A monocular system for automatic face detection and tracking

G.N. Stamou[a], M. Krinidis[a], N. Nikolaidis[a] and I. Pitas[a]

[a]Department of Informatics,
Aristotle University of Thessaloniki,
Box 451, 54124 Thessaloniki, GREECE

ABSTRACT

This paper presents a complete functional system capable of detecting people and tracking their motion in either live camera feed or pre-recorded video sequences. The system consists of two main modules, namely the detection and tracking modules. Automatic detection aims at locating human faces and is based on fusion of color and feature-based information. Thus, it is capable of handling faces in different orientations and poses (frontal, profile, intermediate). To avoid false detections, a number of decision criteria are employed. Tracking is performed using a variant of the well-known Kanade-Lucas-Tomasi tracker, while occlusion is handled through a re-detection stage. Manual intervention is allowed to assist both modules if required. In manual mode, the system can track any object of interest, so long as there are enough features to track.

Keywords: Object tracking, face detection, fusion, occlusion

1. INTRODUCTION - SYSTEM DESCRIPTION

Video-based tracking of the motion of the human body has been a challenging research topic with applications in many domains such as human-computer interaction, surveillance, hand gesture recognition and 3D reconstruction. Such a task is usually preceded by an initialization step that aims at detecting the presence of people. The latter has been often tackled by face detection. However, pose variations (frontal, profile and intermediate poses), skin-color variations, facial structural components, (moustache, beards and glasses), occlusion and poor or variable imaging conditions make this task a rather difficult one. For details on face detection methods, the reader is referred to Ref. 1.

Tracking techniques can be divided into active and passive tracking. For a review of the former, Ref. 2 is recommended. Computer vision researchers have been trying to achieve results comparable to active tracking using passive techniques for a long time, in an effort to produce generally applicable motion tracking systems for uncontrolled (indoor or outdoor) environments. For a comprehensive review of passive tracking methods, the reader is referred to Ref. 3, 4.

The goal of this work is to present a system that aims at robust face detection and tracking, as well as object tracking in general. Our approach for face detection was motivated by Ref. 5 and Ref. 6 and involves fusion of information available from two separate detectors in order to produce more accurate results than each detector alone, as well as to complement each other in case of failures. The tracking algorithm of this system is a variant of the Kanade-Lucas-Tomasi tracker (Ref. 7), capable of dealing with still or slowly moving features. The proposed system can operate in two different modes (automatic and semi-automatic) and is capable of tracking either automatically detected faces or any other manually selected object(s) of interest. In the semi-automatic mode, user intervention is required to initialize the regions to be tracked in the first frame of the video sequence. Manual intervention is also allowed in other cases, such as the initialization of the tracking algorithm for new faces entering the scene, re-initialization if any of the tracked faces is lost and correction of erroneous tracking results. The latter refers to stopping the tracking of erroneously detected objects, as well as correcting the tracked region, so as not to contain portions of the background. Obviously, in case of manual initialization, the system can be used to track any object(s) of interest, other than faces. In its default configuration, it can cope with a range of different environments. However, a number of parameters can be fine-tuned. An overview of the system is illustrated in Figure 1 (a).

Further author information: (Send correspondence to Ioannis Pitas)
I. Pitas: E-mail: pitas@aiia.csd.auth.gr, Telephone: +302310996304

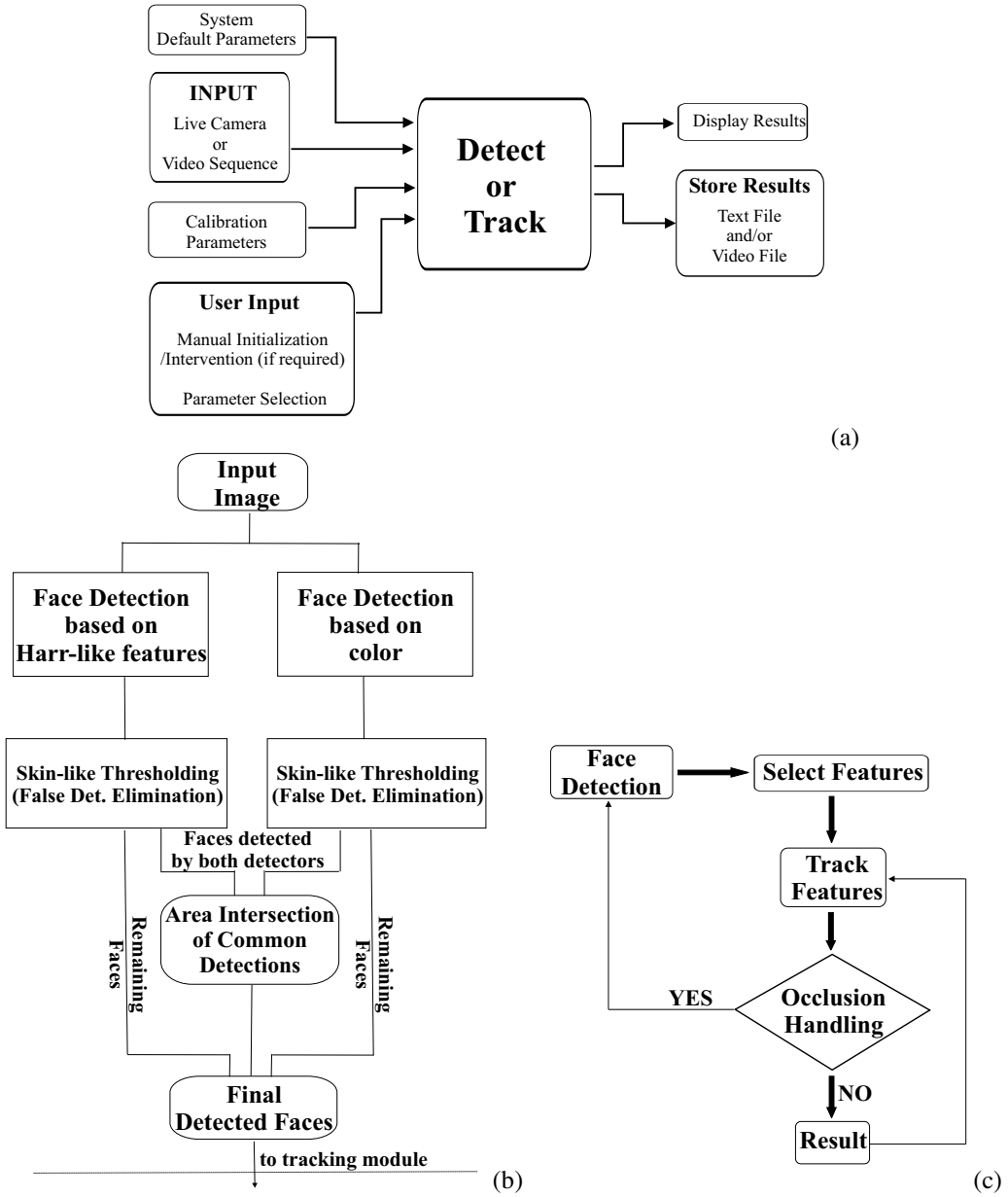

Figure 1. Schematic Diagrams: (a) Overall system, (b) Detection module and (c) Tracking module

Novel contributions of this paper include the addition of a color-based thresholding step into the frontal face detector presented in Ref. 6, in order to reduce false detections in complex scenes. Additional geometrical criteria, as well as a facial feature extraction step are also employed in order to make a color-based face detection algorithm similar to the one presented in Ref. 5 more robust to false detections. Moreover, a fusion scheme that combines the results of the two separate detectors is developed, aiming at reliable detection of faces in various poses (frontal, profile, intermediate) and orientations. However, the main contribution of this paper is the implementation and testing of a complete functional system, which incorporates all the above and aims at detecting and tracking people in live camera input or pre-recorded video sequences.

The remainder of the paper is organized as follows. The face detection algorithm is presented in Section 2. In Section 3 the tracking process is introduced. Section 4 presents experimental results, while in Section 5 the final conclusions are

drawn.

2. FACE DETECTION BASED ON FUSION

The face detection module of this system employs two different face detection algorithms based on color (Ref. 5) and Harr-like features (Ref. 6) respectively. A fusion scheme that combines the two algorithms and employs additional decision criteria to improve the detection rate and reduce false detections is incorporated in order to handle as many different detection scenarios as possible. Fusion is essential, because an automatic system for face detection, especially when applied as an initialization step in a system for tracking people, should be able to cope with frontal to profile face poses, as well as different orientations. However, the computational efficiency should be high enough to allow for fast detection and not limit its applicability in real-world environments.

2.1. Color-based face detection

Using color as the primary source of information for skin detection has been a favorable choice among researchers. Consequently, there have been a number of attempts to determine the optimum color space for skin segmentation. Researchers have concluded that the skin color distribution forms a cluster (the so-called *skin locus*) in various color spaces (Ref. 8, 9), which is however, camera-specific. For a comprehensive discussion on skin color detection techniques, the reader is referred to Ref. 10.

The color-based algorithm used in this paper is similar to the one in Ref. 5. Skin segmentation in the Hue-Saturation-Value (HSV) color space, which has been popular due to its inherent relation to the human perception of color, is used. Moreover, the V component (intensity) is ignored, in order to obtain at least partial robustness against illumination changes, resulting in a 2-D color space. Instead of modelling skin color distribution using non-parametric methods, such as Lookup Tables (LUT), Bayesian classifiers or Self Organizing Maps or parametric methods (single Gaussian, mixture of Gaussians or even multiple Gaussian clusters), the system in this paper employs a skin classifier that explicitly defines the boundaries of the skin cluster in the HS(V) color space.

The input image is first converted into the HSV color space. The H, S values of all the individual pixels are tested against appropriate thresholds (the thresholds used are similar to the ones used in Ref. 5). More specifically:

$$f(h) = \begin{cases} 1 & , \quad 0 < h < 0.15 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{1}$$

and

$$g(s) = \begin{cases} 1 & , \quad 0.2 < s < 0.6 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{2}$$

with h and s values in the interval $[0, 1]$. A pixel will be classified as skin-like only if $f(h)g(s) = 1$. Such a method is attractive because of its simplicity and the ability to construct very fast classifiers. Since the detection method presented in this paper involves a combination of two detectors, it is essential that the computational burden is kept low.

The skin segmentation results are morphologically processed. Connected component analysis is the next step. The number of contour points of each connected component is tested against a threshold, to ensure that the subsequent ellipse fitting process is applied only to large enough regions. The shape for each connected component is then examined by an ellipse fitting algorithm to further reduce the number of candidate regions. The best-fit ellipse is computed using the general conic-fitting method presented in Ref. 11, with additional constraints to fit an ellipse to scattered data. Additional decision criteria (orientation of the ellipse, ratio of the ellipse axes, area occupied by the ellipse) are incorporated to ensure that invalid ellipses will not be fit. The thresholds for the criteria that have been determined by experimentation are the following: $N > 10 * scale$, $1.6 < \frac{b}{a} < 2.5$, $A > 36 * scale$, $45^o < \theta < 135^o$, where N is the number of contour points of the connected component, a and b denote the lengths of the minor and major axis of the ellipse respectively, A is the area occupied by the ellipse, θ is the angle between the horizontal axis and the major ellipse axis (i.e. the orientation of the ellipse), in degrees, and *scale* is a parameter associated with the size of the input images.

Color-based detectors suffer from false detections, due to the presence of other foreground or even background objects that exhibit similar color and shape properties with the objects of interest (e.g. faces). For this reason, the resulting candidate regions are then subjected to a facial feature extraction process to reduce false detections. The first order derivative

with respect to the vertical axis of the input image I is calculated by applying an extended Sobel operator. The resulting image J is then thresholded to produce a binary image B, according to:

$$B(i,j) = \begin{cases} 1 & , \quad J(i,j) > \overline{J(i,j)} \\ 0 & , \quad \text{otherwise} \end{cases}$$

where $\overline{J(i,j)}$ denotes the average grayscale value of all image pixels. The algorithm can correctly detect frontal faces. However, skin-like areas irrelevant to the subsequent tracking process can often be included in the detected faces (i.e. the neck of the subjects), as can be seen in Figure 2 (a). This can cause problems to the tracking module. The algorithm will fail in rare cases (e.g. if the subject wears clothes with skin-like colors, folds in the clothes can potentially confuse the detector, as illustrated in Figures 2 (b) and (c)).

2.2. Face detection based on Harr-like features

The second detector used is the frontal face detector in Ref. 6, with very good results on frontal test datasets. Exposure to real-world conditions might produce false detections, as illustrated in Figure 2(d). To overcome false detections, the algorithm is modified so as to include a color-based thresholding step, identical to the initial skin-like segmentation step of the color-based detection algorithm, as specified by (1) and (2), but applied to each face region detected instead of the whole image. Since a face in any pose or orientation should contain a large portion of skin, thresholding on the number of skin-like pixels is also employed. This eliminates any false detections associated with the background, while maintaining all correctly detected faces, as can easily be seen in Figure 2(e). The algorithm can correctly detect frontal faces, but irrelevant areas (portions of the background) might be included in the detected faces.

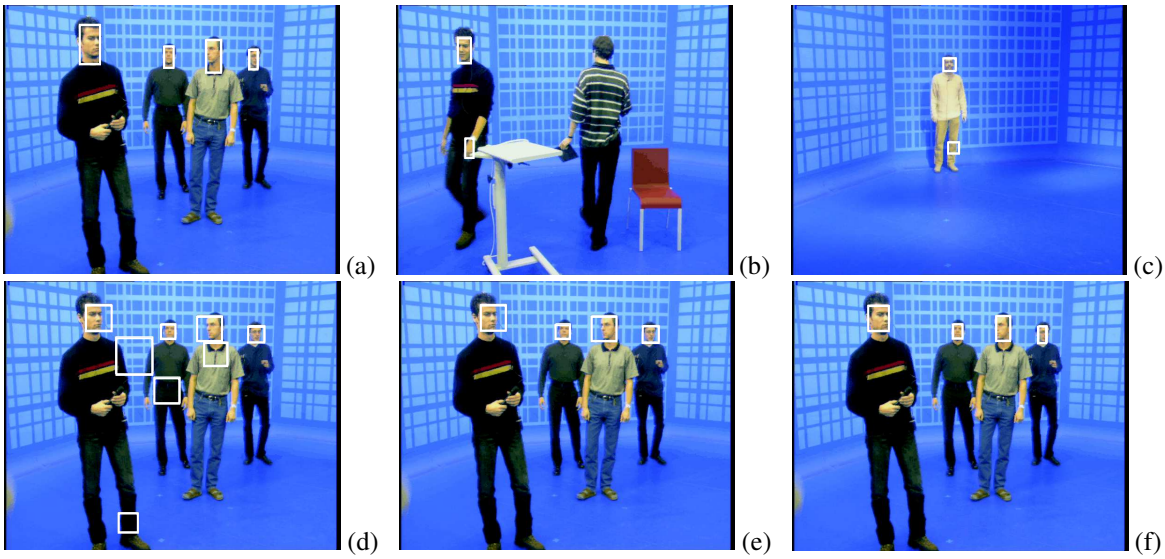

Figure 2. Face detection. (a) False detections produced by the feature based detector in Ref. 6, (b) elimination of false detections by means of a skin-like threshold, (c)-(d) false detections produced by the color-based detector, (e) erroneous detection regions (including the subject's neck), produced by the color-based detector, and (f) results of fusing the two detectors.

2.3. Fusion of color-based and feature-based detectors

The problem of detection is essentially split in two separate tasks: frontal and non-frontal face detection. The frontal case is mainly handled by the frontal face detector used in Ref. 6, modified by incorporating the color-based thresholding step described earlier. The color-based face detection scheme described earlier is responsible for detecting faces in different poses and orientations, as well as for supplementing the results of the frontal face detector. The combined algorithm proceeds as follows. Both algorithms are applied to the input image. The intersections of the frontal face regions detected by both detectors are the ones accepted as frontal faces. However, there exist cases when either of the two detectors will detect frontal faces that the other one has missed. These additional faces are also accepted. Finally, the color-based detector

is responsible for detecting faces in poses and orientations other than frontal and upright. The result of "fusing" the two detectors is illustrated in Figure 2 (f), where it can be clearly seen that original "erroneous" facial regions of both detectors that contained background or "irrelevant" pixels (Figures 2 (a) and 2 (e)) have been corrected. Results are very good, as illustrated in Figure 3 (a)-(d). A schematic description of the overall detection module is depicted in Figure 1 (b).

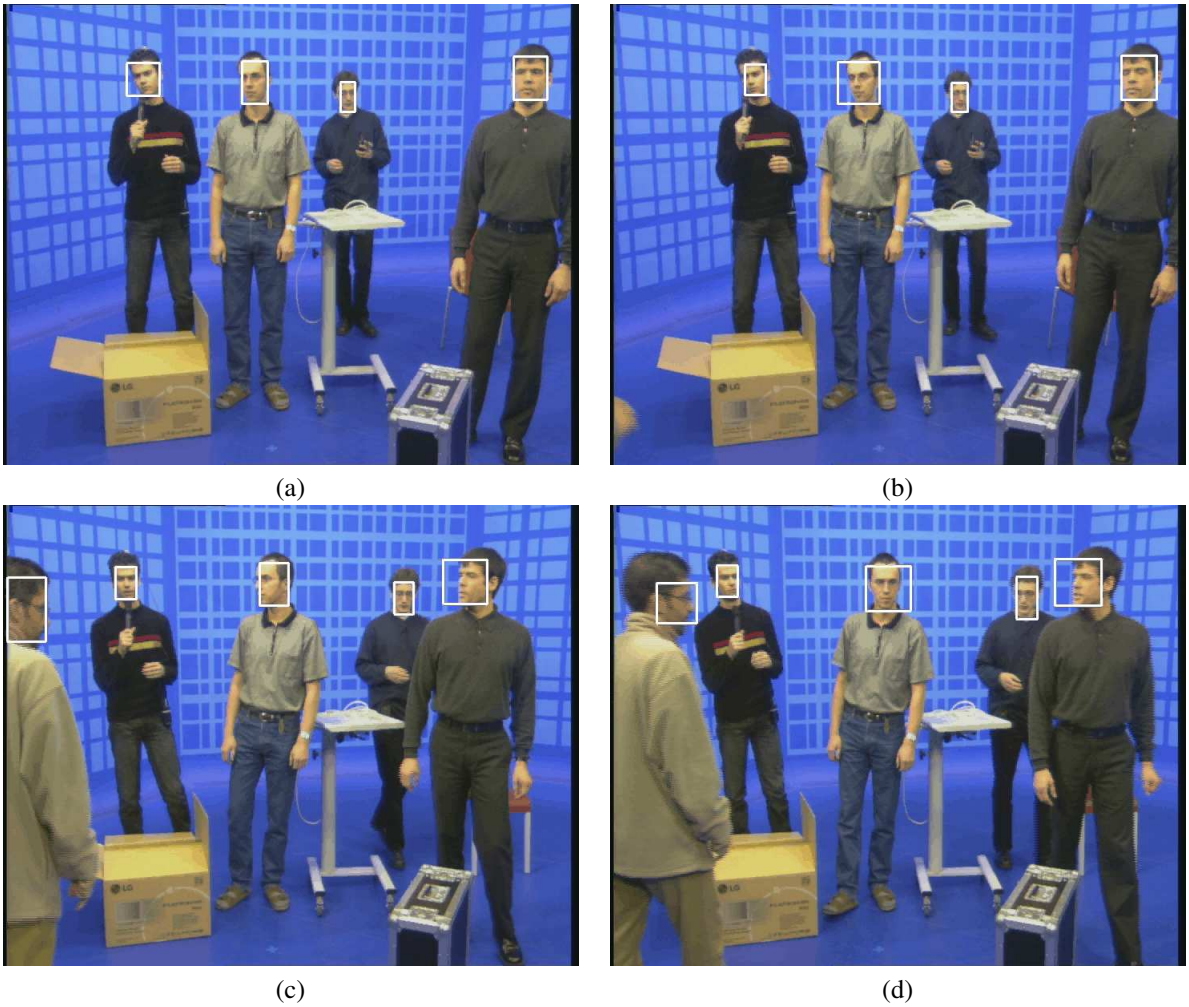

Figure 3. Correct detections produced by the fusion of two detectors in sample frames of a video sequence.

3. REGION BASED FEATURE TRACKING

The algorithm used for tracking faces (or other regions of interest) is based on selecting a large number of point features in the tracking region which are subsequently tracked in the next frames. Tracking is initialized either manually or with the output of the detection module, i.e. the bounding box(es) of the area(s) corresponding to the detected face(s). The result of the tracking algorithm is specified as the bounding rectangle of all the tracked features. Point features are tracked using the Kanade-Lucas-Tomasi (KLT) algorithm (Ref. 7). The displacement $\mathbf{d} = [d_x \;\; d_y]^T$ between two feature windows on images I and J is obtained by minimizing:

$$\varepsilon = \int\int_W [J(\mathbf{x} + \frac{\mathbf{d}}{2}) - I(\mathbf{x} - \frac{\mathbf{d}}{2})]^2 w(\mathbf{x}) d\mathbf{x} \tag{3}$$

where $\mathbf{x} = [x, y]^T$, W is the region of the window and $w(\mathbf{x})$ is a weighting function. In order to perform one iteration of the minimization procedure of (3), the equation $Z\mathbf{d} = \mathbf{e}$ must be solved, where (Ref. 7):

$$Z = \int \int_W \mathbf{g}(\mathbf{x})\mathbf{g}^T(\mathbf{x})w(\mathbf{x})d\mathbf{x} \tag{4}$$

$$\mathbf{e} = 2\int \int_W [I(\mathbf{x}) - J(\mathbf{x})]\mathbf{g}(\mathbf{x})w(\mathbf{x})d\mathbf{x} \tag{5}$$

and

$$\mathbf{g} = \begin{bmatrix} \frac{\partial(I+J)}{\partial x} \\ \frac{\partial(I+J)}{\partial y} \end{bmatrix} \tag{6}$$

To eliminate background features from the tracking process, a clustering procedure is applied (Ref. 12). Let (μ_x, μ_y), (σ_x, σ_y) be the mean and variance of the feature coordinates for all features in frame t and $[x, y]^T$ the coordinates of some feature. This feature is retained in frame t+1 if $x\epsilon[\mu_x - \sigma_x, \mu_x + \sigma_x]$, $y\epsilon[\mu_y - \sigma_y, \mu_y + \sigma_y]$, otherwise it is rejected. Assuming that the tracked object features have similar motion patterns, this enables the algorithm to reject stationary or slowly moving background features, after a number of frames. This is particularly useful if the region used for tracking initialization contains a portion of background, as can be seen in Figure 2 (e).

Feature generation is based on the algorithm used for point feature tracking (Ref. 7), where a good feature is defined as the one whose matrix Z has two large eigenvalues that do not differ by several orders of magnitude. Such a feature assures that equation $Z\mathbf{d} = \mathbf{e}$ is well conditioned. It can be shown that the large eigenvalue prerequisite implies that the partial derivatives $\frac{\partial(I+J)}{\partial x}$ and $\frac{\partial(I+J)}{\partial y}$ are large (Ref. 7).

To overcome the problem of loss of features, especially when the amount of motion between two subsequent frames is above average, the number of features in each tracked region is checked in each frame against a specified threshold. If the number falls below the threshold, features are regenerated. Feature regeneration also takes place at regular intervals, in an effort to further enhance the tracking process.

There exist cases, however, when tracking failure will occur, i.e. a face is lost in a frame. To cope with such problems, re-detection is employed using the combined face detection algorithm presented earlier. However, if any of the detected faces coincides with any of the faces already being tracked, the latter are kept, while the former are discarded from any further processing. Re-detection is also periodically applied to account for new faces entering the field-of-view of the camera. The schematic description of the tracking module is illustrated in Figure 1 (c).

4. EXPERIMENTAL RESULTS

A substantial number of test video sequences with ground truth have been used to test the algorithm presented in this paper. The sequences have been obtained within the framework of the project CARROUSO, IST-1999 20993 and postprocessed within the framework of the SIMILAR European Network of Excellence on multimodal interfaces (http://www.similar.cc), both funded by the Commission of the European Communities. They include scenes shot with different subjects, lighting conditions, motion trajectories and occlusion conditions. Ground truth data were provided by means of the output of a 4 infrared camera system located on the studio ceiling. Both the color-based and the combined face detection algorithms have been tested on a representative sample of 1239 images taken from these sequences. The images contain 1587 facial instances, in various poses, orientations and lighting conditions. In order to calculate the results, two assumptions were made: the whole face should be within the field-of-view of the camera and should be clearly visible (i.e. it should not be occluded) and the subject(s) should not present the back side of their head to the camera (i.e. at least some part of the facial skin should be visible). Examples of these images are illustrated in Figure 2.

The detection rate of the color-based algorithm is 57.9%, while the false alarm rate is 9.6%. When running the combined algorithm, the detection rate increases to 79.1%, whereas the false alarm rate drops to 3.4%. Direct comparison with the feature-based detector, presented in Section 2 would not be fair, because the latter is a frontal face detector that can handle approximately ± 15 degrees of in-plane rotation. However, a qualitative comparison reveals that the false alarm rate, when fusing the two detectors is again significantly lower, while the hit rate is comparable to that of the feature-based detector. The detection rate achieved by the combined algorithm is very satisfactory, especially if the following facts are

taken into consideration: first, detection results refer to facial instances in all possible poses and orientations and second, the computational burden is very low, since a detection scheme that fuses the results of only two simple detectors is employed.

The overall system has also been tested on the test video sequences with very good results. It is capable of processing full PAL video sequences (24-bit-color, resolution 720x576 pixels) at a frame rate of 5 frames/sec using a 2GHz Pentium IV PC with 512 MBytes of RAM. The frame rate can substantially increase (12-15 frames/sec) at the expense of accuracy if the frames are sub-sampled prior to processing or certain internal parameters of the detection algorithms are relaxed. In Figures 4 and 5, the results of automatic face detection and 2-D tracking for two of the video sequences are illustrated. Sample frames are taken at 50-frame intervals. In the first sequence, a single subject is moving parallel to the camera at a distance of 4 meters, from the left to the right, moving in and out of the field-of-view, with optimal lighting conditions and no occlusion. In the second sequence, two subjects are moving randomly, with sub-optimal lighting conditions and at times occluding each other. The different lighting conditions (optimal vs. sub-optimal) refer to two different configurations of the studio lights (where the test sequences were acquired). The first causes soft shadows, whereas the second introduces hard shadows and causes bright and dark areas to appear in the recorded video sequences, thereby hardening the task of detection and tracking.

It can be clearly seen that the system accurately tracks the face of the subject in the 700 frames of the first sequence. Additionally, the system re-detects the subject and re-initiates tracking between frames 600 and 650, as seen in Figure 4 (l) and (m), when the subject moves out of the field-of-view of the camera and later re-enters the scene. This is accomplished through the re-detection stage applied when one of the tracked faces is lost. The re-detection process is illustrated more clearly in Figure 5, because the two subjects are moving randomly in the second sequence and occlusion takes place quite often. The system initially does not detect the second subject, because an inadequate portion of its facial skin is visible. It therefore does not track the second subject until frame 100, depicted in Figure 5 (c), when the subject is detected for the first time. This is due to the fact that the system applies the detection algorithm periodically to account for new faces entering the scene or for faces that were not detected at earlier stages, as in this case. The re-detection period is 100 frames. The two subjects are successfully tracked until frame 300, Figure 5 (g). The system then loses track of the first subject (the taller actor) and can not re-detect him, because he is facing away from the camera. The subject is re-detected later, Figure 5 (i), by which time, the second subject is lost again and re-detected later in the sequence. Both subjects are accurately tracked for the subsequent 200 frames, Figure 5 (j)-(n). The second subject leaves the field-of view of the camera, Figure 5 (o), and re-enters later. The system again re-detects him, Figure 5 (p), and successfully tracks both subjects for the remainder of the sequence.

5. CONCLUSION

In this paper, a complete system for tracking people was presented. The system can operate on either live camera feed or pre-recorded video sequences. Initialization can be automatic, in which case a detection algorithm that is based on fusion of two detectors, based on color and Harr-like features respectively, is employed. The combined algorithm is capable of handling different face orientations and poses (frontal, profile, intermediate). To avoid false detections, a number of decision criteria are employed. Tracking is performed using a variant of the well-known Kanade-Lucas-Tomasi tracker. Manual intervention is allowed to assist both modules if required, while occlusion is handled through a re-detection stage.

Future work could be focused on achieving even higher detection rates and lower false alarm rates for the detection module and on further enhancing the robustness of the tracking module against total occlusions. Increasing the overall speed (frame rate) of the system, without sacrificing accuracy, would also be advantageous.

ACKNOWLEDGMENTS

The work presented was developed within VISNET, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme.

REFERENCES

1. M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(1), pp. 34–58, 2002.

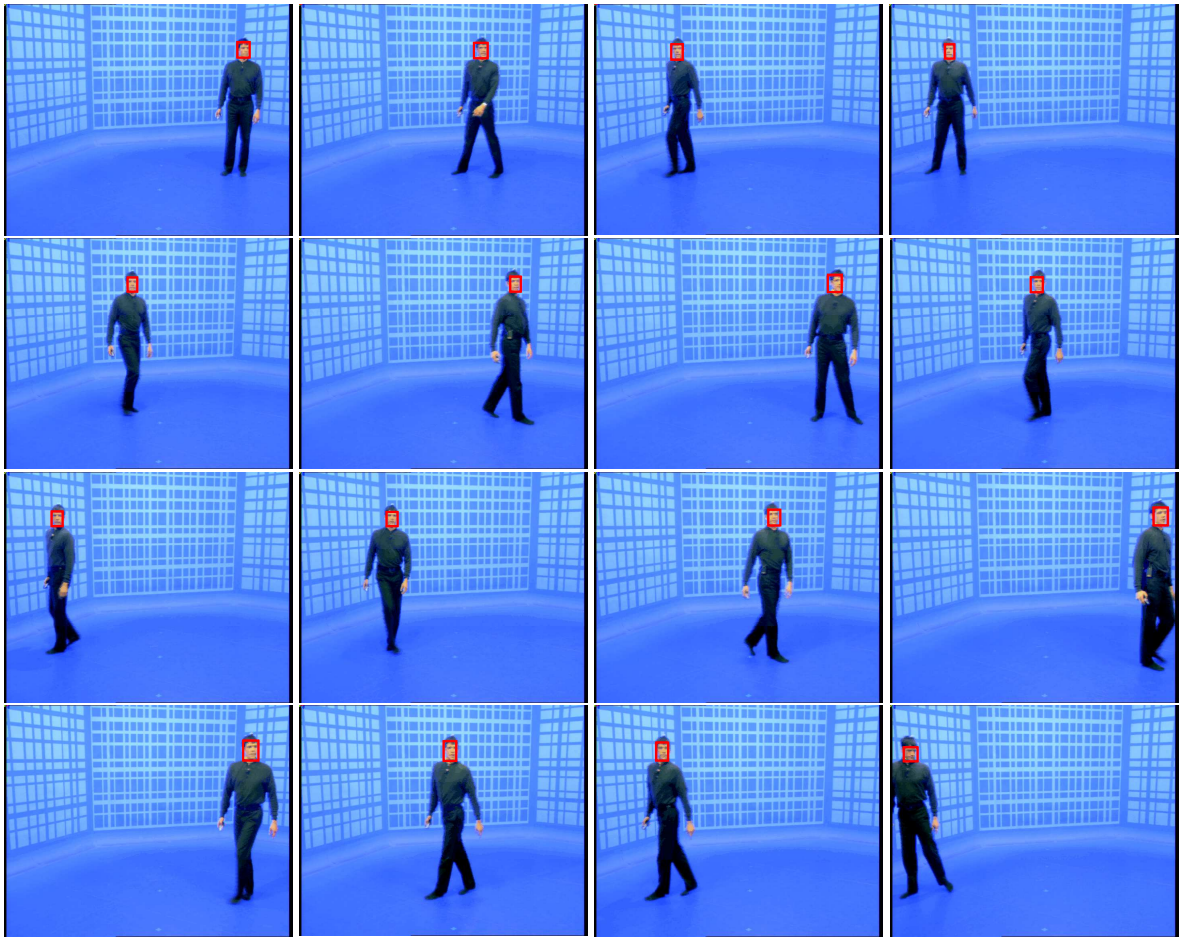

Figure 4. 2-D tracking results on the first test video sequence, 750 frames, sample frames displayed at 50-frame intervals (order: top-to-bottom, left-to-right).

2. G. Welch and E. Foxlin, "Motion tracking: No silver bullet, but a respectable arsenal," *IEEE Computer Graphics and Applications, special issue on "Tracking"* **22**, pp. 24–38, November/December 2002.
3. T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding* **81**, pp. 231–268, 2001.
4. D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding* **73**(1), pp. 82–98, 1999.
5. K. Sobottka and I. Pitas, "Looking for faces and facial features in color images," *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, Russian Academy of Sciences* **7**(1), pp. 124–137, 1997.
6. R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *IEEE International Conference on Image Processing (ICIP02)*, pp. 900–903, (Rochester, New York, USA), September 2002.
7. J. Shi and C. Tomasi, "Good features to track.," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR94)*, pp. 593–600, (Seattle, United States), June 1994.
8. B. D. Zarit, B. J. Super, and F. K. H. Quek, "Comparison of five color models in skin pixel classification," in *ICCV99 International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS99)*, pp. 58–63, (Corfu, Greece), September 1999.
9. B. Martinkauppi, M. Soriano, and M. Laaksonen, "Behavior of skin color under varying illumination seen by different cameras in different color spaces," in *Machine Vision Applications in Industrial Inspection IX, Martin Hunt, Editor Proceedings of SPIE*, **4301**, pp. 102–112, (Coimbra, Portugal), July 1999.

Figure 5. 2-D tracking results on the second test video sequence, 700 frames, sample frames displayed at frame 0, 30, 62 and then at 50-frame intervals (order: top-to-bottom, left-to-right).

10. V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *International Conference on Computer Graphics between Europe and Asia (GRAPHICON-2003)*, (Moscow, Russia), September 2003.
11. A. Fitzgibbon and R. Fisher, "A buyers guide to conic fitting," in *Fifth British Machine Vision Conference (BMVC99)*, pp. 513–522, (Birmingham, UK), 1995.
12. E. Loutas, K. Diamantaras, and I. Pitas, "Occlusion resistant object tracking," in *IEEE International Conference on Image Processing (ICIP01)*, **2**, pp. 65–68, (Thessaloniki, Greece), October 2001.