Generalized Likelihood Ratio-based Face Detection and Extraction of Mouth Features[1]

C. Kervrann[‡], **F. Davoine**[†], **P. Pérez**[‡], **R. Forchheimer**[†] and **C. Labit**[‡]

‡ IRISA/INRIA
Campus Universitaire de Beaulieu
TEMIS project
35042 Rennes Cedex, France
{kervrann,perez,labit}@irisa.fr

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

Submission to PRL special issue on
Audio- and Video-based Biometric Person Authentication

Abstract

In this paper we describe a system to detect the speaker's face and mouth in videophone sequences. A statistical scheme based on a subspace method is described for detecting and tracking faces under varying poses. A matching criterion based on a Generalized Likelihood Ratio is optimized efficiently with respect to a perspective transformation using a coarse-to-fine search strategy combined with a simulated annealing algorithm. Moreover, we analyze the amplitude projections around the speaker's mouth to describe the shape of the lips. All computations are performed on lossy H263-coded images. The proposed algorithms are well-suited to a further real-time implementation.

Keywords : Face detection and tracking, mouth features extraction, principal components analysis, simulated annealing.

AUTHOR TO CONTACT :

FIRST (Personal) NAME :	Charles
LAST (Family) NAME :	KERVRANN
ADDRESS :	IRISA/INRIA, Campus universitaire de Beaulieu 35042 Rennes Cedex, France
TELEPHONE NUMBER :	(+33) 2 99 84 74 29
FAX NUMBER :	(+33) 2 99 84 71 71
E-MAIL ADDRESS :	kervrann@irisa.fr

[1]This work is supported by the European Commission via the ACTS project VIDAS.

1 Introduction

With the rapid development of computer networks, it is now possible to propose automatic systems dedicated to the audio- and video-based authentication of persons. Such fully automatic recognition systems can be usually decomposed into several independent processing steps. First, detecting human faces automatically in a cluttered scene remains an indispensable task. This classical computer vision problem has been proved to be difficult because there can be huge variations in the appearance of face patterns. In this paper, we propose a contribution to this particular problem. We describe a view-based method to detect and track the moving speaker's face and a technique to extract feature points around the mouth. Our system has many potential applications and will be exploited in the context of the European ACTS-VIDAS (*VIDeo ASsisted with audio coding and representation*) project to analyze the labial activity of the speaker.

Traditional rigid template pattern matching techniques and geometrical model-based object recognition approaches [9, 4] tend to perform inadequately for detecting faces. Because of high photometric variations in the appearance of faces that reflect changes in expressions, 3D orientations, lighting conditions, hair styles and so on, it is difficult to parametrize robustly generic models. *Eigenspace* decompositions, applied to an example set, are well-suited to provide this compact parametric description of the object appearance. These approaches extract an orthonormal set of eigenimages. This set spans the visual object space and its dimension defines the degrees of freedom of the deformable photometric template. Rigid template matching using only one reference image can be considered as a degenerate case of this approach. View-based techniques such as *principal components analysis* (PCA) provide *a priori* knowledge about object-specific deformations and have been proved to be efficient for finding deformable objects [24, 23, 14, 13, 11, 5], recognition [4, 11, 14, 13] and tracking [2, 7]. However, the general problem of detecting human faces remains partially unsolved [24, 23, 13, 18] because most of systems detect only vertically oriented and unoccluded frontal views of faces looking at the camera. In the first part of this paper, we present an alternative statistical approach to methods reported in [23, 18, 24, 13]. Our system detects faces under varying poses and orientations and at any scales in grey-level images and requires a low computational cost. We explore how the distribution-based face model of Moghaddam *et al.* [13] can be extended to general viewing conditions. Besides, the starting point of our approach is inspired from the face detection system based on clustering techniques proposed by Sung *et al.* [23] and on the neural network-based system described in [18]. In these works, two training sets containing *face* images and *non-face* images are collected.

However, the view-based technique we propose can not provide accurate positions of facial features, but permits to locate the position of speaker's mouth, by considering a mouth training set. This modular technique permits to return a window around the mouth that can be used to make easier the task of other algorithms that aim to extract the precise shape of the speaker's mouth.

In order to extract feature shapes on a speaker's face, different techniques have been proposed, making use of Karhunen-Loeve expansions [11], elastic or parametric templates [10, 26], active meshes [25], dynamic programming [4] or image invariances. In the second part of this paper, we describe an effective approach based on amplitude projections on straight lines of pixels. Our approach is voluntarily simple because of block distorsions that can appear in the region around the speaker's mouth. Such artefacts usually affect algorithms that strongly rely on textural properties in images like thresholded edges or grey level peaks and valleys. Our aim consists in detecting four extrema points on the mouth. These points can be further exploited to guide face recognition or automatic lipreading tasks. In the context of the project VIDAS, these points are used as a priori knowledge to animate and interpolate the speaker's mouth region by means of a mouth-adapted wireframe.

The remainder of this paper is organized as follows. In section 2, we examine the notations and basic theory related to signal detection. We propose a method to detect a face under variable poses in still images and videophone images sequences. In section 3, we exploit a simplified version of the previous algorithm for the detection of bounding box around the mouth of the speaker, and present a method to define the outline of the mouth. Experimental results on real-world image sequences are shown in section 2 and 3. Tests are performed on H263-coded[2] QCIF (172×144 pixels) video sequences, with a frame rate of 5 Hz and a bitrate of 20 Kbits/s. Such coding schemes introduce visible block (8×8) artefacts at low bitrates, mainly on the moving regions.

2 Statistical Detection of Faces

In this section, we present an unsupervised statistical-based algorithm to detect faces under variable poses on H263-coded QCIF grey-scale videophone images. Two hypotheses are set to indicate presence (hypothesis H_1) or absence (hypothesis H_0) of a face. The decision is taken according to a likelihood ratio test, a fundamental tool in decision theory which has proven to

[2]The H263-standard is based on a hybrid DPCM/DCT video coding method [17], and thus implies DCT and motion compensation on square blocks, as well as variable length coding and scalar quantization (PSTN videophony).

be quite robust and efficient in a number of vision problems :

$$\begin{cases} \mathsf{H}_1 & : \text{ if } \quad \frac{\mathsf{P}_{\mathsf{H}_1}(\mathbf{x})}{\mathsf{P}_{\mathsf{H}_0}(\mathbf{x})} > \zeta \\ \mathsf{H}_0 & : \text{ otherwise} \end{cases} \tag{1}$$

where $\mathsf{P}_{\mathsf{H}_1}(\mathbf{x})$ and $\mathsf{P}_{\mathsf{H}_0}(\mathbf{x})$ are the likelihood functions of a pattern $\mathbf{x}$ associated with H_1 and H_0 respectively and ζ is a predetermined threshold. In our approach, these two likelihood functions are specified by training.

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. We exploit a subspace method to approximate the object appearance using a reduced number of eigenvectors in a Karhunen-Loeve (KL) transform [24, 13, 11, 14, 2]. The main assumption behind this modeling is that the space of grey-level images we consider is linearly spanned by a set of example *face* images. A matching criterion based on a Generalized Likelihood Ratio (GLR) [19] can then be derived for finding faces undergoing geometric distortions in images. Our face finder system combines the advantages of a compact statistical description of illumination in *face* 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 Likelihood functions estimation

We assume that the majority of the face surface (facial mask) can be modeled by a plane. More complex model (*e.g. 3D models* [12]) can be used but the planar model is simple.

In a face detection task, two adverse hypotheses 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 [23, 18]. In our system, we avoid the problem of using a huge training set of *non-face* images by using only a training subset of *non-face* images called ≪*pseudo-face*≫ images. This distinction between the *non-face* class and the *pseudo-face* class is discussed in section 2.1.3. Finally, the visual learning consists in building a distribution-based model of view photographs of *faces* and *pseudo-faces* to capture the full range of permissible variation in patterns.

2.1.1 A distribution-based pattern model

The training procedure relies on the KL transform which enables to identify the degrees of freedom of the statistical variability observed on a training set of representative images. This

KL transform identifies the low dimensional principal subspace in which the final matching field must lie. 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 gathering of pixels in image k. Each element of the vector is a pixel intensity value. The principal components analysis (PCA) is efficient to derive a tractable estimate of the probability distribution $\mathsf{P}(\mathbf{x})$ of a particular pattern $\mathbf{x}$ based on the first M principal components ($M \ll N_T \ll N$) [13]. This decomposition divides the complete vector space $\mathcal{R}^N$ into two orthogonal subspaces : a *principal M-dimensional subspace* spanned by the first M principal components and a *complementary subspace* spanned by the first $N-M$ other eigenvectors. In the following, we assume that $\mathsf{P}(\mathbf{x})$ may be modeled by a multivariate Gaussian density for which the mean vector $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{C}$ are already estimated. We will also assume that $\mathbf{x}_k$ is normalized by its mean and standard deviation to cope with global illumination changes. The distribution $\mathsf{P}(\mathbf{x})$ may be written from the N projections $\{y_i\}_{i=1}^N$ given by the KL transform. The distribution estimate $\widehat{\mathsf{P}}(\mathbf{x})$ is given by a product of two independent Gaussian densities computed from the M principal projections [13] :

$$\widehat{\mathsf{P}}(\mathbf{x}) \;=\; \mathcal{G}_{\mathbf{x}}(\Lambda, M) \;=\; \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right] \cdot \left[\frac{\exp\left(-\frac{\varepsilon^2(\mathbf{x})}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right] \tag{2}$$

where $\Lambda = \{\lambda_i\}_{i=1}^M$ are the M largest 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 :

$$\varepsilon^2(\mathbf{x}) \;=\; \sum_{i=M+1}^{N} y_i^2 \;=\; \|\mathbf{x}-\bar{\mathbf{x}}\|^2 - \sum_{i=1}^{M} y_i^2. \tag{3}$$

In the following, the estimated Mahalanobis distance will be noted as :

$$q(\mathbf{x}) \;=\; \sum_{i=1}^{M}\frac{y_i^2}{\lambda_i} + \frac{\varepsilon^2(\mathbf{x})}{\rho}. \tag{4}$$

The likelihood functions associated with H_1 and H_0 are directly derived from this modeling.

2.1.2 ≪Face≫ Model Building.

The distribution-based *face* model is computed for a 3-dimensional *principal subspace* ($M_f = 3$) from a *face* training set composed of 40 frontal view photographs of different people at fixed scale. The pictures come from a database of 56×46 pixels images created at Olivetti Research Laboratory. The *face* images are roughly geometrically normalized using a global transformation (Fig. 1). We assume reasonably that the likelihood function under H_1 is then modeled by a

multivariate Gaussian density. The likelihood estimate is then given by the product of two independent Gaussian densities as described in the previous section :

$$\widehat{\mathsf{P}}_{\mathsf{H}_1}(\mathbf{x}) \;=\; \mathcal{G}_{\mathbf{x}}(\Lambda_{\mathrm{f}}, M_{\mathrm{f}}). \tag{5}$$

In eigenimages, presented in Fig. 5, only a fraction of the eigenimages is responsible for a specific mode of variation. Arbitrary frontal-view of faces are then generated by a relevant set of coefficients (modal amplitudes) that control the main deformation modes. The first variation mode which controls the lighting conditions and the addition of beards is illustrated in Fig. 2.

Figure 1 to be placed here

Figure 2 to be placed here

2.1.3 ≪Non-face≫ Model Building.

There are many naturally occurring *non-face* patterns in the real world that look like faces when viewed in isolation. Practically, any image can serve as a *non-face* example because the space of *non-face* images is much larger than the space of *face* images. However, collecting a representative set of non-faces is difficult. Here, we propose to select a reduced number of significant negative examples which look like faces (*w.r.t.* $\widehat{\mathsf{P}}_{\mathsf{H}_1}(\mathbf{x})$) collected in a "bootstrapping" manner [23, 18]. It seems more adequate to collect a set of *pseudo-face* images which is a subset of *non-face* images (Fig. 3). The *pseudo-face* images are selected by running a simplified version of our face finder on images containing any faces. The simplified system exploits exclusively a training set of faces and can fail to detect one face in a cluttered scene. The incorrectly diagnosed face detection images are adding to the *pseudo-face* training set (Fig. 3). A distribution-based *pseudo-face* model is further built according to the visual learning procedure described in section 2.1.1. The 3 first *pseudo-face* eigenimages and the first corresponding mode of grey-level variation are respectively presented in Fig. 5 and Fig. 4. A 3-dimensional *principal subspace* ($M_{\mathrm{pf}} = 3$) enables to derive a satisfactory distribution estimate when the training set is composed of 36 examples (56×46 pixels images).

Figure 3 to be placed here

Figure 4 to be placed here

The likelihood estimate is given by the product of two independent Gaussian densities $\mathcal{G}_{\mathbf{x}}(\Lambda_{\mathrm{pf}}, M_{\mathrm{pf}})$. The likelihood function under H_0 is then modeled as a mixture of two densities

[27] :

$$\widehat{\mathsf{P}}_{\mathsf{H}_0}(\mathbf{x}) \;=\; \delta\left(q_{\mathrm{pf}}(\mathbf{x}) < \eta\right) \cdot \mathcal{G}_{\mathbf{x}}(\Lambda_{\mathrm{pf}}, M_{\mathrm{pf}}) \;+\; \delta\left(q_{\mathrm{pf}}(\mathbf{x}) \geq \eta\right) \cdot \xi \tag{6}$$

The distribution ξ in the mixture is meant to model outliers. An outlier describes any texture image that depicts neither a face nor a pseudo-face. One common approach is to choose this distribution to be gaussian with a large variance. We prefer rather to use $0 \leq \xi \leq 1$ to be a constant value inferred from η. Finally, the fixed probability of selecting ζ or $\mathcal{G}_{\mathbf{x}}(\Lambda_{\mathrm{pf}}, M_{\mathrm{pf}})$ is given by the mixture probabilities $\delta\left(q_{\mathrm{pf}}(\mathbf{x}) < \eta\right)$ and $\delta\left(q_{\mathrm{pf}}(\mathbf{x}) \geq \eta\right)$ where $\delta(\cdot)$ designates the Kronecker symbol. A particular layer (arbitrary texture *vs.* pseudo-face) is selecting if the Mahalanobis distance $q_{\mathrm{pf}}(\mathbf{x})$ exceeds a predetermined threshold :

$$q_{\mathrm{pf}}(\mathbf{x}) \;=\; \sum_{i=1}^{M_{\mathrm{pf}}} \frac{y_i^2}{\lambda_i^{\mathrm{pf}}} + \frac{\varepsilon^2(\mathbf{x})}{\rho_{\mathrm{pf}}} \;<\; \eta. \tag{7}$$

$q_{\mathrm{pf}}(\mathbf{x})$ has a distribution which is not chi-square. However, we assume that the weighting sum can be roughly approximated by using only one "scaled chi-square" random variable ν [5]. Statistical moments of ν are then matched with to those of $q_{\mathrm{pf}}(\mathbf{x})$:

$$q_{\mathrm{pf}}(\mathbf{x}) \;=\; \frac{N}{k}\,\nu \quad \text{where} \quad \nu \sim \chi^2(\nu, k) \tag{8}$$

The threshold η is finally inferred from the chi-square distribution with k degrees of freedom given by :

$$k \;=\; \left\lfloor \frac{N^2}{M_{\mathrm{pf}} + \sum_{i\,=\,M_{\mathrm{pf}}+1}^{N} \frac{\lambda_i^{\mathrm{pf}2}}{\rho_{\mathrm{pf}}^2}} \right\rfloor \tag{9}$$

where $\lfloor\cdot\rfloor$ denotes the floor function. The distribution estimate under H_0 is directly influenced by η corresponding to a 95% confidence in practice. Once η is known, the likelihood estimate $\widehat{\mathsf{P}}_{\mathsf{H}_0}(\mathbf{x})$ is derived analytically.

2.2 Matching Criterion

A central problem in object recognition is to determine the transformation that relates the model to appearance of the target object in the image. To recognize objects, we frequently seek to eliminate the effects of view points. Assuming that one face can be approximated by a plane, our method applies a similarity, affine or perspective transformation. The parameters of the geometric transformation are represented by a vector Θ. Our matching criterion aims at estimating the optimal rigid transformation Θ that maximizes the following Generalized Likelihood Ratio [19] :

$$\Theta^* \;=\; \arg\max_{\Theta} \frac{\widehat{\mathsf{P}}_{\mathsf{H}_1}(\mathbf{x}|\Theta)}{\widehat{\mathsf{P}}_{\mathsf{H}_0}(\mathbf{x}|\Theta)} \;>\; \zeta \tag{10}$$

Our system computes several likelihood ratios given the likelihood functions specified in section 2.1 and selects the best one. Fig. 5 describes the overview of our face detection system relying on simple and straightforward computations. Finally, the face detection is validated if the optimal likelihood ratio does not exceed a statistical threshold ζ inferred from the scaled chi-square distributions of $q_{\mathrm{pf}}(\mathbf{x}(\Theta))$ and $q_{\mathrm{f}}(\mathbf{x}(\Theta))$.

Figure 5 to be placed here

2.3 Computational Issues

Figure 6 to be placed here

The algorithmic procedure to estimate the geometric transformation is implemented using a coarse-to-fine strategy [18] described in Fig. 6. For every image in the two training sets we construct a Gaussian pyramid of images by spatial filtering and subsampling. The images at each level in the pyramids form distinct training sets and at each level a PCA is performed to construct the *eigenspace* description of that level. The input QCIF image to be processed is similarly smoothed and subsampled to provide a multiresolution representation over 3 scales. At the coarsest level, the face finder provides an initial estimate for the spatial position and scale parameters. Our search algorithm, based on a fast version of simulated annealing, estimates automatically the 4 parameters and avoids a suboptimal search strategy or an exhaustive search [23, 18]. It uses a Metropolis dynamics and the temperature cooling is inverse linear [1]. Using deterministic refinement techniques, the geometric transformation is estimated at each pyramid level, from the coarser level to the finer one, and the matching procedure stops when the algorithm converges at the finest resolution. We have implemented our face detector on a SUN SPARC20 workstation. Once the input 144×176 pixels image is loaded, the multiresolution processing takes respectively 1.5s and 3s to estimate an affine or a perspective transformation.

2.4 Experimental Results

Finally, we present a set of experiments illustrating the performance of our face finder. To test the parametrized matching technique, we run the system on arbitrary still images (Fig. 7). The white window denotes the region selected as the best match by the detection algorithm. Our algorithm can detect about 90% of faces in a set of 40 total images. The system was tested on a wide variety of images, with many faces and unconstrained backgrounds. All the face patterns

that the face finder missed had either strong illumination shadows or low quality scans.

Figure 7 to be placed here

The matching scheme can be used to track faces undergoing changes in viewpoint. The algorithm tracks a frontal-view of the face and normalizes it using the geometric transformation for each frame. The initialization of Θ is random on the first frame of the sequence. A prediction scheme based on Kalman filtering techniques is used to propagate the initial guess for $\boldsymbol{\Theta}$ from frame to frame. The time computing is then notably reduced to $0.2s$ by frame on a SUN SPARC20 workstation to estimate the affine transformation. In this case, the use of a SA algorithm and a multiresolution search strategy is avoided (except for the first frame). In our experiments, we consider the problem of tracking faces in H263-coded QCIF image sequences in which the speaker is moving against an cluttered background (Fig. 8) or a more complex textured moving background (Fig. 9). The low temporal resolution of video sequence is 5 Hz. The two sequences are originated in the audio-video corpus of the ACTS-VIDAS project. Fig. 8 shows excerpts from the run. In the first row of images, the white box denotes the region selected as the best match by the tracking algorithm. The second row of images is a magnified view of the region in the white box after warping by the estimated geometric (affine) transformation. All normalized images should be quite identical if the geometric transformations are robustly estimated. In fig.9, several key frames from a more difficult sequence are shown. The 2 first rows of images show the tracking with an affine transformation and the 2 next ones with a perspective transformation. The affine model seems limited to track the face when the view points are quite far from the frontal orientation. In conclusion, we observed that the system succeeds to extract reliably the face as long as the view of the speaker belongs to intervals spaced from $\pm 45^o$ along the horizontal and vertical lines.

Figure 8 to be placed here

Figure 9 to be placed here

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 proposed by [13]. In this section, only a Maximum Likelihood estimate of the position of the mouth is presented. The

visual learning relies exclusively on a mouth training set in this case. The detected region of interest will be considered as an 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 it by affine or perspective transformation, yielding a rectangular (56×46) box containing the normalized 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 because the face is non-planar. As a result, the *a priori* location of the mouth is refined using a deterministic optimization algorithm with a low computational cost. The results of mouth detection on normalized face images are presented on Fig. 10a and Fig. 10c. The windows containing the mouth are warped into the original image domain on Fig. 10b and Fig. 10d.

Figure 10 to be placed here

3.2 Detection of the mouth outline

Extraction of the mouth shape is useful in many applications such as speaker authentication based on biometric data, facial expressions recognition, microphone orientation for videoconferencing, video-aided audio processing, bimodal speech recognition and real-time computer graphics animation. The detection task is however made difficult because of the wide range of allowable shape variations of the mouth. Typical extraction algorithms are making use of snakes or deformable correlation templates [26, 10] that build models based on prior knowledge of the mouth structure. The templates are aligned with detected properties in the image intensity, such as edges, valley and peak fields. A wide class of methods are moreover considering the specific color of the human skin and human mouths, by using statistical chromaticity models [20, 15, 21]. Other methods are making use of active meshes in order to fit a set of interconnected points on lips edges [25]. These methods need reference points that are usually detected by analysing the summation of the magnitude of the image gradient along vertical and horizontal lines, in the mouth region. Such methods based on local intensity profiles or gradient projections have been broadly investigated in the literature for facial features extraction [4, 22, 16, 6], and was initially proposed by T. Kanade in [8] for face recognition.

We describe in this section 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 poor resolution of the H263-coded QCIF images that we consider,

and visible artefacts (block effects) on the mouth, limit the performance of typical image analysis tools. Our aim is consequently to propose a simple method and robust to partial occlusions of the mouth. The search is confined to the bounding box around the mouth, which has been previously detected (see section 3.1). The extraction is performed by examining amplitude integral projections on cross lines of image intensities and gradients, inside the mouth bounding box (see Fig. 11). The correct orientation of the bounding box makes our algorithm independent of the orientation of the mouth. The points detection can be performed on the normalized facial images, after compensation by the geometric transform associated to the parameter Θ, or directly 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 quadrilateral bounding box. We use a method 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.

Figure 11 to be placed here

Let us describe the extraction algorithm, confined to an horizontal and rectangular bounding box $[x_1, x_2] \times [y_1, y_2]$ on a normalized face, around the mouth (the same approach is used to process the interior of a quadrilateral in the non compensated face). The image is first bandpass-filtered to reduce the effect of the noise. We call this new image $I(x, y)$.

1. Compute the sums of the grey levels on each horizontal line, starting from the top of the box. This point gives the function $sum_H(y)$ (top row on Fig. 12)

$$sum_H(y) = \sum_{x=x1}^{x2} I(x, y)$$

2. Look for the first maximum negative slop of the function $sum_H(y)$, starting from the left side. This point 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 side, consider two cases:

   (a) Compute the first local minimum of the function (see Fig. 12).

   (b) If this point does not exist, detect the maximum negative slop of the function (see Fig. 14.a).

   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 the gradient image $I'(x, y)$, 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, in order to obtain the function $sum_V(x)$ (bottom row on Fig. 12)

$$sum_V(x) = \sum_{y=y1}^{y2} I'(x, y)$$

5. Compute the extreme left and right maximum values of the function $sum_V(x)$. These two points give 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.

Figure 12 to be placed here

Figure 13 to be placed here

We present on Fig. 13 and Fig. 15 results obtained from the proposed features extraction algorithm, performed on QCIF H263 videophone scenes with fixed and moving backgrounds, and a bitrate equal to 20 Kbits/s. Our experiments confirm that the algorithm is robust, as well on the original image as on the normalized image (the detection is respectively done inside an oriented quadrilateral, or inside an horizontal rectangle). The global shapes of the functions $sum_H(x)$ and $sum_V(y)$ stay approximately constant in time, provided they are obtained from a box suitably centered on the mouth, and that the intensity structures (grey level shape of the mouth) stay visible in the decoded frames. We notice that this is not the case when too many objects are moving in the background of a video sequence encoded at a fixed bitrate. This given example gives rise to false detections, with points typically on the borders of the detected mouth bounding box. Under normal conditions, the algorithm allows an effective detection of the position of the four points with a precision of one or two pixels. The system works well for both open and closed mouths, since we consider only outlines.

Figure 14 to be placed here

Figure 15 to be placed here

4 Conclusion

We have successfully developed an example-based learning technique for representing and automatically detecting and tracking views of human faces under variable poses and orientations in H263-coded sequences. We proposed distribution-based models to capture pattern variations in *face* and *pseudo-face* images. We have described an original matching criterion based on a *Generalized Likelihood Ratio* and an efficient multiresolution implementation which is planned to be adapted for real-time applications. The method is simple and straightforward and provides solution to the problem of tracking faces undergoing geometric distortions. The system is completed by a robust 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 *Int. Conf. on Computer Vision (ICCV95)*, pp.523–530, Boston, USA, June 1995.

[2] M.J. Black and A.D. Jepson. – Eigentracking : robust matching and tracking of articulated objects using a view-based representation. – In *European Conf. on Computer Vision (ECCV96)*, pp.329–342, Cambridge, UK, April 1996.

[3] R. Brunelli. – Face recognition: Dynamic Programming for the detection of face outline. – *Tech. Report 9104-06*, I.R.S.T, 1991.

[4] R. Brunelli and T. Poggio. – Face Recognition : Features versus Templates. – *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **15**(10): 1042–1052, 1993.

[5] T. Cootes, G.J. Page, C.G. Jackson and C.J. Taylor. – Statistical grey-level models for object location and identification. – *Image and Vision Computing*, **14**: 533–540, 1996.

[6] G. Galicia and A. Zakhor. – Depth based recovery of human facial features from video sequences. – In In *IEEE Int. Conf. on Image Processing (ICIP95)*, pp. 603–606, Washington D.C., USA, October 1995.

[7] G. Hager and P. Belhumeur. – Real-time tracking of image region with changes in geometry and illumination. – In *Computer Vision and Pattern Recognition (CVPR96)*, pp.403–410, San Francisco, USA, June 1996.

[8] T. Kanade. – Picture processing system by computer complex and recognition of human faces. – In *Tech. Report, Kyoto University, Dept. of Information Science*, November 1973.

[9] T. Kanade. – Computer recognition of human faces. – Basel and Stuttgart : Birkhauser 1977.

[10] M. Kass, A. Witkin and D. Terzopoulos. – Snakes : active contour models. – In *Int. J. Computer Vision*, **1**: 321–331, 1988.

[11] A. Lanitis, C.J. Taylor and T.F. Cootes. – An unified approach to coding and interpreting face images. – In *Int. Conf. on Computer Vision (ICCV95)*, pp.368–373, Boston, USA, June 1995.

[12] H. Li, P. Roivonen and R. Forchheimer. – 3D motion estimation in model-based facial image coding. – *IEEE Trans. on Pattern Analysis and Machine Intelligence* **15**: 545–555, 1993.

[13] B. Moghaddam and A. Pentland. – Maximum Likelihood detection of faces and hands. – In *Int. Conf. on Computer Vision (ICCV95)*, pp.786–793, Boston, USA, June 1995.

[14] H. Murase and S.K. Nayar. – Visual learning and recognition of 3D objects from appearance. – *Int. J. Computer Vision*, **14**: 5–24, 1995.

[15] N. Oliver, A. Pentland, and F. Bérard. – LAFTER: Lips and Face Real Time Tracker. – In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR97)*, Porto Rico, USA, 1997.

[16] 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**: 116–127, 1994.

[17] K. Rijkse. – ITU standardization of very low bitrate video coding algorithms. – *Signal Processing : Image Communication*, **7**: 553–565, 1995.

[18] H.A. Rowley, S. Baluja and T. Kanade. – Neural network-based face detection. – In *Computer Vision and Pattern Recognition (CVPR96)*, pp. 203–208, San Francisco, USA, June 1996.

[19] S. Sista, C.A. Bouman and J.P. Allebach. – Fast image search using a multiscale stochastic model. – In *IEEE Int. Conf. on Image Processing (ICIP95)*, Vol.II, pp. 225–228, Washington D.C., USA, October 1995.

[20] M.U. Ramos Sanchez, J. Matas and J. Kittler. – Statistical Chromaticity Models for Lip Tracking with B-splines. – In *IAPR Int. Conf. on Audio- and Video-based Biometric Person Authentication (AVBPA97)*, pp. 69–76, Crans-Montana, Switzerland, March 1997.

[21] K. Sobottka and I. Pitas. – Face localization and facial feature extraction based on shape and color information. – In *IEEE Int. Conf. on Image Processing (ICIP96)*, Vol.III, pp. 483-486, Lausanne, Switzerland, September 1996.

[22] K. Sobottka and I. Pitas. – A Fully Automatic Approach to Facial Feature Detection and Tracking. – In *IAPR Int. Conf. on Audio- and Video-based Biometric Person Authentication (AVBPA97)*, pp. 77–84, Crans-Montana, Switzerland, March 1997.

[23] K. Sung and T. Poggio. – Example-based learning for view-based human face detection. – In *Technical Report AIM-1521*, MIT, 1994.

[24] M. Turk and A. Pentland. – Eigenfaces for recognition. – *J. of Cognitive Science*, **3**(1): 1–24, 1991.

[25] Y. Wang, R.-S. Wang, O. Lee, T. Chen, H. Chen and B.G. Haskell. – Mouth Shape Detection and Tracking Using an Active Mesh. – *SPIE, Visual communications and image processing*, **2501**: 1141–1152, 1995.

[26] 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.

[27] X. Zhuang, T. Wang and P. Zhang. – A highly robust estimator through partially likelihood function modeling and its application in computer vision. – *IEEE Trans. on Pattern Analysis and Machine Intelligence* **14**(1): 19–35, 1992.

List of figures

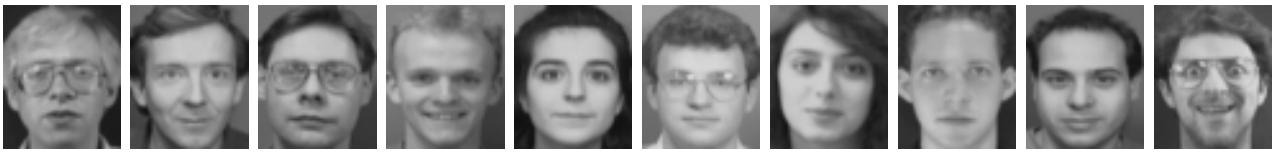


Figure 1:

$-3\,\lambda_1^{\mathrm{f}} \triangleleft$ ——————————— $\bar{\mathbf{x}}_{\mathrm{f}}$ ——————————— $\triangleright + 3\,\lambda_1^{\mathrm{f}}$

Figure 2:

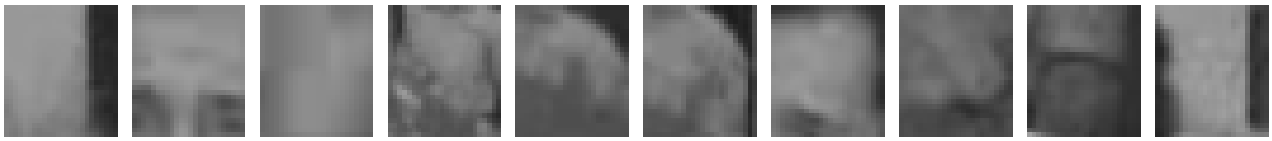
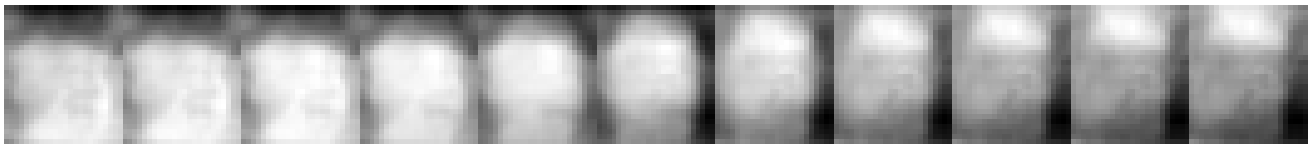


Figure 3:



$-3\,\lambda_1^{\mathrm{pf}} \triangleleft$ ——————————— $\bar{\mathbf{x}}_{\mathrm{pf}}$ ——————————— $\triangleright + 3\,\lambda_1^{\mathrm{pf}}$

Figure 4:

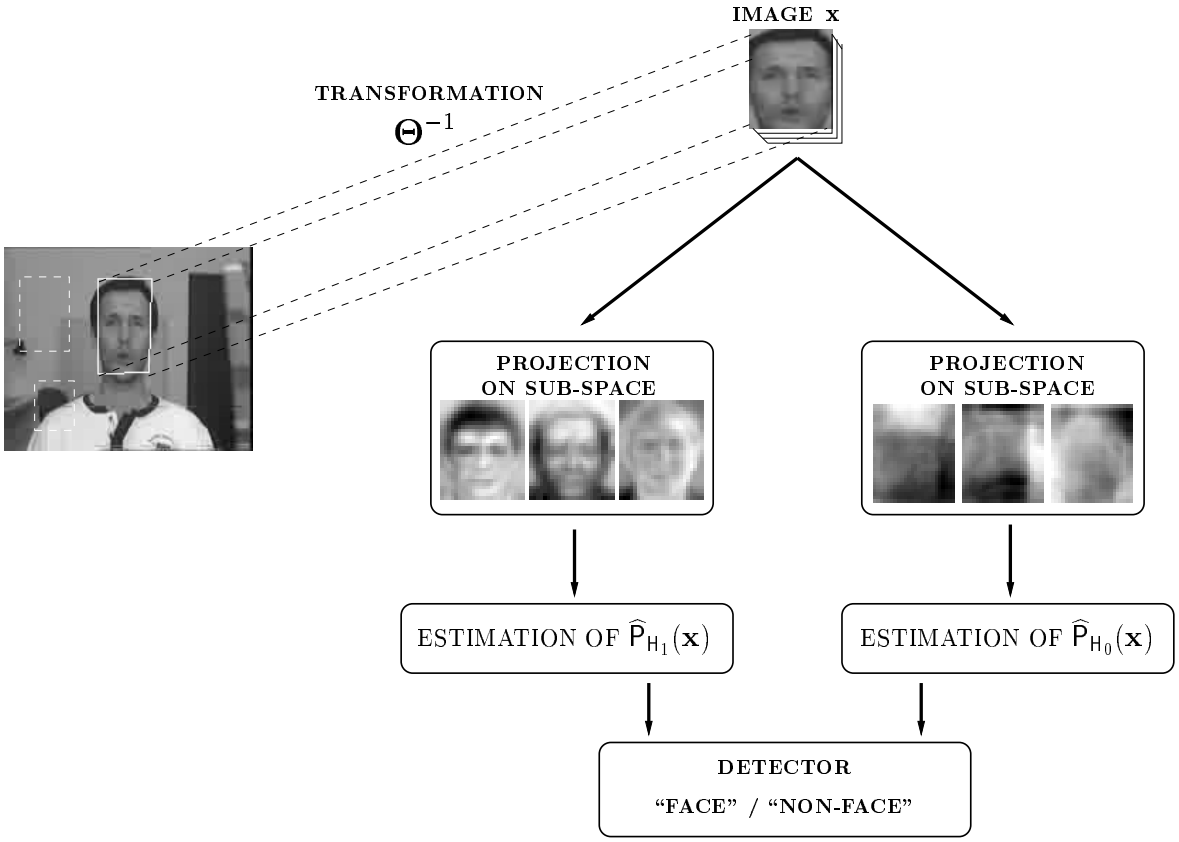


Figure 5:

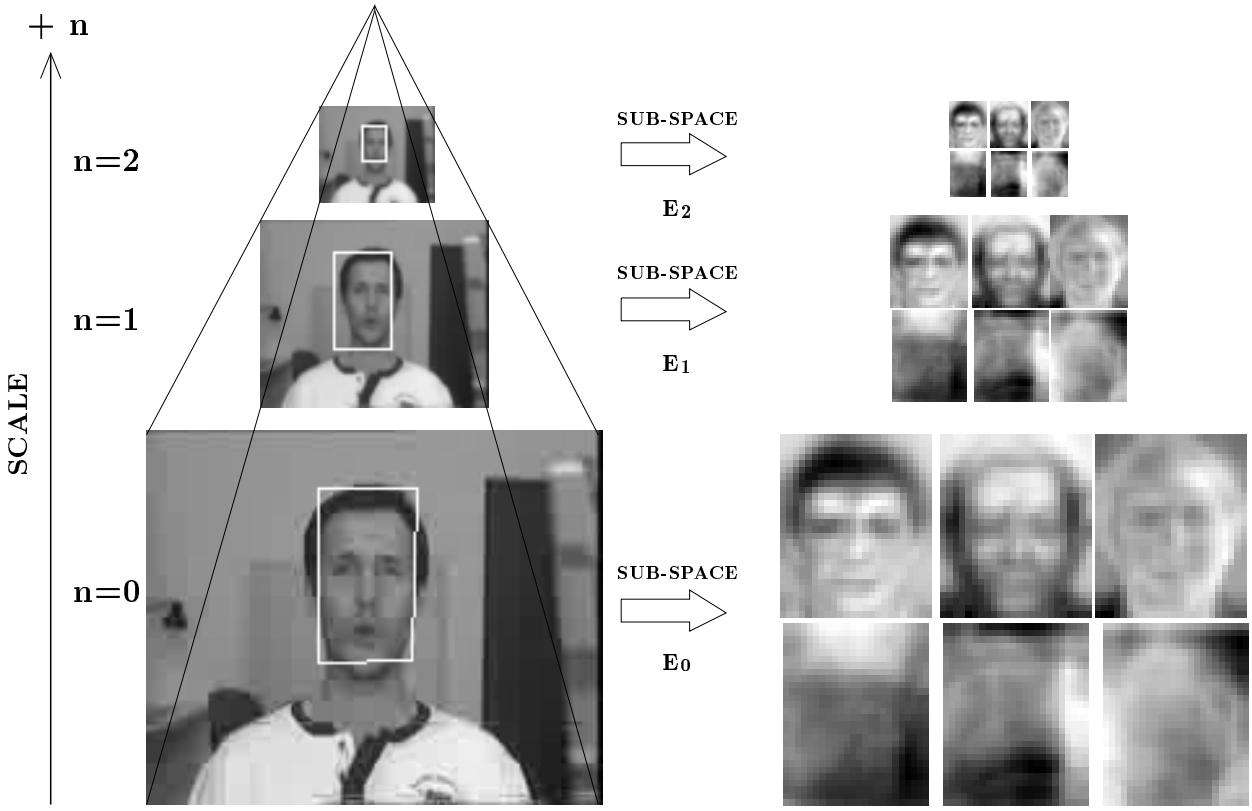


Figure 6:

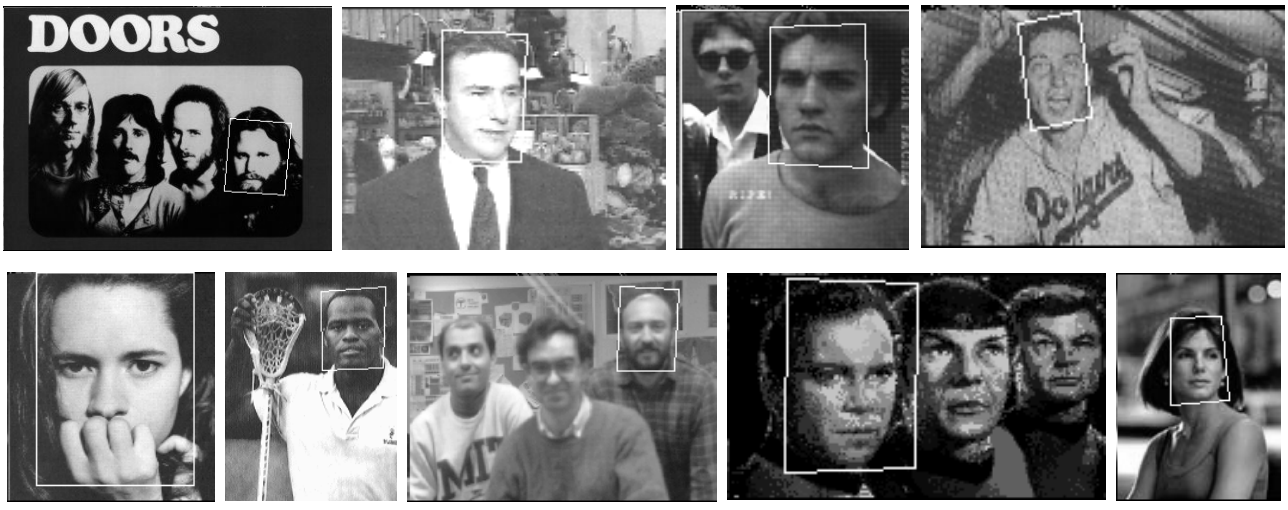


Figure 7:

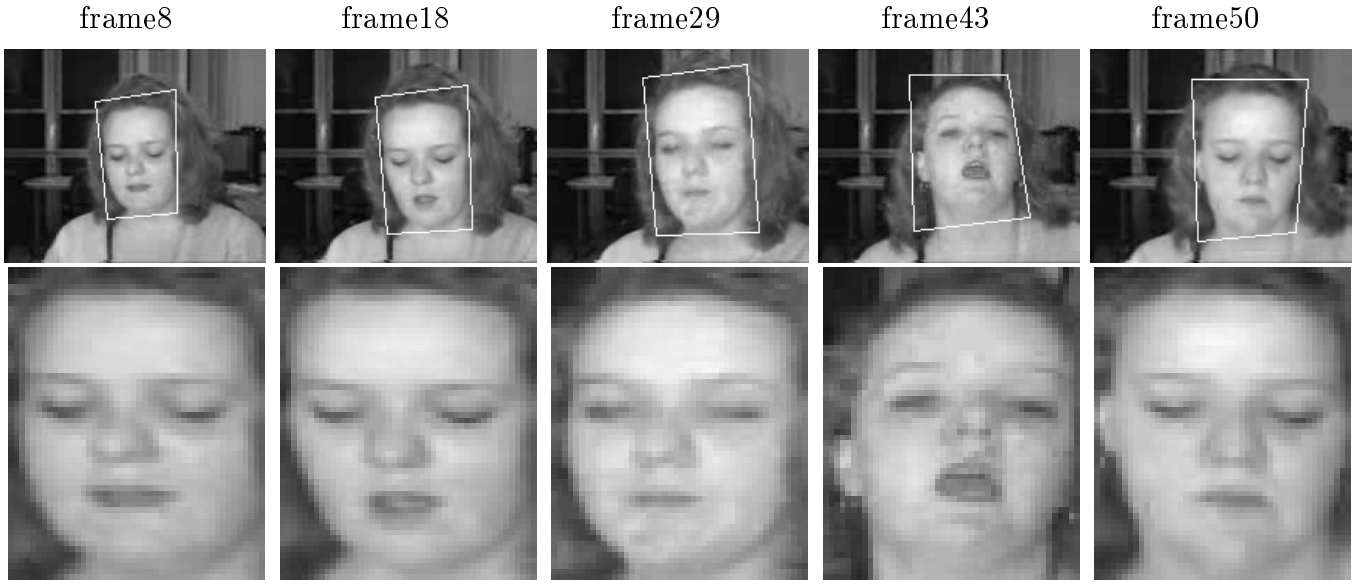


Figure 8:

Figure 9:

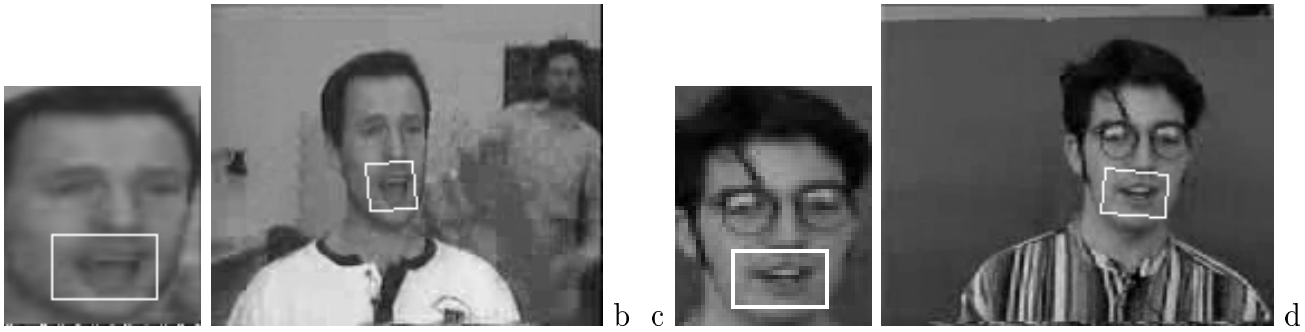


Figure 10:

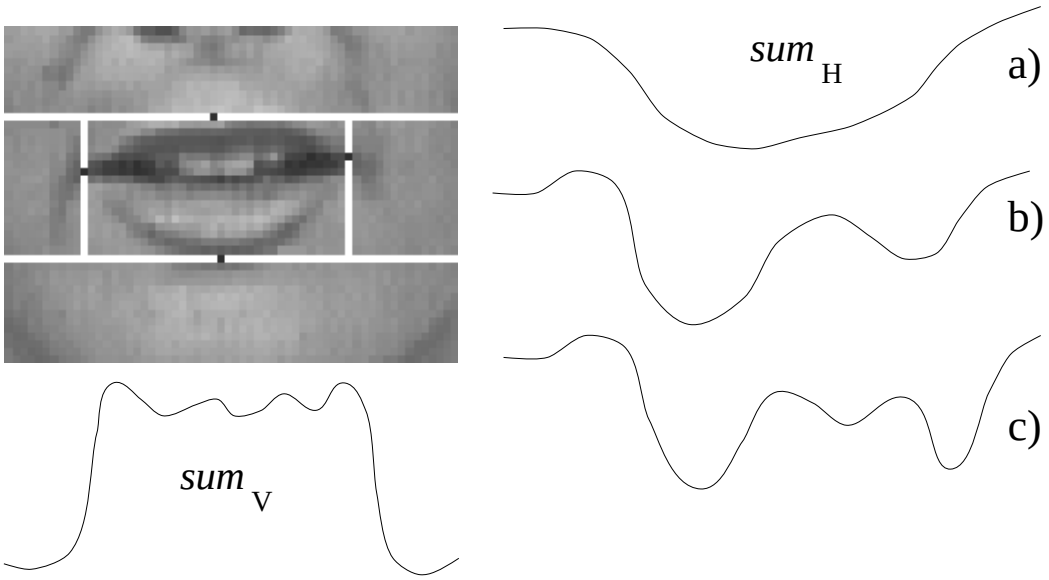


Figure 11:

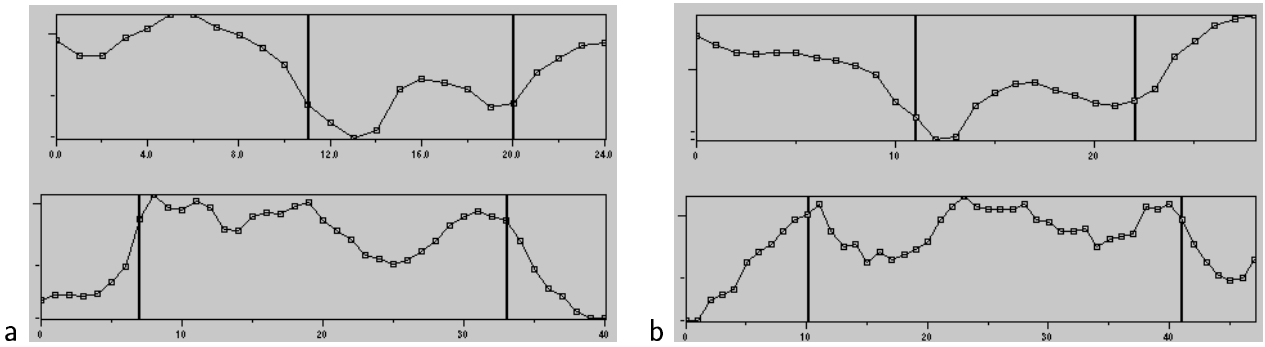


Figure 12:

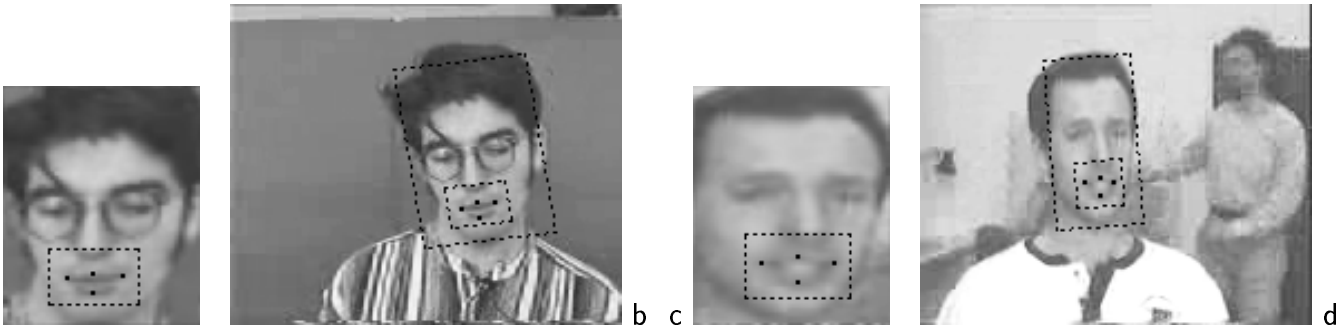


Figure 13:

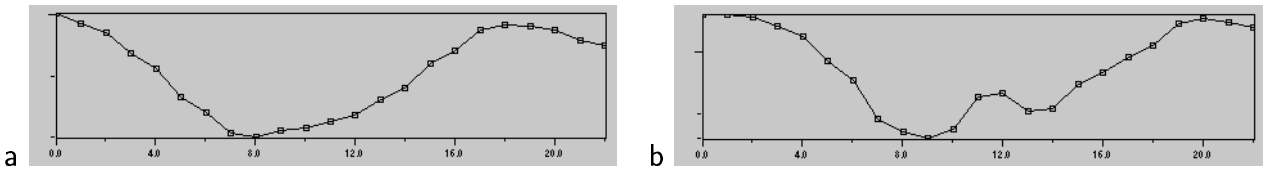


Figure 14:

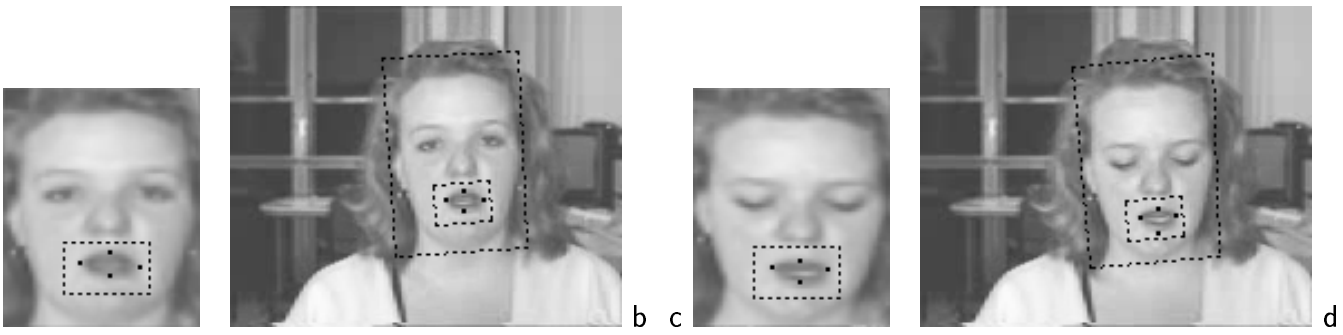


Figure 15: