Face detection in a video sequence - a temporal approach

K. Mikolajczyk R. Choudhury C. Schmid
INRIA Rhône-Alpes GRAVIR-CNRS, 655 av. de l'Europe, 38330 Montbonnot, France
{Krystian.Mikolajczyk,Ragini.Choudhury,Cordelia.Schmid}@inrialpes.fr

Abstract

This paper presents a new method for detecting faces in a video sequence where detection is not limited to frontal views. The three novel contributions of the paper are :
1) Accumulation of probabilities of detection over a sequence. This allows to obtain a coherent detection over time as well as independence from thresholds. 2) Prediction of the detection parameters which are position, scale and pose. This guarantees the accuracy of accumulation as well as a continuous detection. 3) The way pose is represented. The representation is based on the combination of two detectors, one for frontal views and one for profiles.

Face detection is fully automatic and is based on the method developed by Schneiderman [13]. It uses local histograms of wavelet coefficients represented with respect to a coordinate frame fixed to the object. A probability of detection is obtained for each image position, several scales and the two detectors. The probabilities of detection are propagated over time using a Condensation filter and factored sampling. Prediction is based on a zero order model for position, scale and "pose" ; update uses the probability maps produced by the detection routine. Experiments show a clear improvement over frame-based detection results.

1 Introduction

In the context of video structuring, indexing and visual surveillance, faces are the most important "basic units". If the application is, for example, to identify an actor in a video clip, face detection is required as the first step.

Existing approaches either detect faces for every frame without using the temporal information or they detect a face in the first frame and then track the face through the sequence with a separate algorithm. This paper presents a novel approach which integrates detection and tracking in a unified probabilistic framework. It uses the temporal relationships between the frames to detect human faces in a video sequence, instead of detecting them in each frame independently.

The proposed algorithm first detects regions of interest which potentially contain faces based on detection probabilities [13]. These probabilities are propagated over time using a Condensation filter and factored sampling for prediction and updating [8]. We predict the detection parameters which are position, scale and "pose". The information about the predicted face location and scale can significantly accelerate the algorithm by narrowing the search area and scale. The accumulation of probabilities allows to continuously detect faces even in the frames where the probability based detector fails. Prediction also helps to obtain a stable detection over time, which is independent of scale dependent thresholds. Our pose representation allows to handle out of plane rotations which is considered to be a challenging task for any tracker. We can also handle the appearance and disappearance of faces by updating with the probabilities produced by the detection routine.

Experiments have been carried out on various video sequences with multiple faces occurring at different sizes. In addition, faces disappear from the sequence or get occluded and the pose of the face changes. We have applied the proposed algorithm on each of these sequences. Experimental results show a clear improvement over frame-based detection results.

Related work Past work has concentrated either on detection [13, 15] or on exploiting the temporal aspect in the form of face tracking [2, 7]. Detection can be feature based [18], based on a statistical model [13, 15], on colour [16], geometric shape [3] or motion information [7, 17]. Features may not convey complete information about the face. Model based approaches suffer from the drawbacks of the specific model - skin colour is susceptible to changes in lighting conditions and motion information may be distracted by alternate motion in the video.

In order to incorporate the face changes over time in terms of changes in scale, position and to localize the search for the face, it is essential to exploit the temporal information inherent in the videos. Face tracking [2, 7, 12] exploits the temporal content of image sequences. There have been a variety of trackers proposed (for faces or general objects). They involve feature tracking (contours, points) [7] or use information within the contour (colour [12]). Birchfield [1] combines these approaches to obtain a tracker which exploits the elliptical contour fitted to the face and the colour information enclosed. This can handle out-of-plane rotations and occlusions but is unable to handle multiple faces. Tracking can be also categorized on the basis of the face model : shape [3], colour [14] and statistical models [4]. It involves prediction and update for which filters like Kalman filter [6] and Condensation filter [8] have been used. Tracking requires initialization, which is mostly done manually. Furthermore, it does not handle the appearance of new targets in the scene.

Approaches which combine tracking and detection, mainly use detection for initializing the tracker [2] and then track the detected features through the sequence. This may be difficult for features like pupils [5] which are difficult to detect. Liu et. al [9] propose a method which incorporates non-frontal faces by updating. Pose variation has been handled via detection [11]. The framework of [10] can be used to track multiple faces, but it does not permit the addition of new faces.

The above survey highlights the contribution of our approach : we integrate detection with the temporal aspect of a video sequence. We propose a procedure for face detection which is robust to changes in scale and pose. Detection is used for initialization and then the detection information is integrated at each time step based on the propagated parameters. In addition, the detection probabilities provide information about new faces, which can be incorporated whenever they appear.

Overview The paper is organized as follows. In Section 2, we briefly describe our approach. Section 3 describes the detection algorithm and the pose representation. This is followed by a description of the temporal propagation in Section 4. Section 5 presents the results of our approach and compares them to results obtained by frame-by-frame detection.

2. A temporal approach to face detection

The paper proposes a method for identifying multiple faces in a video sequence. Initially a detector based on a local histogram of wavelet coefficients is used to associate a detection probability with each pixel. A probability value is computed for different scales and for two different views. By "pose" we mean the frontal and the profile views and the positions of the face in between these two views. The parameters which characterize the face are therefore the position, the scale and the "pose".

These parameters could be computed using a frame-by-frame detection but the detector response can decrease due to different reasons (occlusions, lighting conditions, face pose). Without any additional information these responses can be easily rejected even if they still indicate the presence of a face. This is due to a fixed threshold. Lowering the threshold increases the number of false detections, see for example the results presented by [13]. Furthermore, frame-by-frame detection is not continuous over time.

We therefore propose a temporal approach to detection which avoids these problems. The idea is to propagate detection parameters over time using the Condensation filter [8]. Prediction is based on a zero order model for position, scale and pose. The update at each stage uses the probability map generated by the detection routine.

Our proposed procedure is divided into two phases. The first phase is the detection which produces the probabilities for each image location, scale and viewpoint. This is described in section 3. The second phase is the prediction and update stage which predicts the detection parameters and uses the probabilities to track the face through the sequence. Temporal propagation is described in section 4.

3 Face detection

The face detection algorithm used in this paper is based on the method developed by Schneiderman and Kanade [13]. We use a local implementation of this detector. In the following, we briefly explain the detector as well as the computation of the probability score used in our temporal approach. Furthermore, we give implementation details and explain how to represent pose.

3.1 Detection algorithm

The detector uses the statistics of both face and non-face appearance. Statistics are based on joint probabilities of visual attributes. Visual attributes, referred to by $pattern_a$, are obtained by combining quantized wavelet coefficients at different sub-bands. We obtain 11 visual attributes representing different orientations and frequencies. Visual attributes are computed at different positions k: $pattern_a(k)$, where k is 16 for our experiments. Histograms are used to represent the joint statistics of visual attributes $pattern_a(k)$.

3.2 Probability score

The response of the detector is given by a weighted combination of visual attribute probabilities for the face and the non-face model. The weights are estimated from the training data in order to minimize the classification error. These responses are then normalized to associate face probabilities with each pixel. A probability corresponds to a given location and a given scale. Probabilities have to be computed for all image positions and for different scales. The ratio between two consecutive scale levels has been chosen empirically as 1.2. In order to detect frontal and profile views we use two detectors (cf. below). The probabilities of the frontal and profile detectors are $P_f(I, x, y, s)$

Figure 1. Varying face pose. The frame numbers are 1, 3, 4, 6 and 8. A cross indicates the location of local maxima for the frontal detector and a circle the location of local maxima for the profile detector.

and $P_p(I, x, y, s)$ respectively. The coordinates (x, y) is the position at which the detector is applied and s is the scale.

We obtain a probability map for each image location where the local maxima correspond to potential face locations. In the frame-by-frame approach a face is detected, if the maximum is above a threshold. Maxima over scales are merged by collision removal.

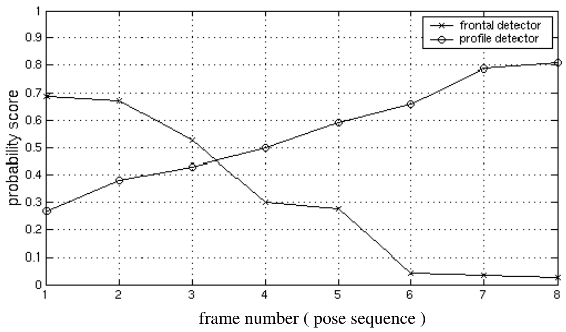

Figure 2. Probability scores for frontal and profile detectors. The sequence and the frame numbers for face pose correspond to those of figure 1.

3.3 Implementation details

The probability of a pattern is estimated from the training data. The frontal face and profile face model was built with about 200 different faces each. The faces were normalized with respect to size and orientation. To increase the training set, we created for each face several smoothed and rotated versions. The profile faces in the training set included images of faces with angles between 45 and 90 degrees. The non-face model was initially learnt from 200 images containing no faces. We then reapplied the detection procedure to the non-face images and updated the non-face model with the locations that gave the highest detection responses. The training set of face and non-face images was collected from the internet and none of the test video sequences were included in our training set.

3.4 Pose representation

In order to detect faces from any viewpoint, we have used two detectors. One was trained to detect frontal views, the other to detect profiles. Note that the profile detector is also applied to the mirror-reversed image. We keep the higher response as the probability of profile. We will see in the following that using these two detectors is sufficient to detect all intermediate views. Figure 2 shows the probabilities of the two detectors for varying pose. The corresponding pose images are displayed in figure 1. We can see that the probability of the frontal view detector decreases, as the face turns sideways and vice versa. The combination of the two detectors allows us to approximatively estimate pose. This approximation is sufficient for prediction/update. If we only use one of the detectors, we cannot predict whether the face is turning or disappearing in the case of decreasing probability. Note that the maximal response for the two detectors, when applied to intermediate face poses, has slightly different localizations (figure 1). We have estimated the displacement vector from our training set.

4. Temporal propagation

We have developed a framework for prediction and update which propagates the probabilities of detection and the detection parameters over time. The CONDENSATION (CONditional dENsity propagaTION) algorithm proposed by Isard and Blake [8] has been used. The probability distribution over all the detection parameters is represented by random samples. The distribution then evolves over time as the input data / observations change. The Condensation filter, unlike the Kalman filter, can represent non-Gaussian densities and can handle multiple modes and alternate hypotheses. In addition, the present state of detection parameters is conditional to their past state, which are estimated from the data sequence by the filter. Factored sampling helps in propagating over time the samples with higher associated weights. This is required in our scenario as the faces with higher probability score need to be propagated over time. Thus the Condensation filter combined with factored sampling is most appropriate for our purpose.

4.1 The adaptation of Condensation

The detection parameters which we need to predict are
• (x, y) : position. Position ranges over the entire image.
• s : scale at which the face occurs. We use a discrete range of scales which have been empirically chosen (cf. section 3).
• θ : the parameter indicating the pose. θ can take any value between 0 and 90 degrees and indicates the angle. $0°$ indicates a frontal face and $90°$ the profile. Note that we do not distinguish between the two profiles (left and right).

The *state* at time t, $\mathbf{s_t}$, is defined to be a vector of parameters $\mathbf{s_t} = (x_t, y_t, s_t, \theta_t)$. The *observations* at each stage are the probability values computed by the detector in section 3. The probability $P(I_t, x_t, y_t, s_t)$ is the value associated with each pixel of the image and associated with a particular scale. There are two different probabilities associated with the two "poses" of the head ; $P_f(I, x, y, s)$ corresponding to the frontal face detector and $P_p(I, x, y, s)$ corresponding to the profile detector. The observation z_t is then given by :

$$z_t = (P_f(I, x, y, s), \quad P_p(I, x, y, s))$$

These probability values indicate the likelihood of observations. The conditional probability $P(z_t|\mathbf{s_t})$ is the probability of observation z_t given the state $\mathbf{s_t}$. Given this conditional probability distribution, a discrete representation of the entire probability distribution can be constructed over the possible states. Our proposed algorithm is divided into 4 steps. Note that probabilities are denoted by $P(I, x, y, s)$ using the suffixes for the frontal and side views used only if we need to distinguish them.

Step I : Initialization We initialize with the local maxima of detection in the initial frame of the video sequence. Each of the maxima is propagated separately. Note that we do not use a threshold as in the case of detection. Maxima corresponding to non-faces are eliminated over time.

1. We pick a Gaussian random sample around each of the maxima (x, y). We keep the scale fixed in the initial sampling to consolidate the maxima over the scale. We pick 300 samples around each maxima.

2. We pick the probabilities corresponding to the samples from the respective positions of the frontal and profile faces. We initialize pose with $\theta_0 = \frac{P_p}{P_f + P_p} \times 90°$. The corresponding total probability P is then given by

$$P(I_0, x_0, y_0, s_0) = \begin{cases} P_f(I_0, x_0, y_0, s_0) & \text{if} \quad 0° < \theta_0 \leq 45° \\ P_p(I_0, x_0, y_0, s_0) & \text{if} \ \ 45° < \theta_0 \leq 90° \end{cases}$$

   The set of probabilities are normalized to produce the weights $w_i = \frac{P_i}{\Sigma_{i=1}^{S} P_i}$ where S is the total number of samples which are being propagated.

The sample states and the weights are used to predict the probability distribution at the next time instant. The next three stages set up a new probability distribution at time t given the distribution at time t-1.

Step II : Selection We use factored sampling [8] to sample the states at stage t-1. These are then propagated to the next time instant depending on the associated weights.

Step III : Prediction We use a zero order temporal model for the prediction of a new state

$$\begin{aligned} x_t &= x_{t-1} + N(\sigma_x) \\ y_t &= y_{t-1} + N(\sigma_y) \\ s_t &= s_{t-1}(1.2)^{\text{round}(k)} \ \text{ with } \ k \in N(\sigma_s) \\ \theta_t &= \theta_{t-1} + N(\sigma_\theta) \end{aligned}$$

Note : In the above, the scaling factor of 1.2 has been empirically chosen. It has shown to give the best results for scale. For our experiments we have set the parameters σ_x and σ_y to 5, σ_s to 0.5 and σ_θ to 1. The prediction approximates the conditional probability of the present state given the previous state.

Step IV : Updating The probabilities of the predicted states are obtained by weighting the responses P_f and P_p with the difference between the predicted and observed pose. The combined response associated with each state is
$P(I_t, x_t, y_t, s_t, \theta_t) = \max(f(\lambda_t, \theta_t)P_f, f(\lambda_t, \theta_t)P_p)$
where θ_t is the predicted pose and $\lambda_t = \frac{P_p}{P_f + P_p} \times 90$ is the observed pose at time t. $f(\lambda_t, \theta_t)$ is a linear function which is 1 when $\lambda_t = \theta_t$ and decreases as the difference between the two increases. The aim is to maintain a high response if the predicted pose is in the direction of the observed pose and the response should decrease if the prediction is incorrect. A lower response causes the sample to be rejected in the next stage.

To consolidate the propagated samples at each stage we find the weighted mean $Mean(\mathbf{S}) = \Sigma_{i=1}^{S} w_i \mathbf{S}_i$ and variance $Var(\mathbf{S}) = \Sigma_{i=1}^{S} w_i \mathbf{S}_i^2 - Mean(\mathbf{S})^2$ of the states $\mathbf{S}_i = (x_i, y_i, s_i)$. The mean value indicates the position and the scale of the face. This allows stabilization over time. The variance is stable over time for the faces and increases for non-faces, that is gets diffused over time.

4.2 Incorporating the appearance of new faces

To handle the situation when new faces appear in the scene, we update our set of local maxima at every nth frame where n is equal to 5 in our experiments. This allows to incorporate new faces which appear in the sequence. Faces which are lost are the ones which get diffused over the image, that is for which the variance increases.

5. Experimental results

Experiments have been carried out on different video sequences with multiple faces occurring at various sizes and positions. Furthermore, faces may disappear from the sequence (Fig. 3, frame 20). In the following we compare the frame-by-frame detection method with our "temporal"

Figure 3. The first column : frame-by-frame detector. The second column : "temporal" detector. The frame numbers are marked above the images.

detection. Fig. 3 and Fig. 4 show the results for two sequences. The first column displays the frame-by-frame detection results and the second column gives the results of the "temporal" detector.

Frame-by-frame detection The detection algorithm of Section 3 was applied to each frame of the video sequences. The occurrence of a local maximum indicates the presence of a face. A face is actually detected if the maximum is above a fixed threshold. Results for individual frames are shown in the first columns of Fig. 3 and Fig. 4. The scale at which the face is detected determines the size of the bounding box. It has been previously observed that the results are very sensitive to the threshold used [13]. They have shown that the missing detection and the false detection rate vary significantly depending on the threshold.

"Temporal" detection The temporal framework is applied to each of the video sequences. The strongest local maxima associated with the first frame of the sequence are used to initialize the procedure. Without loss of generality, we have used the 4 strongest maxima for our experiments. These maxima are then propagated throughout the temporal sequence. Weighted mean and variance is computed for each frame using all the samples. The mean value for position and scale are used to display the position of the face and to draw the corresponding bounding box. This smoothes the size and eliminates the discontinuities of the frame-based detection. Results are shown in the second columns of figure 3 and figure 4. Figure 5 shows the results of the pose detection applied to a sequence. The combination of the two detectors frontal and profile is shown. The response from the frontal detector rapidly diminishes as the face turns away from the camera and the profile detector responses (indicated by the triangle embedded in a square) increases. These are combined by the temporal approach to give an integrated response.

Observations

• The face of the man was not detected by the simple detector in frame 1 of Fig. 3. There is however a local maximum with a low probability at the location of the man. In the case of the "temporal" detector, this maximum is picked for initialization and propagated/increased over time. The frame-by-frame detection fails, because the probability score is below the detection threshold.

• In frame 1 of Fig. 3 two of the four maxima used for initialization are non-faces. The probability of non-faces goes to zero in subsequent frames and are therefore eliminated.

• The faces continue to be tracked even in the case of a head movement and out-of-plane rotation, while the frame-by-frame detection loses some of these faces as the associated probability score drops below a threshold (cf. Fig. 4, frame 23).

• When occluded faces reappear in the sequence, they are detected much later by the simple detector compared with the temporal detector (Fig. 4, frames 53 and 67).

Evaluation We applied the frame-based approach and the temporal approach to 760 images of various video sequences. The frame-based approach gave 10.6% false detections and 20.8% missing detections. The temporal approach was able to remove all false detections and all faces were continuously detected through the video sequence.

Our experiments show that "temporal" detector clearly improves results over the frame-based detection results.

6. Conclusions and perspectives

We have presented a novel approach for temporal face detection which gracefully integrates detection for individual frames with the temporal aspect. This allows to improve on individual detection results and at the same time to track through the sequence.

In the future, we propose to include a motion model, incorporate alternate cues like colour etc. and investigate the gain obtained by an additional (third) detector for intermediate views (45 degrees).

Acknowledgements This work was supported by the French project RNRT AGIR and the European FET-open project VIBES. We would like to acknowledge the "Institut National de l'Audiovisuel" for giving permission to use their video material

Figure 4. Sequence with occluded faces. First column : frame-by-frame detector. Second column :"temporal" detector. The frame numbers are marked above the images.

Figure 5. Results of pose detection. As the face turns the response of the frontal detector (square) decreases and the profile detector (triangle embedded in a square) increases. We have the displayed the detector for which the maximum response it obtained.

displayed in figures 1, 3, 4 and 5.

References

[1] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR*, pp. 232–237, 1998.

[2] D. Comaniciu, V. Ramesh, and P. Meer. Real time tracking of non-rigid objects using mean shift. In *CVPR*, pp. 142–149, 2000.

[3] D. Decarlo and D. Metaxas. Deformable model based face shape and motion estimation. In *FG*, 1996.

[4] C. J. Edward, C. J. Taylor, and T. F. Cootes. Learning to identify and track faces in an image sequence. In *FG*, pp. 260–265, 1998.

[5] R. S. Feris, T. E. de Campos, and R. M. C. Junior. Detection and tracking of facial features in video sequences. In *Lecture notes in AI*, pages 197–206, 2000.

[6] A. Gelb, editor. *Applied Optimal Estimation*. MIT Press, 1992.

[7] G. Hager and K. Toyama. X vision : A portable substrate for real-time vision applications. *CVIU*, 69(1):23–37, 1998.

[8] M. Isard and A. Blake. Condensation-conditional density propagation for visual tracking. *IJCV*, 29:5–28, 1998.

[9] Z. Liu and Y. Wang. Face detection and tracking in video using dynamic programming. In *ICIP*, 2000.

[10] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *ICCV*, pp. 572–578, 1999.

[11] S. McKenna, S. Gong, and J. J. Collins. Face tracking and pose representation. In *BMVC*, pp. 755 – 764, 1996.

[12] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying lighting conditions using color. In *FG*, pp. 228–233, 1998.

[13] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *CVPR*, volume I, pp. 746–751, 2000.

[14] K. Schwerdt and J. Crowley. Robust face tracking using colour. In *FG*, pp. 90–95, 2000.

[15] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *PAMI*, 20(1):39–51, 1998.

[16] J.-C. Terrillon, M. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *FG*, pp. 54–61, 2000.

[17] J. Yang and A. Waibel. Tracking human faces in real time. TR CMU-CS-95-210, CMU, 1995.

[18] K. C. Yow and R. Cipolla. Feature based human face detection. TR 249, University of Cambridge, 1996.