Probabilistic multiple face detection and tracking using entropy measures

Evangelos Loutas Ioannis Pitas Christophoros Nikou

University of Thessaloniki
Department of Informatics
Box 451, Thessaloniki 54124, Greece
Tel.+30 2310 996 361, Fax.+30 2310 996 304
E-mail: pitas@zeus.csd.auth.gr

Address for correspondence :
Professor Ioannis Pitas
University of Thessaloniki
Department of Informatics
BOX 451, 54006 Thessaloniki, Greece
Tel. ++ 30 2310 99 63 04
Fax ++ 30 2310 99 63 04
Email: pitas@zeus.csd.auth.gr

Abstract

A joint probabilistic face detection and tracking algorithm combining likelihood estimation and a prior probability is proposed in this paper. The likelihood estimation scheme is based on the statistical training of sets of automatically generated feature points and a mutual information tracking cue, while the prior probability estimation is based on a Gaussian temporal model. The likelihood estimation process is the core of a multiple face detection scheme used to initialize the tracking process. The resulting system was tested on real image sequences and is robust to significant partial occlusion and illumination changes.

Index terms: Mutual information, arbitration scheme, occlusion, illumination changes.

I. Introduction

Automatic detection and tracking of human body parts (e.g. face, arms) is a challenging research topic with applications in many domains such as human computer interaction, surveillance, face recognition and human joint audio and video localization systems.

In this framework, Bayesian approaches express the posterior probability of the motion parameters in terms of a prior probability and a likelihood function [1]. The prior probability is representative of the previous history of the tracked object and the likelihood is representative of its similarity to an appearance based model learnt through statistical training. Bayesian approaches are considered as an effective way for updating prior information by estimating the posterior probability and using it as a prior in the next stage of the tracking process. They also allow the fusion of different tracking cues in order to provide a joint tracking output.

The main characteristics of the relevant published work are the use of an image model learnt through statistical training and the fusion of different tracking cues. An appearance model consisting of a stable component, a transient component and an outlier detection process is proposed in [2], while the use of exemplar based models in object tracking is introduced in [3]. Object tracking is performed using color, texture, and edge information in [4], while edge and ridge information is used in [5]. Grayscale information and motion models are combined in [6] to perform tracking of 3D articulated figures.

Head orientation is calculated by using either feature based methods [7],[8] or appearance based methods [9],[10]. The latter rely on using training sets of face images under various poses, while the former do not require statistical training. Appearance based methods are particularly interesting, as they can be combined in a probabilistic framework to obtain a single perceptual output.

The face tracking scheme proposed in this paper relies on calculating the posterior probability of motion parameters as the product of a prior probability and a likelihood function. The construction of the likelihood function relies on an appearance based model of automatically generated feature point sets and a mutual information based tracking cue, while the prior probability is constructed by using a temporal model term. Mutual information has been widely used in image registration [11]. It has also been used as a cue selection criterion in multiple cue tracking systems [12], [13]. The novelty of our approach lies in the use of mutual information as a separate cue in a probabilistic face tracking framework. Furthermore, the probability of face observation is constrained by using a temporal model based on the automatically generated feature point sets. Head orientation calculation is performed using a mutual information based scheme as well. The proposed approach does not require training for head orientation estimation and has shown good results in determining pose under facial appearance changes and illumination variations.

The tracking initialization algorithm uses a likelihood function estimation framework and can be interpreted as a probabilistic face detector. An arbitration scheme is also used to obtain an extension of the algorithm to cover multiple face cases.

The main contributions of the current work are the use of a novel probabilistic model based on automatically generated feature point sets in an object tracking scheme, the introduction of mutual information as a separate cue in a probabilistic face tracking framework and the head orientation calculation method using mutual information.

The proposed tracking scheme was tested on real image sequences. The tracker performs well in partial occlusion and illumination changes, because it combines the robustness of the mutual information systems to

illumination changes and the appearance based face detection systems to partial occlusion.

The remainder of the paper is organized as follows: The estimation of the likelihood function term based on statistical training is described in section 2. The estimation of the probability based on mutual information is presented in section 3. The temporal model is presented in section 4. The tracking process and the tracking initialization procedure are described in sections 5 and 6 respectively. Experimental results are presented in section 7 and conclusions are drawn in section 8.

II. Estimation of the likelihood function term based on statistical training

The likelihood function term based on statistical training is learnt through training using automatically generated feature point sets. Each image of the training set is reduced to a set of automatically generated feature points [14], [15]. The feature points represent image corners and are characterized by large gradient variations in both horizontal and vertical directions.

A. Face feature generation and training

The feature points $\mathbf{v} = [v_x, v_y]^T$ [14], are generated using a matrix:

$$\mathbf{Z} = \begin{bmatrix} \sum_W J_x^2 & \sum_W J_x J_y \\ \sum_W J_x J_y & \sum_W J_y^2 \end{bmatrix}, \tag{1}$$

where J_x and J_y are the image gradients of an image point in the x and y direction respectively and W is a $n \times n$ window centered on the candidate feature point. Matrix $\mathbf{Z}$ is zero-positive one by definition with two eigenvalues $\lambda_2 > \lambda_1 \geq 0$ and is calculated for every candidate feature point. The selected feature points have two large eigenvalues of their matrix $\mathbf{Z}$. Furthermore, the geometrical distance between two feature points must not be smaller than a predefined threshold (feature point neighborhood threshold) to ensure that the feature points do not concentrate on small image neighborhoods. The feature point set is assumed to be comprised of N pixels. Most of them represent contour corners or local intensity patterns not corresponding to obviously visible scene features [16]. In the case of faces, the feature points are expected to lie on facial areas containing intensity variations, such as facial contours, eyes, nose and mouth (see Figure 1).

The training procedure involves the feature point set generation from a number of training images. The ”ORL Database of Faces”[17] containing a total number of 400 images of 40 different persons was used for training. Feature point sets were generated on the facial region of each training image. The facial region bounding rectangle was manually defined during the training process. The feature point set generation process was performed inside the manually defined facial region for each image belonging to the training set.

As stated in [18], a major difficulty in the application of statistical feature point training is the efficient establishment of a rough registration between the different instances of the training set. Therefore, to avoid tedious manual interaction on very large feature point sets we have resorted to a semi-automatic registration procedure. A bounding box containing the face was drawn on each image of the training set and registered with respect to the bounding box of an arbitrarily chosen reference image.

Let $\mathbf{v}_{r_i}^1$ be the geometrical coordinates of the $i-th$ feature point with respect to the upper left corner of the face bounding box belonging to the first image of the training set and $\mathbf{v}_{r_j}^l$ the geometrical coordinates with respect to the upper left corner of the face bounding box of a feature point belonging to the $l-th$ image of the

training set and has not been matched yet. We have assumed correspondence for features $\mathbf{v}^1_{r_i}$ and $\mathbf{v}^l_{r_{j^*}}$, with j^* satisfying:

$$j^* = \arg\min_j \| \mathbf{v}^1_{r_i} - \mathbf{v}^l_{r_j} \| \tag{2}$$

among the feature points of image l not yet matched. Fully automatic unsupervised registration algorithms also proposed in literature may be applied. Their output strongly depends on the similarity metric used [19].

The feature point generation can be seen as a mapping procedure. Each image of the training set is mapped to a "feature point set" space with reduced dimensionality. The number of feature points N is selected to be much less than the total number of image pixels N_1, ($N << N_1$). It is convenient to set $N < N_T$, where N_T is the cardinality of the training image set. In our case $N_T = 400$.

A second dimensionality reduction step can be accomplished by using standard PCA. This step is necessary, if further dimensionality reduction is desired without reducing the number of feature points. The Gaussian probability density function of facial observation can be computed using the first M principal components, typically $M = 0.2N$. If N is chosen as $N \simeq 0.1N_1$, then $M \simeq 0.02N_1$. This corresponds to a significant data reduction of a factor of 50:1. The level of reduction can be controlled by appropriately selecting the number of feature points N.

Let $J(\mathbf{v})$ be the image intensity at pixel $\mathbf{v}$. The feature vector $\mathbf{x} = [J(\mathbf{v}_1), \ldots, J(\mathbf{v}_N)]^T$ can be expressed as:

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}\mathbf{b}, \tag{3}$$

where $\mathbf{P}$ is the matrix whose columns are the eigenvectors of the covariance matrix:

$$\mathbf{C} = \mathbf{E}[(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T], \tag{4}$$

$\overline{\mathbf{x}}$ is the mean feature vector:

$$\overline{\mathbf{x}} = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbf{x}_i \tag{5}$$

and

$$\mathbf{b}_i = \mathbf{P}^T(\mathbf{x} - \overline{\mathbf{x}}) \tag{6}$$

are the coordinates of $\mathbf{x} - \overline{\mathbf{x}}$ in the eigenvector basis. A principal component feature vector $\mathbf{y} = [y_1, \ldots, y_M]^T$ is obtained by:

$$\mathbf{y} = \mathbf{P}_M^T(\mathbf{x} - \overline{\mathbf{x}}), \tag{7}$$

where P_M is a submatrix of P containing the M principal eigenvectors.

B. Probability estimation

The estimation of the likelihood function term based on statistical training is accomplished by using the multiscale extension of the face detection procedure presented in [20]. Let $\overline{\mathbf{x}}$ be the mean and $\mathbf{C}$ be the covariance matrix of the feature vector $\mathbf{x}_t = [J_t(\mathbf{v}_1), \ldots, J_t(\mathbf{v}_N)]^T$ obtained by the statistical training procedure. The likelihood of an feature vector $\mathbf{x}_t$ at time instant t, under the assumption of Gaussian distribution is given by:

$$p(\mathbf{x}_t | \underline{\phi}_t, \Omega) = \frac{exp[-\frac{1}{2}(\mathbf{x}_t - \overline{\mathbf{x}})^T \mathbf{C}^{-1}(\mathbf{x}_t - \overline{\mathbf{x}})]}{(2\pi)^{\frac{N}{2}} |\mathbf{C}|^{\frac{1}{2}}} \tag{8}$$

Where Ω is the face class and $\underline{\phi}_t$ is the position, scale and rotation vector to be defined in the next section. If PCA has been performed on the feature vectors, $p(\mathbf{x}_t|\underline{\phi}_t, \Omega)$ can be approximated by [20]:

$$\widehat{p}(\mathbf{x}_t|\underline{\phi}_t, \Omega) = p_M(\mathbf{x}_t|\underline{\phi}_t, \Omega)\widehat{p}_{N-M}(\mathbf{x}_t|\underline{\phi}_t, \Omega) \tag{9}$$

where:

$$p_M(\mathbf{x}_t|\underline{\phi}_t, \Omega) = \frac{\exp(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i})}{(2\pi)^{\frac{M}{2}}\Pi_{i=1}^{M}\lambda_i^{\frac{1}{2}}} \tag{10}$$

is the term estimated from the M principal components and:

$$\widehat{p}_{N-M}(\mathbf{x}_t|\underline{\phi}_t, \Omega) = \frac{\exp(-\frac{\epsilon^2(\mathbf{x}_t)}{2\rho})}{(2\pi\rho)^{\frac{(N-M)}{2}}} \tag{11}$$

is the estimated contribution of the remaining components. $\epsilon^2(\mathbf{x}_t)$ is the residual reconstruction error:

$$\epsilon^2(\mathbf{x}_t) = \sum_{i=M+1}^{N} y_i^2 = \| \mathbf{x}_t - \overline{\mathbf{x}} \|^2 - \sum_{i=1}^{M} y_i^2 \tag{12}$$

and:

$$\rho = \frac{1}{N-M}\Sigma_{i=M+1}^{N}\lambda_i. \tag{13}$$

In order to estimate $\widehat{p}(\mathbf{x}_t|\underline{\phi}_t, \Omega)$ over a new face region in video frame t, a set of feature points should be generated using the previously described algorithm. The probability $\widehat{p}(\mathbf{x}_t|\underline{\phi}_t, \Omega)$ of a pattern $\mathbf{x}_t$ belonging to a face is generally normalized with respect to its maximum value. The normalized probability is compared to a predefined threshold in order to perform facial region assignment.

III. Estimation of the probability based on mutual information

The tracking process can be modeled as a communication between a transmitter (the reference face region at time $t-1$) and a receiver (the target face region at time t) with a symbol alphabet having cardinality N_{max} (the maximum number of grayscale levels). The mutual information is a measure of the amount of information transmitted through the communication channel. Let $U(\underline{\phi}_{t-1}), V(\underline{\phi}_t)$ be two random variables with $p(u,v), p(u), p(v)$ their joint and marginal probability mass functions and $\underline{\phi}_t = [\mathbf{V}_t, s_t, \vartheta_t]^T$ be the tracked face parameter vector at time t. $\mathbf{V}_t$ contains the feature point set geometrical coordinates at time t, $\mathbf{V}_t = [\mathbf{v}_{1_t}, \ldots, \mathbf{v}_{N_t}]^T$, while s and ϑ represent face scale and rotation parameters at time t with respect to the face location at time $t-1$.

The mutual information of two random variables U, V with a joint probability mass function $p(u,v)$ is defined as [21], [22]:

$$I(U(\underline{\phi}_{t-1}), V(\underline{\phi}_t)) = \sum_{i=1}^{N_{max}}\sum_{j=1}^{N_{max}} p(u_i, v_j)\log_2\frac{p(u_i, v_j)}{p(u_i)p(v_j)}, \tag{14}$$

The maximal mutual information for a particular prior $p(u)$ is [23]:

$$I_{max}(U(\underline{\phi}_{t-1}), V(\underline{\phi}_t)) = -\sum_{i=1}^{N_{max}} p(u_i)\log_2 p(u_i) \tag{15}$$

and reaches its maximal value when:

$$p(u_i) = \frac{1}{N_{max}}, \qquad 0 \leq i < N. \tag{16}$$

We define the prior probability based on the mutual information tracking cue as:

$$p_{MI}(U, V|\underline{\phi}_t, \underline{\phi}_{t-1}) = \frac{I(U(\underline{\phi}_{t-1}), V(\underline{\phi}_t))}{I_{max}(U(\underline{\phi}_{t-1}), V(\underline{\phi}_t))}, \tag{17}$$

Since $I(U, V) \geq 0$ [21], [22], $0 \leq p_{MI}(U, V|\underline{\phi}_t, \underline{\phi}_{t-1}) \leq 1$. A large value of $p_{MI}(U, V|\underline{\phi}_t, \underline{\phi}_{t-1})$ indicates a strong match between the reference and the target regions, while a small value of $p_{MI}(U, V|\underline{\phi}_t, \underline{\phi}_{t-1})$ indicates a weaker match. As can be seen, the calculation of p_{MI} does not require previous training and is calculated using the facial position in the previous frame (reference face region) and current frame (target face region).

IV. Temporal model

The temporal model describes the probability face appearance at a certain location given its location at the previous time instant. The temporal model is used as a constraint factor in the tracking process [6]. Scale variation s is modeled as a Gaussian distribution:

$$p(s_t|s_{t-1}) \sim c_1 e^{-(c_2(s_t - s_{t-1})^2)}. \tag{18}$$

In order to model the facial position variation, the feature point sets generated on the reference and target regions are used. The overall facial position variation is modeled as:

$$p(\mathbf{V}_t|\mathbf{V}_{t-1}) \sim c_3 e^{-(c_4 \sum_k (x_k(t) - x_k(t-1))^2 + (y_k(t) - y_k(t-1))^2)} \tag{19}$$

where $x_k(t), y_k(t)$ are the x and y coordinates of feature point k respectively at time instant t. Finally, rotation is modeled by:

$$p(\vartheta_t|\vartheta_{t-1}) \sim c_5 e^{-(c_6(\vartheta(t) - \vartheta(t-1))^2)} \tag{20}$$

The overall temporal model term is defined as the product of the terms:
$p(s_t|s_{t-1}), p(\mathbf{V}_t|\mathbf{V}_{t-1})$ and $p(\vartheta_t|\vartheta_{t-1})$.

$$p_{TEMP}(\underline{\phi}_t|\underline{\phi}_{t-1}) = p(s_t|s_{t-1}) p(\mathbf{V}_t|\mathbf{V}_{t-1}) p(\vartheta_t|\vartheta_{t-1}) \tag{21}$$

Prior probabilities are not informative if the prior pdf has a larger variance than the likelihood function [1]. Therefore, too small values of c_2, c_4 and c_6 will render the temporal model non informative and, thus, not useful for the tracking process.

V. Face Tracking

In order to track the detected faces to the next frame, the observation probabilities $p(\underline{\phi}_t|\mathbf{x}_t, \underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega)$ are calculated for each detected face [6]. Let us recall that $\underline{\phi}_t = [\mathbf{V}_t, s_t, \vartheta_t]^T$ is the vector containing the feature points and their rotation and scaling parameters at time instant t.

Using the Bayesian formulation it can be easily found that:

$$p(\underline{\phi}_t|\mathbf{x}_t, \underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega) = \frac{p(\underline{\phi}_t|\underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega) p(\mathbf{x}_t|\underline{\phi}_t, \underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega)}{p(\mathbf{x}_t|\underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega)} \tag{22}$$

In (22) $p(\underline{\phi}_t|\mathbf{x}_t, \underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega)$ represents the posterior density, $p(\underline{\phi}_t|\underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega)$ represents the prior density function, $p(\mathbf{x}_t|\underline{\phi}_t, \underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega)$ represents the likelihood function. The term $p(\mathbf{x}_t|\underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega)$ is considered a constant [1]. In the context of present work, the likelihood function is calculated as a the product of a term based on statistical training and a mutual information tracking term, while the prior density function is calculated by using a temporal model term. The mutual information term is used as a measure of the similarity between the face in the previous and in current frame. The observation probability $p(\underline{\phi}_t|\mathbf{x}_t, \underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega)$ of the parameter vector $\underline{\phi}_t$ is approximated, as the product of a prior probability term and a likelihood term [1] by:

$$p(\underline{\phi}_t|\mathbf{x}_t, \underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega) = c_7 p(\underline{\phi}_t|\underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega) p(\mathbf{x}_t|\underline{\phi}_t, \underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega). \tag{23}$$

The term c_7 is a normalizing factor [1]. In the context of the present work:

$$p(\mathbf{x}_t|\underline{\phi}_t, \underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega) \approx p_{MI}(U, V|\underline{\phi}_t, \underline{\phi}_{t-1}) p(\mathbf{x}_t|\underline{\phi}_t, \Omega), \tag{24}$$

$$p(\underline{\phi}_t|\underline{\phi}_{t-1}, \mathbf{x}_{t-1}, \Omega) \approx p_{TEMP}(\underline{\phi}_t|\underline{\phi}_{t-1}) \tag{25}$$

In order to obtain an estimate of the head location parameters $\underline{\phi}_t$ we start by finding the vector $\mathbf{V}_t$. The estimate of $\underline{\phi}_t$ is then refined by calculating the scale factor s_t and the rotation angle ϑ_t. Better results may be obtained by adopting a recursive refining process.

In the context of present work, the scale estimate is determined by using the procedure described in section II-B. Head orientation is estimated by using the mutual information cue presented in section III. A search for the best rotation angle is performed after the translation parameters are estimated. The probability $p_{MI}(U, V|\underline{\phi}_t, \underline{\phi}_{t-1})$ (17) incorporates the orientation information. The obtained estimates are refined by using the temporal model presented in section IV.

VI. Face tracking initialization

The face tracking algorithm initialization procedure is based on the estimation of the probability $\widehat{p}(\mathbf{x}_t|\underline{\phi}_t, \Omega)$. The probability is obtained by using the process described in section II-B and is extended to handle multiple faces. Candidate facial regions are considered all those for which the normalized probability $\widehat{p}(\mathbf{x}_t|\underline{\phi}_t, \Omega)$ with respect to its maximum value exceeds a predefined threshold, whose value is set empirically by obtaining tracking results on image sequences acquired using the same camera under similar illumination conditions. In order to eliminate false facial region candidates an arbitration scheme similar to that presented in [24] is implemented and is described subsequently. The initialization steps of the multiple face tracking algorithm are:

- Calculate the probabilities $\widehat{p}(\mathbf{x}_t|\underline{\phi}_t, \Omega)$ over the entire image (9).
- Reject all the candidate regions whose normalized probability is below a predefined threshold. Mark these candidate regions as non facial regions.
- **Repeat**
  - Mark as a face the unmarked image region assigned to the maximum probability.
  - Perform the *arbitration scheme*:
    - Reject any candidate facial region whose center lies within a previously defined facial region.
    - Reject any candidate facial region overlapping with a previously defined facial region.

    - Reject any candidate facial region, when the number of less probable candidate facial regions within them is less than a predefined threshold.

- **until** all candidate regions are marked as facial or non facial ones.

The first two rules of the arbitration scheme impose the rejection of candidate facial regions that are considered outliers, based on previous detection results. Nevertheless, wrong rejection of facial regions lying very closely to each other is a possible side effect. The third rule of the arbitration scheme was motivated by the work of Rowley et al. [24] and is based on the assumption that a strong facial candidate should be accompanied by neighboring less strong facial candidates. The absence of the less strong facial candidates implies that the facial candidate under examination is an outlier and should be rejected.

VII. Experimental Results

The proposed algorithm was tested on a variety of real face image sequences under various illumination and occlusion conditions. The image sequences were obtained, using a simple video-conference camera. They can contain multiple faces per video shot.

The kernel parameters versus time, such as the likelihood function term based on statistical training (9), the mutual information based probability p_{MI}, (17) and the observation probability (23) are presented in time sequences (Figures 4, 5 and 6), for each face in the image sequences presented in Figures 2 and 3. The results for all the kernel parameters are calculated from the second frame of each testing sequence onwards. Note that in Figure 6, the final observation probability and the likelihood term based on statistical training are normalized with respect to the maximum probability after the first face localization.

Results on a single face sequence without illumination changes or partial occlusion are presented in Figure 7. As can be observed, the face position and orientation are correctly determined. Tracking results on a similar sequence with illumination changes are presented in Figure 8. A slight drift in the estimated facial position is noticed in very dark image sequences, when the tracking process is prolonged for too long. Results on multiple face image sequences suffering from lightening changes and partial occlusion are presented in Figures 9 and 10 respectively. Facial position is correctly determined in the multiple face case, even under severe partial occlusion or illumination changes. In general, the face tracking algorithm proposed in this paper can effectively track multiple faces under significant illumination changes and partial occlusion.

VIII. Conclusions

A probabilistic face tracking scheme was presented in this paper. Likelihood function estimation is performed using sets of automatically generated feature points and a mutual information tracking cue, while the prior density function estimation is based on a Gaussian temporal model.

The main contributions of the proposed scheme are the introduction of a novel appearance based model for likelihood function estimation and the use of a mutual information tracking cue in conjunction with a Gaussian temporal model. Moreover, the implementation of an arbitration scheme to face tracking initialization is also important, since it allows a multiple face tracking extension.

The proposed algorithm was tested on real face sequences acquired using a video-conference camera under different illumination and occlusion conditions. Results have shown that the facial position is correctly determined

even in image sequences presenting important illumination changes and partial occlusion. The face orientation was correctly determined under normal illumination conditions and slight illumination changes. Robustness to illumination changes is obtained by using the mutual information tracking cue, while robustness to partial occlusion is obtained by the use of the appearance based model.

IX. Acknowledgements

This study has been partially supported by the Commission of the European Communities, in the framework of the project IST-1999 20993 CARROUSO (Creating, Assessing and Rendering of High Quality Audio-Visual Environments in MPEG-4 context).

References

[1] J. Ruanaidh and W. Fitzgerald, *Numerical bayesian methods applied to signal processing*, Springer-Verlag, 1996.

[2] A. Jepson, D. Fleet, and T. Maraghi, "Robust online appearance models for visual tracking," in *Proc. of 2001 Int. Conf. on Computer Vision and Pattern Recognition*, 2001, vol. I, pp. 415–422.

[3] K. Toyama and A. Blake, "Probabilistic tracking in a metric space," in *Proceedings of the International Conference on Computer Vision*, 2001, pp. 50–57.

[4] C. Rasmussen and G. D. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 560–576, 2001.

[5] H. Sidenbladh and M. Black, "Learning image statistics for bayesian tracking," in *IEEE International Conference on Computer Vision (ICCV), Vancouver, Canada.*, 2001, vol. 2, pp. 709–716.

[6] H. Sidenbladh, F. De la Torre, and M. Black, "A framework for modeling the appearance of 3d articulated figures," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG), Grenoble, France.*, 2000, pp. 368–375.

[7] R. Lopez, A. Colmenarez, and T. Huang, "Head and feature tracking for model-based video coding," in *International workshop on Synthetic-Natural Hybrid coding and 3-D imaging, Rhodes Greece 1997*, 1997.

[8] A. Nikolaidis and I. Pitas, "Facial feature extraction and pose determination," *Pattern Recognition, Elsevier*, vol. 33, no. 11, pp. 1783–1791, 2000.

[9] T. Darrell, B. Moghaddam, and A. Pentland, "Active face tracking and pose estimation in an interactive room," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1996, pp. 67–72.

[10] Y. Wu and K. Toyama, "Wide-range, person and illumination-insensitive head orientation estimation," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG), Grenoble, France.*, 2000, pp. 183–188.

[11] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[12] H. Kruppa and B. Schiele, "Context-driven model switching for visual tracking," in *9th International Symposium on Intelligent Robotic Systems, Toulouse, France.*, 2001.

[13] H. Kruppa and B. Schiele, "Using mutual information to combine object models," in *8th International Symposium on Intelligent Robotic Systems 2000, Reading, UK.*, 2000.

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

[15] K. Rohr, *Landmark-based image analysis*, Kluwer Academic Publishers, 2001.

[16] A. Verri E. Trucco, *Introductory techniques for 3-D Computer Vision*, Prentice Hall, 1998.

[17] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota FL*, 1994, pp. 138–142.

[18] T. F. Cootes and C. J. Taylor, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 1, no. 1, pp. 38–59, 1995.

[19] F. Heitz C. Nikou and J. P. Armspach, "Robust registration of dissimilar single and multimodal images.," in *Proceedings of the 5th European Conference on Computer Vision (ECCV'98)*, 1998, vol. 2, pp. 51–65.

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

[21] S. Haykin, *Communication Systems-3rd ed.*, J. Wiley, 1994.

[22] F. M. Reza, *An introduction to information theory*, Dover, 1994.

[23] M. Skouson, Q. Guo, and Z. Liang, "A bound on mutual information for image registration," *IEEE Transactions on Medical Imaging*, vol. 20, no. 8, pp. 843–846, 2001.

[24] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–37, 1998.

(a) (b) (c)

Fig. 1. (a) Feature point set of 100 feature points. Feature neighborhood threshold=5 pixels. (b) Feature point set of 100 feature points. Feature neighborhood threshold=3 pixels. (c) Feature point set of 300 feature points. Feature neighborhood threshold=3 pixels.

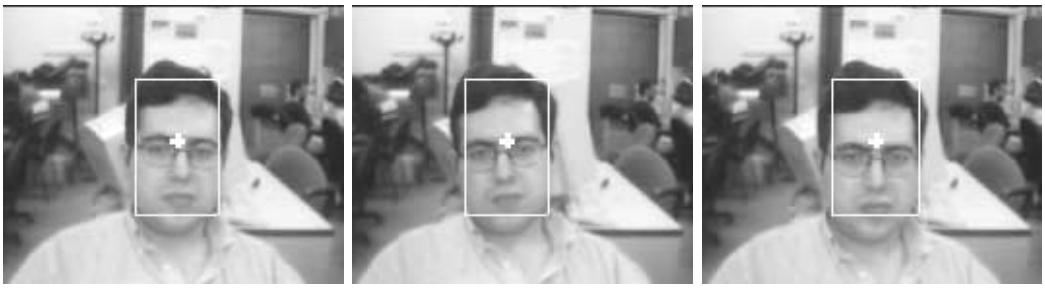

Fig. 2. Single face tracking image sequence.

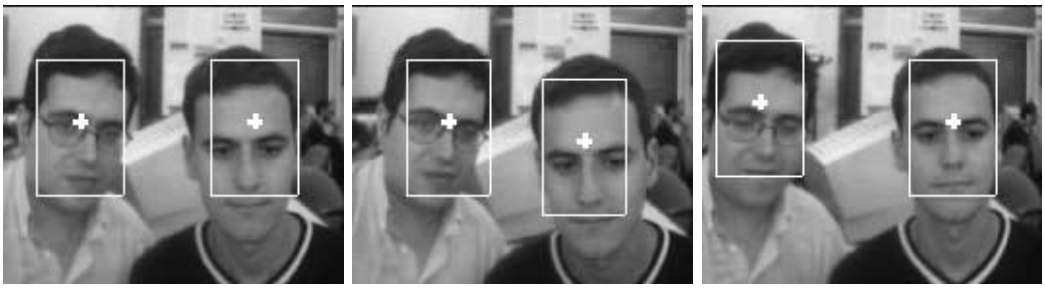

Fig. 3. Face tracking image sequence containing two faces.

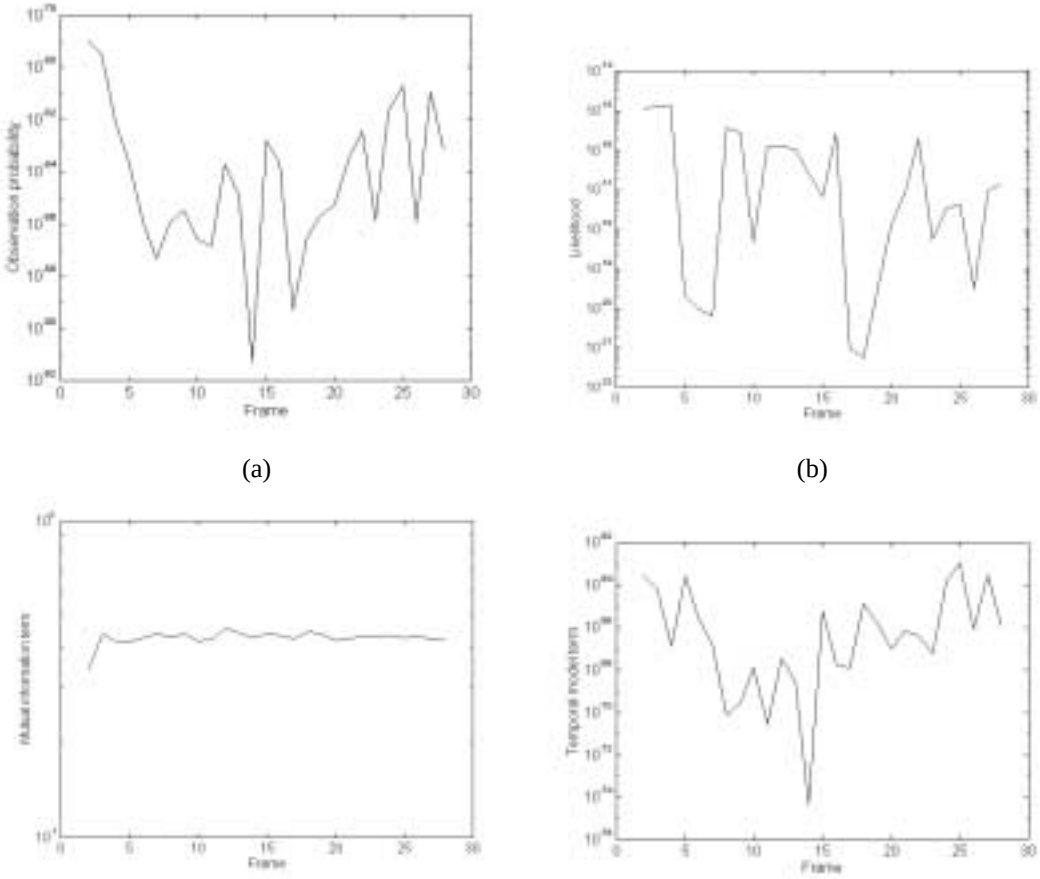

Fig. 4. (a) Observation probability (23), (b) likelihood function term based on statistical training (9) , (c) mutual information term (17), (d) temporal model term (18-20) versus time for the image sequence shown in Figure 2.

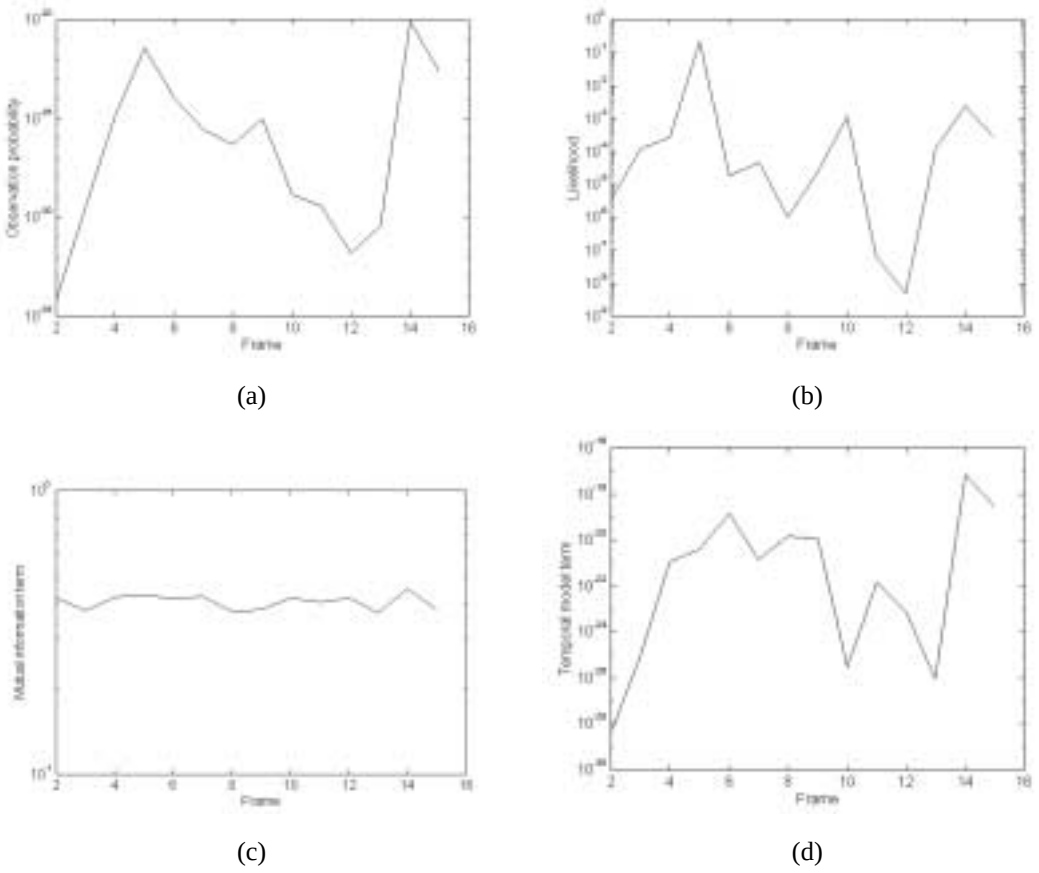

Fig. 5. (a) Observation probability (23), (b) likelihood function term based on statistical training (9) , (c) mutual information term (17), (d) temporal model term (18-20) versus time for the left-hand face of the image sequence presented in Figure 3.

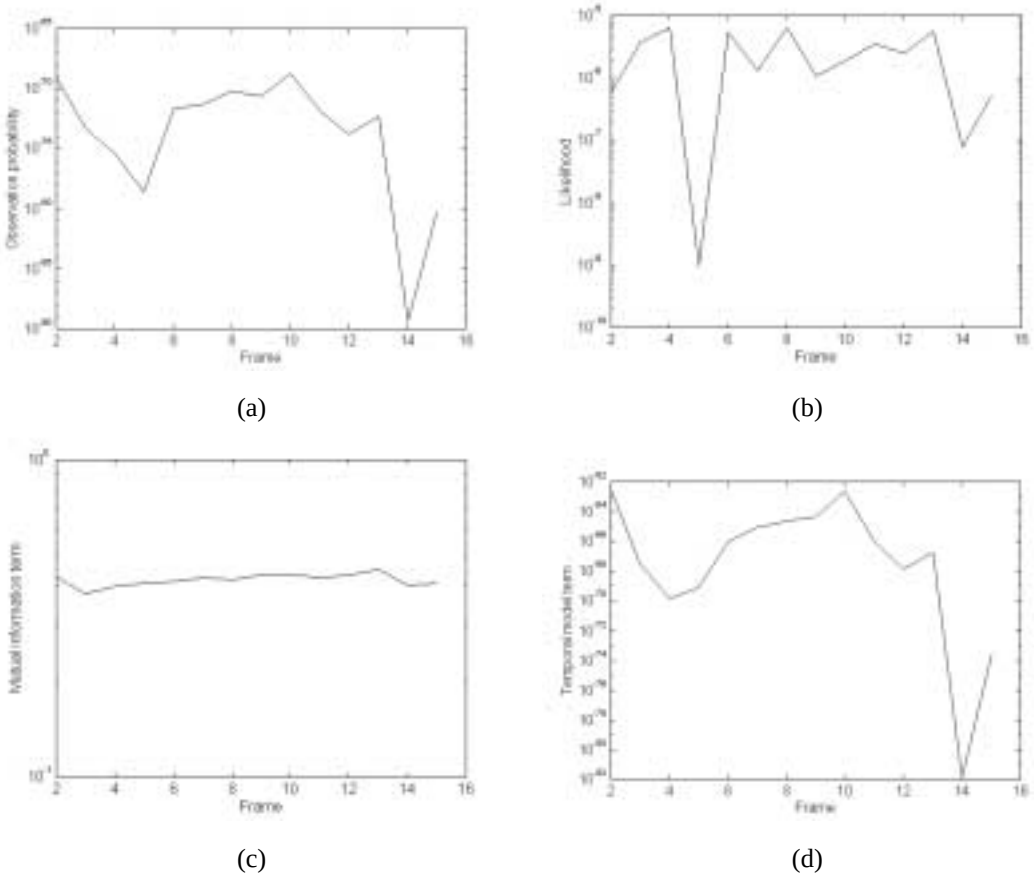

Fig. 6. (a) Observation probability (23), (b) likelihood function term based on statistical training (9) , (c) mutual information term (17), (d) temporal model term (18-20) versus time for the right-hand face of the image sequence presented in Figure 3.

Fig. 7. Face tracking under constant illumination conditions.

Fig. 8. Face tracking under varying illumination conditions.

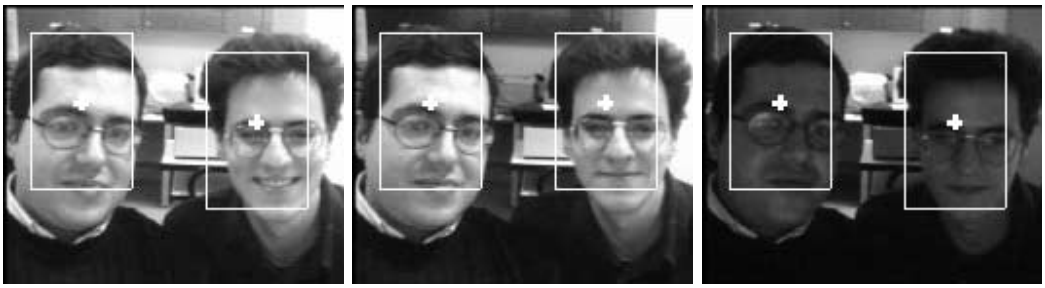

Fig. 9. Tracking of two faces under varying illumination conditions.

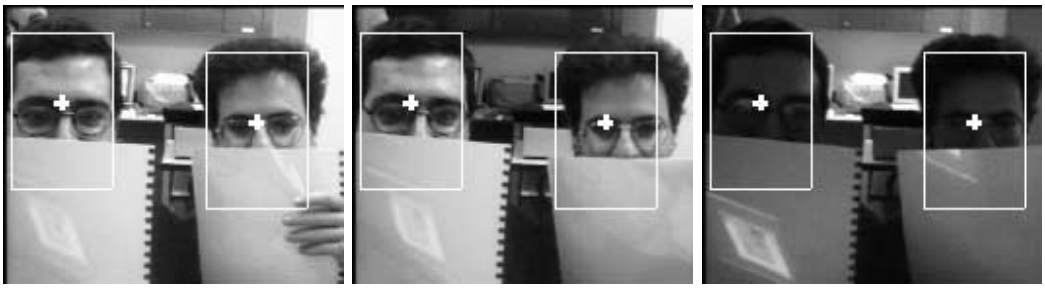

Fig. 10. Tracking results of two faces under varying illumination conditions and partial occlusion.