SPATIAL-TEMPORAL VIDEO ANALYSIS FOR PEDESTRIAN DETECTION: APPLICATION TO DENSITY ESTIMATION AND TRACKING IN DEMONSTRATIONS.

Josep R. Casas, Aleix Puig, Pere Puig

Signal Theory and Communications Department
UPC – Technical University of Catalonia
Campus Nord UPC, edifici D5
Jordi Girona 1-3, 8034 Barcelona, Spain

ABSTRACT

Crowded video sequences like those of demonstrations offer an interesting challenge for object extraction and tracking due to their complexity: taken outdoors, often in extreme illumination conditions, with faces not in frontal view, perspective, complex background, etc. Tracking of the individuals becomes a difficult task due to the high number of occlusions. In order to deal with these problems a mutual feedback spatial-temporal detection algorithm is proposed. The system improves its efficiency thanks to a cooperative approach between spatial detection and temporal tracking. Spatial detection is based on skin color classification and shape analysis by morphological tools. Temporal tracking is based on the analysis of the optical flow. The mutual feedback approach improves both spatial detection and temporal tracking. In order to deal with multiple occlusions, a graph-based approach taking advantage of the neighborhood consistency has been introduced.

1. INTRODUCTION

Many techniques have been developed for the detection and tracking of people in crowded scenes [9,10]. In order to obtain a reliable and objective estimate of the number of people attending a demonstration, one may follow different approaches, depending on the type of demonstration. For static demonstrations, a usual strategy is to estimate the local density of people along the streets and places filled by the crowd. Then, given the total surface of this area, one can obtain an estimate of the overall number of attendees. For demonstrations where the crowd walks along the streets, a simpler strategy is to estimate the flow of people who cross one or several reference lines along the way. In the latter case, video analysis can contribute powerful tools in order to automate the people count.

Video sequences under analysis look like the image presented in Figure 1. These images pose interesting challenges to state of the art techniques for video object extraction and tracking, basically due to the complexity of the objects under analysis, and the different effects due to occlusion, shadowing, perspective deformation, complex motion and insufficient image definition. In this paper, we approach the problem of extracting and tracking people in a demonstration by means of a spatial detection and a temporal tracking approach.

This material is based upon work partly supported by the IST programme of the EU through the NoE IST-2000-32795 SCHEMA and by the grant TIC2001-0996 of the Spanish Government

We apply spatial detection in selected key-frames in order to locate people in the crowd. Then we track the detected persons along the sequence, until the next key-frame. **Spatial detection and temporal tracking complement each other by mutual feedback in a synergetic approach**. Our target is to show the improvement in robustness contributed by the spatial-temporal feedback, which overcomes the aforementioned problems.

Figure 1. Example of a demonstration image

2. STRATEGY

The dual spatial/temporal processing is represented in Figure 2. The **Spatial Detection (SD) block** and the **Temporal Tracking (TT) block** have symmetric structure, as described below.

The SD block analyzes the original video image using an initial estimate of the detected objects. The evolution of the features for each object is stored in a feature vector and also input into the SD block. From these data, the SD block updates the estimated mask of the detected objects by means of a spatial segmentation algorithm. The feature vector and the initial estimate are used to control the spatial segmentation and detection parameters.

The TT block analyzes each original video image and its variation with respect to the previous frames. The feature vector is also input into the TT block. From these data, the TT block updates the estimated mask of the detected objects by means of a temporal tracking algorithm. In this case, the feature vector and the initial estimate control the object-tracking parameters.

The cooperation of the two independent blocks takes place at processing time, as shown in figure 3. The following sections explain the techniques used in each one of the building blocks and the cooperation process in more detail.

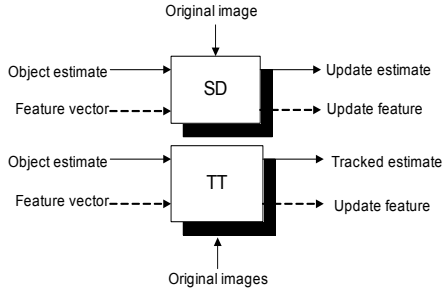

Figure 2. Spatial Detection (SD) and Temporal Tracking (TT) blocks

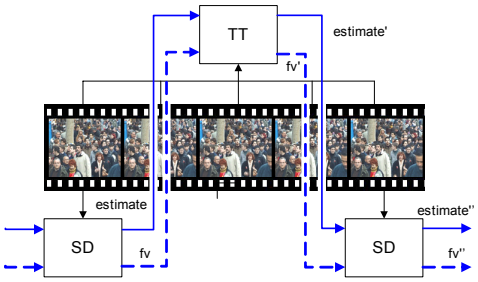

Figure 3. Interaction of the SD/TT blocks at processing time

3. SPATIAL DETECTION

The Spatial Detection block aims at detecting the faces of the people in the demonstration, as faces are the most distinctive features of the objects to be extracted. Color, shape and size criteria are used to assess the segmentation algorithm.

Figure 4 shows the overall processing procedure for the Spatial Detection block.

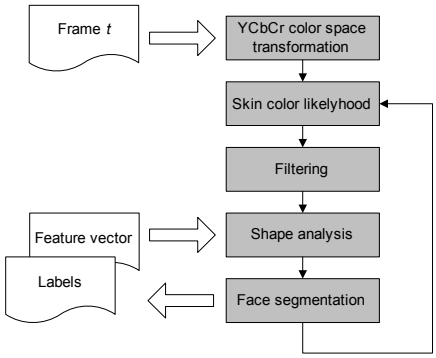

Figure 4. Spatial detection algorithm

3.1 Skin color likelihood

A two dimensional Gaussian function on the chrominance plane of the YC_bC_r color space is used as human skin color model [4]. In order to improve performance, the Gaussian parameters are made adaptive to the characteristics of the scene. This results in a skin-color-likelihood image, which will be used once filtered and binarized in the morphological shape analysis step.

3.2 Morphological shape analysis

Most shape analysis techniques for face detection are based on the explicit knowledge that the shape of a face is approximately an ellipse. In crowded scenes, mostly due to partial-occlusion problems, it is quite common to detect regions including two or more overlapping faces, which do not have an elliptical shape. The algorithm handles this problem by analyzing the skeleton of the detected skin regions. We apply the definition of the skeleton transformation proposed by Calabi [1], as the set $S(X)$ of maximal balls $S_n(X)$.

The first step is to eliminate false skin positive detections (x_i) from the original set of detected regions (X), whose size is smaller than faces, like hands or handbags, or larger, like coats.

$$X' = X - \bigcup_i x_i \quad | \quad \begin{cases} \max\left\{ S(x_i) = \bigcup_n S_n \right\} > m + \sigma_{\sup} \\ \text{or} \\ \max\left\{ S(x_i) = \bigcup_n S_n \right\} < m - \sigma_{\inf} \end{cases}$$

where m is the mean size of a human face, and σ is a threshold. This analysis considers the perspective deformation inferred from the lines shown in figure 1. A perspective factor is applied on any decision taken on shape and size features thereon.

To deal with regions potentially including multiple overlapping faces, the skeleton is iteratively analyzed. Region shapes can be decomposed into p (a priori unknown) faces, which are allowed to overlap up to factor r.

$$\begin{cases} S'_1(X') = S(X') \setminus S_{k_1} \\ S'_i(X') = S'_{i-1}(X') \setminus S_{k_i} \quad i = 2..p \end{cases}$$

$$S_{k_i} = \bigcup s \quad | \quad \delta_{n_s - r_s}(s) \cap \delta_{n_k + 1 - r_k}(k_i) \neq \varnothing$$

$$\text{where } s \neq k_i \quad s, k_i \in S'_{i-1}(X') \text{ and } \begin{cases} r_s = \dfrac{n_s}{n_s + n_k} \cdot r \\ r_k = \dfrac{n_k}{n_s + n_k} \cdot r \end{cases}$$

The resulting skeleton is a set of p values, which correspond to the center of the detected faces. The result of the inverse skeleton transformation of $S'(X')$ is used as the marker image required for the segmentation step. The inverse skeleton transformation takes into account the overlapping factor $r>0$ in order to avoid overlapping markers.

$$X'' = \bigcup_k \delta_{n_k - r_{k_{\max}}}(k) \text{ where } \begin{cases} r_{k_{\max}} = \max\left\{ r_k = \dfrac{n_k}{n_s + n_k} \cdot r \right\} \forall s \in S'(X') \\ k \in S'(X') \end{cases}$$

3.3 Face segmentation

Once the markers for the faces have been detected, their shape must be accurately determined in order to track them in the video sequence. The watershed algorithm is applied on the morphological gradient of the luminance component of the original image for face reconstruction.

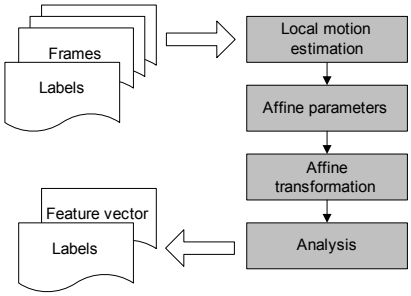

Figure 5. Temporal Tracking algorithm

4. TEMPORAL TRACKING

Local motion estimation is performed using the algorithm presented by Lucas-Kanade [6]. It consists of a multi-scale coarse-to-fine procedure based on a gradient approach. After linearizing and minimizing respect to $V_x(x,y)$ and $V_y(x,y)$ the following system is obtained:

$$\begin{bmatrix} \hat{V}_x(t) \\ \hat{V}_y(t) \end{bmatrix} = \begin{bmatrix} \sum_{x\in B} Ex^2 & \sum_{x\in B} ExEy \\ \sum_{x\in B} ExEy & \sum_{x\in B} Ey^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_{x\in B} ExEt \\ -\sum_{x\in B} EyEt \end{bmatrix}$$

where *Ex*, *Ey* and *Et* are the estimated gradients.

As in the Lucas-Kanade algorithm described by Barron et al [7], the gradients are estimated by using a 1x5 kernel, so that a set of 5 frames is needed –i.e. the current frame, two previous frames and two subsequent frames. As suggested by Simoncelli, Adelson and Heeger [5], the gradients are also smoothed in order to weight their contribution. Due to the complex nature of the images and the large error that would be introduced, we omit any pre-smoothing of the original images, which otherwise would be normally applied in order to reduce errors in the gradient estimation.

Using the labelled image provided by the Spatial Detection Block, the **affine parameters** within each region are estimated by standard regression techniques. Affine motion is defined by the following equations:

$$\begin{pmatrix} V_x \\ V_y \end{pmatrix} = \begin{pmatrix} a_{xx} & a_{xy} \\ a_{yx} & a_{yy} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_{xo} \\ a_{yo} \end{pmatrix}$$

where a_{ij} stands for the affine motion parameters.

As defined in [8], the linear least squares solution for the affine motion parameters within a given region, R_i, is as follows:

$$\begin{bmatrix} a_{xi} & a_{yi} \end{bmatrix} = \left[\sum_{R_i} \phi\phi^T \right]^{-1} \sum_{R_i} \left(\phi \begin{bmatrix} V_x(x,y) & V_y(x,y) \end{bmatrix} \right)$$

The size of the regions depends on the labels obtained in the Spatial Detection Block. A region might have multiple motion modalities if its label contains multiple faces. The analysis of the homogeneity of the affine approximation of the velocity helps us in the detection of the labels conveying multiple faces (those regions with multiple motion modalities).

The next step in the motion block is the **affine transformation**, which propagates the labels to the next frame. Considering $V_x=x'-x$ and $V_y=y'-y$, being (x',y') the next frame and (x,y) the current frame, then:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a_{xx}+1 & a_{xy} \\ a_{yx} & a_{yy}+1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_{xo} \\ a_{yo} \end{pmatrix}$$

is the equation relating current frame and next frame positions. We apply backward motion compensation instead of forward motion prediction: by evaluating each point of a neighbourhood N_{label} of the given label L, we determine which $(x',y') \in N_{label}$ comes from a $(x,y) \in L$. Backward motion compensation ensures that the spatial connectivity of the label is kept.

The main purpose of the **analysis block** is to obtain motion information in order to improve the global performance of the tracking algorithm. Information such as the direction of propagation, the module of the velocity vector, the global optical flow, the splitability of the labels, etc. can be crucial in disambiguating occlusion situations. All the data is stored in the feature vector, updated at each iteration of the algorithm.

Finally the predicted motion error needs to be evaluated. We use a normalized DFD to provide, apart from the motion data, the reliability of the prediction for each region. Given a label R_i, we compute its prediction error as follows:

$$DFD_{Ri} \equiv \left| \frac{\hat{I} - I_t}{size_{Ri}} \right|$$

where $\hat{I}$ is the interpolated prediction of the initial content of the label and I_t are pixel values under the label in the current frame.

5. MUTUAL FEEDBACK

The SD block uses temporal information to improve its detection rates without increasing false positive rates. This is achieved by increasing its sensitivity in those areas where the TT block predicts the position of a tracked face. Temporal tracking provides information of the number of faces p that might be included in a given skin region, which allows increased robustness in partial occlusion situations. In turn, the TT block improves its performance thanks to the analysis carried out in the SD block. Tracking errors are corrected at every key frame, when spatial detection is performed.

The mutual feedback approach not only increases the performance of each individual block. The fusion of spatial and temporal information in a unique feature vector allows face identification along the sequence. Occlusions are dealt with by using temporal and spatial coherence in consecutive frames.

After a labeling step in the current frame, a label-matching algorithm is applied to map the faces in the last frame to the faces detected in the current frame. This provides valuable dynamic information like size evolution, motion vectors and neighborhood information, which is used as the feature vector in the feedback schema in order to improve detection in forthcoming frames. A graph-based tracking technique [3] is also applied to take advantage of this neighborhood consistency. The graph-based tracking technique solves region occlusion, reappearance, region merging and region splitting situations.

The first step of the face matching algorithm tries to map detected faces in the predicted partition $\hat{P}_t$ (based on detection P_{t-k}), to the faces detected in partition P_t. Face similarity between faces $f_i \in \hat{P}_t$ and $f_j \in P_t$ is given by the probability $P(i \rightarrow j)$, which is a weighted function of face size S, the position of the center of mass M_C and a neighborhood measure N:

$$P(i \rightarrow j) = \alpha \cdot (S_j - \hat{S}_i) + \beta \cdot (M_{Cj} - \hat{M}_{Ci}) + \gamma \cdot (N_j - \hat{N}_i)$$

The next step deals with *disappearances*, *appearances* and *reappearances* by exploiting the information in the feature vector, base of the communication between the two blocks. This vector contains up to date information of each matched region and updated parameters characterizing the features of every face. When a face disappears in the video sequence, it is labeled in the feature vector as *occluded*, and all its attributes are kept, including its hypothetically predicted position for every frame. This allows taking occluded faces into account for the matching step, when a face appearing in P_t but not in $\hat{P}_t$ has to be labeled as new or as a reappearance of a face that was occluded in some frame between the interval [0,t). Only occluded faces which predicted position leaves the camera field of view are discarded, which helps to keep the memory of the feature vector at a reasonable size.

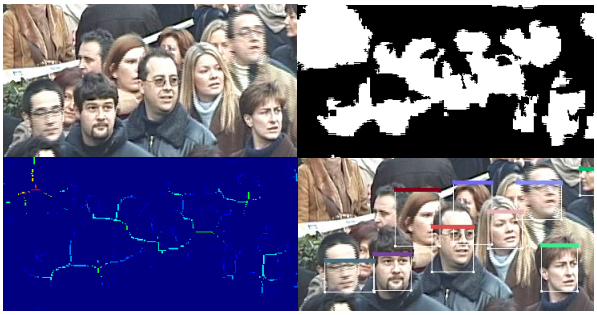

Figure 6. Partial occlusion. From left to right: original image, binarized skin-likelihood masks, skeleton transformation and detected faces.

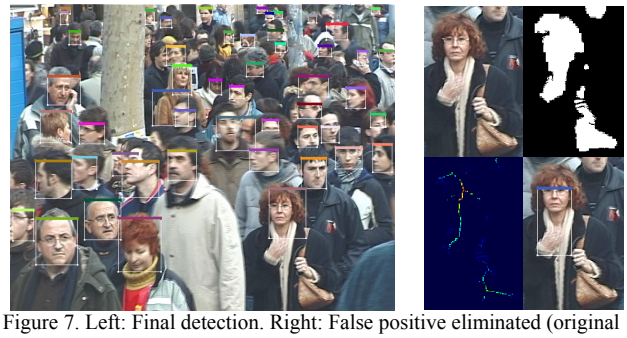

Figure 7. Left: Final detection. Right: False positive eliminated (original image, binarized skin-likelihood mask, skeleton and detected face).

6. RESULTS

As shown in figure 6, the algorithm solves occlusion problems, where partial overlapping causes connection between two or more faces in the skin likelihood image. Figure 7 shows how morphological analysis is able to eliminate false positives from skin color detection.

The Spatial Detection block uses temporal tracking information to improve detection (figure 8). If a face has been tracked to a given area, the skin color detector lowers its sensitivity threshold in that area of the image at the following frames. Thus, correct detection rates increase while false positive rates remain unchanged. Figure 9 shows the prediction error, obtained by computing the difference between the centers of mass of a given region in the ground truth ('Dist GT') and the predicted one. It can be observed that the mutual feedback improves tracking performance, decreasing the error at each key frame.

7. CONCLUSIONS AND FURTHER RESEARCH

We have presented a cooperative spatial detection – temporal tracking procedure aimed at people counting in video sequences of demonstrations. The originality of the approach is its ability to combine positive contributions of both algorithms. Spatial detection is improved by searching for target objects in the areas predicted by the temporal tracking algorithm. In this way, we can deal with occlusions, splits and crossovers. Temporal tracking is updated every N frames with the results from spatial detection, so that tracked object positions can be corrected when misalignments occurs. In this paper we are presenting the results of an on-going research project targeting the challenge of object extraction and tracking in complex video sequences. Further work will focus on illumination problems and graph-based techniques in order to use neighborhood information (neighboring objects) in order to disambiguate false detections or miss-detections.

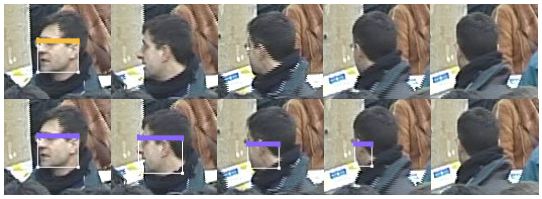

Figure 8.First row: results of SD without mutual feedback. Second row: results of SD with mutual feedback

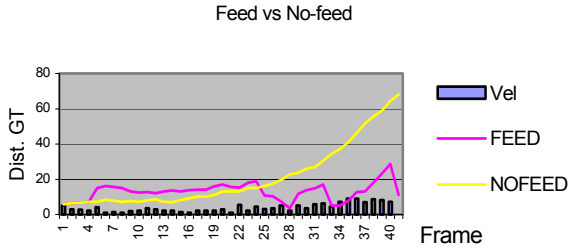

Figure 9. Improvement in Temporal Tracking

8. REFERENCES

[1] L.Calabi and W. Harnett. *Shape recognition, prairie fires, convex deficiencies and skeletons*. Technical Report 1, Parke Math. Lab. Inc., One River Road, Carlisle MA, 1966.

[2] Vincent, L., *Mathematical Morphology in Image Processing*, E. Dougherty, Editor, Marcel-Dekker, New York, September 1992.

[3] C. Gomila and F. Meyer, *Graph-based object tracking*, Proc. of International Conference on Image Processing (ICIP), Barcelona, Spain, September 2003.

[4] L. Gu and D. Bone, "Skin Colour Region Detection in MPEG Video Sequences", Proc. of International Conference on Image Analysis and Processing (ICIAP), Venice, Italy, September 1999.

[5] Simoncelli, E.P., Adelson, E.H., Heeger,D.J., *Probability distributions of optical flow. Computer Vision and Pattern Recognition*. Proceedings CVPR'91, IEEE Computer Society Conference on Pattern Recognition, 3-6 June 1991.

[6] Lucas, B., and Kanade, T. *An Iterative Image Registration Technique with an Application to Stereo Vision*, Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI) 1981, pp. 674-679.

[7] Barron, J.L., Fleet, D.J., Beauchemin, *Performance of Optical Flow Techniques*. Proc. Of International Joint Conference of Computer Vision, 1994, Vol.12, No.1, pp. 43-77.

[8] Wang, J.Y.A, and Adelson E.H., *Spatio-temporal segmentation of video data*, Proc of the SPIE: Image and video processing II, vol. 2182 1994.

[9] B. A. Boghossian, S. A. Velastin, *Real-time motion detection of crowds in video signals*, IEE Colloquium on High Performance Architecture for real-time image processing, pp. 12/1- 12/6, February 1998.

[10] S. Regazzoni, A. Tesei, V. Murino *A real-time vision system for crowding monitoring* Proc. of the International Conference on Industrial Electronics 1993 (IECON'93) pp/1860-1964 vol.3