AN INFORMATION THEORETIC APPROACH TO JOINT PROBABILISTIC FACE DETECTION AND TRACKING

E. Loutas *C. Nikou* *I. Pitas*

Department of Informatics
University of Thessaloniki
Box 451, Thessaloniki 540 06
GREECE
E-mail:{eloutas,nikou,pitas}@zeus.csd.auth.gr

ABSTRACT

A joint probabilistic face detection and tracking algorithm for combining a likelihood estimation and a prior probability is proposed in this paper. Face tracking is achieved by a Bayesian framework. The likelihood estimation scheme is based on statistical training of sets of automatically generated feature points, while the prior probability estimation is based on the fusion of an information theoretic tracking cue and 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

1. INTRODUCTION

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

In that 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 tracked object previous history and the likelihood is representative of the similarity to an appearance based model learnt through statistical training. Bayesian approaches are considered an effective way of updating prior information by forwarding the posterior probability and using it as the prior in the next stage of the process. They also allow the fusion of different tracking cues in order to provide a joint tracking output.

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

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).

Head orientation is calculated by using either feature based methods [6, 7] or appearance based methods [8, 9]. The latter rely on using training sets of face images under varying pose, while the feature based methods do not require statistical training. Appearance based methods are particularly interesting as they can be combined in a probabilistic framework to obtain a single perceptory output.

The Bayesian face tracking scheme proposed in this paper relies on an appearance based model of automatically generated feature point sets for construction the likelihood function [10] and a mutual information tracking cue for constructing the prior probability. Our approach introduces the use of mutual information as a separate cue in a Bayesian face tracking framework. Also, the probability of face observation is constrained using a temporal model based on the automatically generated feature point sets. Head orientation calculation is performed using a mutual information based scheme. The proposed approach doesn't require training for head orientation estimation and has shown good results in determining pose under facial appearance changes and illumination variations.

The tracking algorithm is initialized using a likelihood function estimation framework and is interpreted as a probabilistic face detector. An arbitration scheme is also used to obtain a multiple face extension of the algorithm.

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 Bayesian 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 change situations as it combines the robustness of mutual information systems to illumination changes and the appearance based face detection systems to partial occlusion.

2. LIKELIHOOD ESTIMATION

The acquisition of the likelihood estimates is an important part of a Bayesian tracking framework. Moreover, it can be used in order to construct a face detection scheme. The face detection scheme is used as a tracking initialization procedure and is applied at the beginning of the tracking process or in the case of tracking failure.

Likelihood is learnt through training of automatically generated feature points. Each image of the training set is described by a set of automatically generated feature points [10, 11]. The feature points represent image corners and are characterized by large gradient variations in both horizontal and vertical directions and is presented in [12] as an edge detection algorithm.

2.1. Face feature generation and training

The feature set [10], is generated using a matrix:

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

Matrix $\mathbf{Z}$ is constructed for every candidate feature point. J_x and J_y are the image gradients of an image point in the x and y direction respectively and W is a nxn window centered on the candidate feature point. Matrix $\mathbf{Z}$ is zero-positive with two eigenvalues $\lambda_2 > \lambda_1 \geq 0$

Features having two large eigenvalues of their matrix $\mathbf{Z}$ are selected and the inter feature distance must not exceed a predefined threshold (feature neighborhood threshold).

The feature set is assumed to be comprised of N features. Most of them represent corners generated by the intersection of the object contours or corner of the local intensity pattern not corresponding to obvious scene features [12]. In the case of faces, the feature set is expected to lie on face areas containing intensity variations such as the face contour, the eyes area, the nose area and the mouth area (see figure 1).

The training procedure involves the feature set generation from a number of training images. The "ORL Database of Faces"[13] containing a total number of 400 images of 40 different persons was used for training. The number of features, 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 size of the training image set. In our case $N_T = 400$.

2.2. Face observation probability estimation

The estimation of the first cue face observation probability is accomplished by calculating the likelihood $P(\mathbf{x}|\Omega)$ of a target, where $\mathbf{x}$ is the input pattern in the "feature point set space" and Ω represents the face class. The multiscale extension of the face detection procedure used in [14] is adopted. Using the results obtained by PCA, $P(\mathbf{x}|\Omega)$ can be approximated with [14]:

$$\widehat{P}(\mathbf{x}|\Omega) = P_M(\mathbf{x}|\Omega)\widehat{P}_{N-M}(\mathbf{x}|\Omega) \quad (2)$$

where $P_M(\mathbf{x}|\Omega)$ is the term estimated from the M principal components and $\widehat{P}_{N-M}(\mathbf{x}|\Omega)$ is the estimated contribution of the remaining components.

In order to estimate $\widehat{P}(\mathbf{x}|\Omega)$ over a new image region a set of feature points should be generated using the previously described algorithm. An estimate of the face position and scale is thus obtained. The probability $\widehat{P}(\mathbf{x}|\Omega)$ of a pattern $\mathbf{x}$ belonging to a face is generally normalized with respect to its maximum value $\widehat{P}(\mathbf{x}|\Omega)_{max}$. The normalized probability is compared to a predefined threshold in order to perform facial region assignment.

2.3. Tracking algorithm initialization

The face tracking algorithm initialization procedure is based on the estimation of the facial observation probability. The facial observation probability calculation process is extended to handle multiple faces. Candidate facial regions are considered all those for which the normalized face observation probability exceeds a predefined threshold. In order to eliminate false facial region candidates an arbitration scheme similar to that presented in [15] is implemented. The steps of the initialization of the multiple face tracking algorithm are:

- Calculate the facial observation probabilities over the whole the image (eq. 2).
- Reject all the candidate regions whose normalized facial observation probability is below a predefined threshold. Mark these candidate regions as non face regions.
- **Repeat**
  - Mark as a face the unmarked image region assigned to the maximum facial observation 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 face or non face.

3. PRIOR PROBABILITY ESTIMATION

The prior probability is representative of the previous knowledge acquired through the tracking process. The estimation of the prior is based on a mutual information tracking cue and a temporal model.

3.1. Mutual information cue

The tracking process can be modeled as a communication between a transmitter (the reference face region A_1) and a receiver (the target face region A_2) with a N_{max} symbol alphabet (the maximum number of grayscale levels). Mutual information is a measure of the amount of information transmitted through the communication channel. Let $U((\phi_{t-1}), V(\phi_t))$ be two random variables with $p(u), p(v)$ their marginal probability mass functions and $u_i = J_1(\mathbf{p}_{t-1})$, $v_j = J_2(\mathbf{p}_t)$ their possible outcomes, where J_1 and J_2 are the reference and target images respectively, $\mathbf{p}_{t-1} \in A_1$, $\mathbf{p}_t \in A_2$ and $\phi_t = [\mathbf{x}_t, s_t, \vartheta_t]^T$ is the tracked face parameter vector at time t. $\mathbf{x}_t$ is the feature point set at time t, while s and ϑ represent scale and rotation respectively.

The mutual information of two random variables U, V with a joint probability mass function $p(u, v)$ is defined as [16]:

$$I(U(\phi_{t-1}), V(\phi_t)) = \sum_{i=1}^{N_{max}} \sum_{j=1}^{N_{max}} p(u_i, v_j) \log_2 \frac{p(u_i, v_j)}{p(u_i)p(v_j)}, \quad (3)$$

The maximum mutual information for a particular prior $p(u)$ is [17]:

$$I_{max}(U(\phi_{t-1}), V(\phi_t)) = -\sum_{i=1}^{N_{max}} p(u_i) \log_2 p(u_i) \quad (4)$$

and reaches its maximum value when

$$p(u_i) = \frac{1}{\log_2 N_{max}}, \qquad 0 \leq i < N. \quad (5)$$

Let the prior probability based on the mutual information tracking cue be:

$$p_{MI}(\phi_t|\phi_{t-1}) = c_1 \frac{I(U(\phi_{t-1}), V(\phi_t))}{I_{max}(U(\phi_{t-1}), V(\phi_t))}, \quad (6)$$

where c_1 is a constant.

Since $I(U,V) \geq 0$ [16], $0 \leq p_{MI}(\phi_t|\phi_{t-1}) \leq 1$. A large value of $p_{MI}(\phi_t|\phi_{t-1})$ indicates a strong match between the reference and the target regions, while a small value of $p_{MI}(\phi_t|\phi_{t-1})$ indicates a weaker match.

3.2. Temporal model

The temporal model part of the prior describes the probability of a face to appear given its location at the previous time instant. The temporal model is used as a constraint factor [5] in the tracking process. Scale variation s is modeled as a gaussian distribution:

$$p(s_t|s_{t-1}) \sim e^{-(c_2(s_t - s_{t-1})^2)}. \quad (7)$$

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 also modeled as a gaussian distribution:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) \sim e^{-(c_3 \sum_k (x_k(t) - x_k(t-1))^2 + (y_k(t) - y_k(t-1))^2)}, \quad (8)$$

where $x_k(t), y_k(t)$ are the x and y coordinates of feature point k at time instant t respectively.

Finally, rotation is modeled by:

$$p(\vartheta_t|\vartheta_{t-1}) \sim e^{-(c_4(\vartheta(t) - \vartheta(t-1))^2)} \quad (9)$$

Constants c_2, c_3 and c_4 are empirically determined. 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_3 and c_4 will render the temporal model non informative and thus unimportant to the tracking process.

4. FACE TRACKING

In order to track the detected faces to the next frame the observation probabilities $p_i(\phi_t|K_t)$ are calculated for each detected face [5]. Let us recall that $\phi_t = [\mathbf{x}, s, \vartheta]^T$ is the vector containing the feature points and their rotation and scaling parameters at time instant t and K_t is representing a subimage of image J_t. The observation probability $p(\phi_t|K_t)$ of the parameter vector is expressed by:

$$p(\phi_t|K_t) = C_d p(K_t|\phi_t) \widehat{P}(\mathbf{x}|\Omega). \quad (10)$$

The term C_d is a normalizing factor [1], while the term $p(K_t|\phi_t)$ represents the prior:

$$p(K_t|\phi_t) = p_{MI}(\phi_t|\phi_{t-1}) p(\phi_t|\phi_{t-1}) \quad (11)$$

As it can be observed, the prior is constructed from the mutual information contribution $p_{MI}(\phi_t|\phi_{t-1})$ and the temporal model contribution $p(\phi_t|\phi_{t-1})$.

In order to obtain the full estimate of the head orientation ϑ a coarse estimate is obtained at first by finding the translation vector. The estimate is then refined by calculating the scale factor and the rotation angle and the final estimate is obtained. Better results may be obtained by adopting a recursive refining process.

5. EXPERIMENTAL RESULTS

The proposed algorithm was tested on a variety of real face image sequences under different lightening and occlusion conditions. Results on a single face sequence without lightening changes or partial occlusion are presented in Figure 2. As it can be observed, the face position and orientation are correctly determined. Tracking results on a similar sequence with lightening changes are presented in Figure 3. 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 4 and 5 respectively. Facial position is correctly determined in the multiple face case even under severe partial occlusion and illumination changes. In general, the face tracking algorithm proposed in this paper can effectively track multiple faces under significant illumination changes and partial occlusion.

6. CONCLUSIONS

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

The main contributions of the proposed scheme are the introduction of a novel appearance based model for likelihood estimation and the use of a mutual information tracking cue in order to estimate the prior combined 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. 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.

7. 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] 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.

[4] 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.

[5] 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.

[6] T. Jebara and A. Pentland, "Parametrized structure from motion for 3d adaptive feedback tracking of faces," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1997, pp. 144–150.

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

[8] 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.

[9] 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.

[10] C. Tomasi and T. Kanade, *Shape and Motion from Image Streams: a Factorization Method - Part 3 Detection and Tracking of Point Features*, 1991.

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

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

[13] 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.

[14] 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, 2001.

[15] 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.

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

[17] 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.

(a) (b) (c)

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

Fig. 2. Tracking results under normal lightening conditions.

Fig. 3. Tracking results under different illumination conditions.

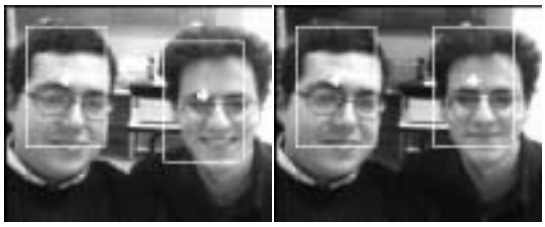

Fig. 4. Tracking results in a face image sequence containing two faces under varying lightening conditions.

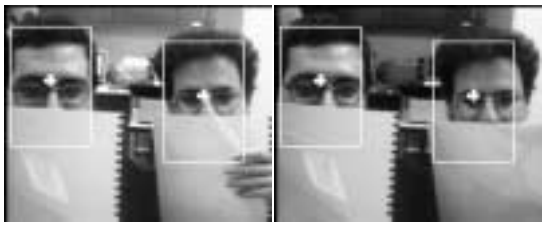

Fig. 5. Tracking results in a face image sequence containing two faces under varying lightening conditions and partial occlusion.