The time computing is then

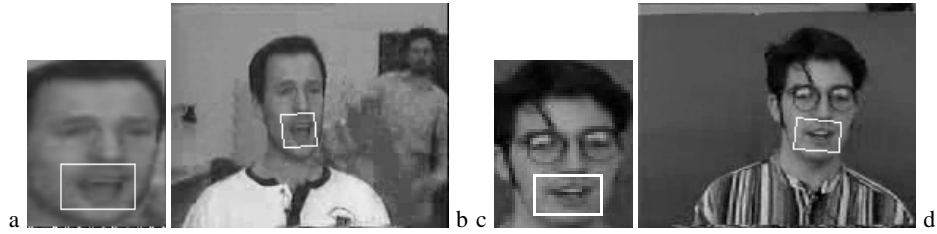


Fig. 3. Maximum Likelihood mouth detections ; a-c) normalized images ; b-d) QCIF images.

notably reduced to $0.2s$ by frame on a SUN SPARC20 workstation to estimate the affine transform. In this case, the use of a SA algorithm and a multiresolution search strategy is avoided (except for the first frame).

3 Detection of Salient Points on the Mouth

3.1 Maximum Likelihood Detection of Mouth

The *eigentemplate* approach to the detection of facial features was presented in [5]. A Maximum Likelihood estimate of the position of the mouth is only presented in this section. The visual learning relies exclusively on a mouth training set in this case. The detected region of interest will be considered as input to the mouth parameters extraction algorithm described in the next section. Once the location of the face has been estimated, the vector parameter Θ^* is used to compensate by affine or perspective transform, yielding a rectangular (56×46) box containing a face. A second feature detection stage operates at this level to estimate the scale and the position parameters of the mouth using a reduced version of our matching algorithm. The effects of view point are not totally eliminated during the face detection step and the *a priori* location of the mouth is then refined using a deterministic optimization algorithm with a low computational cost. The results of mouth detection on normalized images are presented on Fig. 3a and Fig. 3c. The windows containing the mouth are warped into the original image domain on Fig. 3b and Fig. 3d.

3.2 Extraction of Mouth Features

We describe here a fast algorithm which permits to detect the outline of the mouth, composed of the following four feature points: the top of the upper lip, the bottom of the lower lip, and the two corners. The search is confined to the mouth bounding box given from section 3.1.

The poor spatial resolution of the H263 QCIF coded images, and visible artefacts (block effects) on the mouth, limit the performance of typical image analysis tools. Our aim is consequently to propose a simple and effective method which stays robust even if a part of the mouth is blurred. The extraction is done by examining amplitude projections on cross lines inside the bounding box. It can be performed on the normalized images or on the original video sequence. In the first case, we consider vertical and

horizontal lines, and in the second case, we consider lines parallel to the borders of the bounding box. We use an algorithm similar to the Bresenham's algorithm to find intersections between the uniform grid of pixels and oblique line segments having grid vertices as their endpoints.

Let us describe the extraction algorithm, confined to a rectangular bounding box on a normalized face, around the mouth (the same approach is used to process the interior of a parallelogram):

1. Compute the sums of the grey levels on each horizontal line, starting from the top of the box. This point gives the function $sum_H(y)$ (see top of Fig. 4).
2. Look for the first maximum negative slope of the function $sum_H(y)$, starting from the left. This gives the position of the horizontal line on the top of the upper lip.
3. Considering the function $sum_H(y)$, starting from the right, consider two cases:
 - (a) compute the first local minimum of the function (see Fig. 4);
 - (b) if this point does not exist, detect the first maximum negative slope of the function.This point gives the position of the horizontal line on the bottom of the lower lip.
4. Limit now the search to the rectangular part between the two previously detected lines, and consider gradient image, in order to detect the left and right horizontal external points of the mouth. Compute the sums of the gradient values on each vertical line, starting from the left. This gives the function $sum_V(x)$ (see bottom of Fig. 4).
5. The computation of the left and right maximum values of the function $sum_V(x)$ returns the positions of the two vertical lines passing through the corners of the mouth.
6. The minimum values on these lines return the spatial positions of the corners.
7. The two points on the top and bottom of the external contours are then localized on the vertical symmetry axis of the mouth between the two corners.

We present on Fig. 5 a result obtained from our features extraction algorithm, performed on two 30 frames videophone scenes with a fixed and moving background. Our experiments show that the algorithm is robust, and works well on the original QCIF image as well as on the normalized image. The shapes of the functions $sum_H(y)$ and $sum_V(x)$ stay approximately constant in time, provided they are obtained from a box suitably centered on the mouth. This allows an effective detection of the position of the four features with a precision of one or two pixels. The system works for both open and closed mouths, since it considers only the outlines.

4 Conclusion

We have successfully developed an example-based learning technique for representing and automatically detecting views of human faces under variable poses and orientation in H263-coded 255 sequences. The distribution-based model captures pattern variations in *face* and *face-like* images. We have proposed an original matching criterion based on *Generalized Likelihood Ratio* and an efficient multiresolution implementation which is planned to be extended for a real-time version. The system is completed by a robust

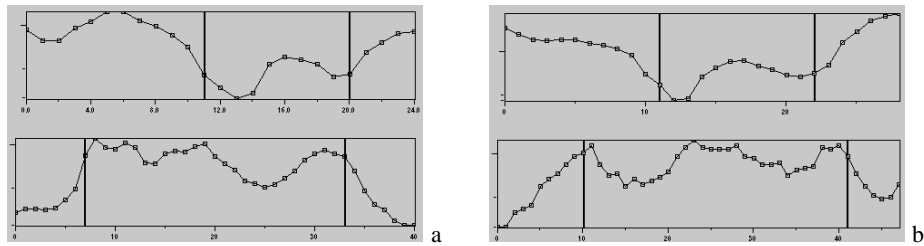


Fig. 4. Functions $sum_H(y)$ (top) and $sum_V(x)$ (bottom) computed from the normalized images on Fig. 5a (left), and Fig.5c (right).

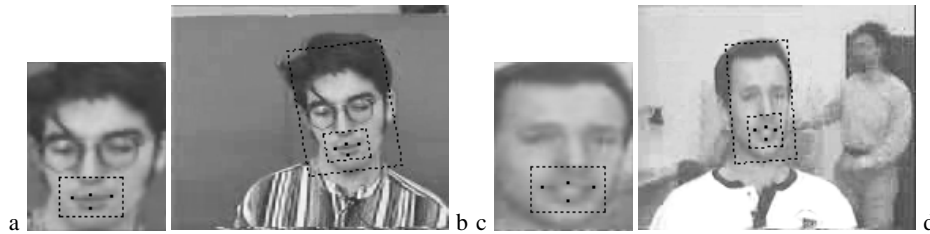


Fig. 5. Extraction of 4 feature points on the mouth ; a-c) normalized images ; b-d) QCIF images.

algorithm which detects salient points on the mouth yielding promising prospects as concerns the characterization and the interpretation of both audio and video signals for a person authentication task.

References

1. M. Betke and N.C. Makris. – Fast object recognition in noisy images using simulated annealing. – In *ICCV95*, pp.523–530, Boston, June 1995.
2. M.J. Black and A.D. Jepson. – Eigenttracking: robust matching and tracking of articulated objects using a view-based representation. – In *ECCV96*, pp.329–342, Cambridge, April 1996.
3. R. Brunelli and T. Poggio. – Face Recognition: Features versus Templates. – *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **15**(10):1042–1052, 1993.
4. A. Lanitis, C.J. Taylor and T.F. Cootes. – An unified approach to coding and interpreting face images. – In *ICCV95*, pp.368–373, Boston, June 1995.
5. B. Moghaddam and A. Pentland. – Maximum Likelihood detection of faces and hands. – In *ICCV95*, pp.786–793, Boston, June 1995.
6. H. Murase and S.K. Nayar. – Visual learning and recognition of 3D objects from appearance. – *Int. J. Computer Vision*, **14**: 5–24, 1995.
7. K.V. Prasad, D.G. Stork and G.J. Wolff, Preprocessing video images for neural learning of lipreading. – In *SPIE Proc. Substance Identification Analytics*, **2093**, pp.116–127, 1994.
8. K. Rijkse. – ITU standardization of very low bitrate video coding algorithms. – *Signal Processing: Image Communication*, **7**: 553–565, 1995.
9. H.A. Rowley, S. Baluja and T. Kanade. – Neural network-based face detection. – In *CVPR96*, pp.203–208, San Francisco, June 1996.

10. K. Sung and T. Poggio. – Example-based learning for view-based human face detection. – In *Technical Report AIM-1521*, MIT, 1994.
11. M. Turk and A. Pentland. – Eigenfaces for recognition. – *J. of Cognitive Science*, **3**(1):1–24, 1991.
12. A.L. Yuille, P.W. Hallinan and D.S. Cohen. – Feature Extraction from Faces Using Deformable Templates. – *Int. J. of Computer Vision*, **8**(2):99–111, 1992.