A MONOCULAR SYSTEM FOR PERSON TRACKING: IMPLEMENTATION AND TESTING

Georgios N. Stamou, Michail Krinidis, Nikos Nikolaidis and Ioannis Pitas

Aristotle University of Thessaloniki, Greece
{gstamou;mkrinidi;nikolaidis;pitas}@aiia.csd.auth.gr

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. The system caters for calibrated cameras and can provide 3-D coordinates of any tracked object(s) of interest. It has been tested with very good results on a variety of video sequences, including a database of studio video sequences, for which 3-D ground truth data, originating from a 4-camera infrared tracking system, exist.

KEYWORDS

Object tracking – Face Analysis – Face detection – Fusion – Camera Calibration – Multimodal Interfaces

1. INTRODUCTION

Tracking the motion of people in video sequences has been a topic of active and intense research for the past two decades. Such a task is usually preceded by an initialization step that aims at detecting the presence of people.

Detecting people has been tackled with using a range of methods. Various methods aim at recovering the human figure, i.e. the silhouette [1], [2], whereas others focus on recovering the position of certain parts of the human body, e.g. arms, hands, limbs etc. [3], [4]. A large number of attempts focus on face detection, due to its obvious importance as a pre-processing step in applications such as intelligent human-computer interfaces, content-based image retrieval, surveillance, video coding, face recognition, face authentication, pose estimation etc. A number of factors that include pose variations (frontal, profile and intermediate poses), skin-color variations, facial structural components, such as moustache, beards and glasses, occlusion and poor or variable imaging conditions make this task a rather difficult one. These factors essentially force researchers to make a number of assumptions that enable them to successfully handle the task in hand, but at the same time limit the application scope of such algorithms. Face detection methods can be classified into a number of different categories, ranging from knowledge-based methods, aiming mainly at face localization, to appearance-based methods, where models are learned from a set of representative training images and used for face detection. For details on these methods, the reader is referred to [5], [6].

Video-based tracking of the motion of the human body, either viewed as a single object or comprising an articulated structure consisting of a number of rigid body parts, is also a challenging research topic with applications in many domains such as human-computer interaction, surveillance, hand gesture recognition and 3-D reconstruction. There exist alternative approaches, which could be divided into two broad classes, active and passive tracking. Active trackers employ wearable devices, which simplify further processing and are mainly suitable for well-controlled environments. Passive trackers, on the other hand, use at the most simple markers attached to the subject(s) or no such devices at all. 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, free of special markers or devices, able to function in uncontrolled (indoor or outdoor) environments. However, a number of difficulties arise, including but not limited to projection ambiguities, computational burden, self occlusion, unconstrained motion, clutter, poor or varying lighting conditions, use of a single camera etc. These difficulties have led researchers to adopt a number of assumptions in order to focus on tackling specific aspects of an overall very complex problem. Assumptions can be either related to the motion of the camera or subjects (e.g. fixed camera, single-person scenes, occlusion-free scenes, known motion models, front-to-parallel movement with respect to the camera) or refer to the appearance of the environment (constant lighting conditions, uniform background etc.) or the subject(s) (known initial position, tight clothing etc.). For a comprehensive review of different methods, the reader is referred to [7] and earlier surveys [8, 9, 10].

It is obvious that building tracking systems that can be used in real-world environments is far from being a simple process. In [11], a real-time tracking system was built and tested on a variety of physical locations, without any special devices. The proposed system uses simple 2-D models of the human body to perform detection and tracking, as well as a priori knowledge to recover from failures. However, the system performance deteriorates when specific assumptions adopted by the authors, namely that the background is much less dynamic than the subject and that there is only one subject within the camera field-of-view, do not hold. [12] presents a system capable of detecting and tracking multiple people in the context of video-conferencing. The system continuously applies a neural-network-based face detector to account for new subjects entering the field-of-view of the camera or to handle occlusion and tracks the detected faces. Limitations include detecting portions of the background as valid faces, as well as loss of valid faces when the regions corresponding to faces merge with regions of falsely detected faces, due to proximity. In [13], a real-time system for face and facial feature detection and tracking in video sequences was proposed. The system can automatically detect and subsequently track up to four faces in specific orientations only. In [14], a face tracker similar to the automatic tracker used in this paper was integrated into a video encoding environment, in an attempt to allocate more coding bits to the face regions of interest, with the use of additional cues to lower the number of false alarms. However, in scenes with multiple faces or in scenes where faces move quickly in and out of the field-of-view, the authors decided that the best strategy is not to track such faces at all, which limits the applicability of such a system in more generic environments. In [15], a real-time visual surveillance system for detecting and tracking multiple people and monitoring their activities in an outdoor environment was implemented. The system employs shape analysis and tracking to locate certain parts of the human body (head, hands, feet, torso) and can track multiple people. Additionally, it can detect and track objects other than people and consequently monitor interactions between people and such objects. In [16] the image intensity was represented by a $3D$ deformable surface model. The system relies on selecting and tracking feature points by exploiting a feature vector which is an intermediate step of the deformation governing equations. This vector is proven to be a combination of the output of various line and edge detection masks, thus leading to distinct, robust features.

The goal of this work is to present an automatic/semi-automatic system, that integrates enhanced versions of a number of different algorithms originating from different fields of computer vision and aims at robust face detection and tracking, as well as object tracking in general, covering both 2-D and 3-D cases, with performance comparable to expensive commercial tracking systems that utilize special devices for tracking. Our approach for face detection was motivated by [17] and [18] 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 with respect to failures. The first detector is based on color whereas the second employs the so-called Harr like features. The tracking algorithm of this system is a variant of the Kanade-Lucas-Tomasi tracker [19], [20], which can successfully deal with still or slowly moving features and large displacements of features between consecutive frames, in order to make the tracking process more robust to occlusions. The proposed system can operate in 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 its default configuration, the system can cope with a range of different environments. However, a number of parameters can be fine-tuned to produce even better results.

One of the novel contributions of this paper is a fusion scheme that combines the results of the two separate detectors, aiming at reliable detection of faces in various poses (frontal, profile, intermediate) and orientations. However, the main contribution 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. Tracking using a calibrated camera is described in Section 4. A brief description of the main features of this system can be found in Section 5. Section 6 presents experimental results, while in Section 7 the final conclusions are drawn.

2. FACE DETECTION BASED ON FUSION OF INFORMATION

In this Section, two different face detection algorithms based on color and Harr-like features are described. Their strengths and weaknesses are identified and the two approaches are shown to be complementing each other. A fusion scheme that combines the two algorithms and employs additional decision criteria to improve the detection rate and reduce false detections is derived. 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.

A number of detection methods use color to perform skin segmentation and then post-process the segmentation results, to compensate for any errors originating from substantial changes in foreground and background lighting. This process typically involves a connected component analysis step, followed by shape analysis and detection of multiple facial features in each connected component [17], [21]. The facial features used in such methods include the eyes, eyebrows, hair, nose and the mouth. Symmetry of the human face [22], as well as biometric information, i.e. distances between facial features [23] have also been exploited.

Another category of methods ignore color information and search for salient features by means of edge detectors [24], [25]. Such features, however, can be easily corrupted by noise and illumination changes. Other researchers [26], [27], [28] use standard face patterns as templates and evaluate the correlation between a new image and the pattern images for a number of different features (eyes, nose, mouth, face contour). Limitations in scale, pose and shape variations can be overcome at the expense of increased computational burden. All the above mentioned methods share a common characteristic, that is, they are mainly focused on frontal face detection. A limited number of attempts to build profile or non-frontal face detectors are reported [29], [30], [31], [32].

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 [33], [34], which is however, camera-specific. For a more comprehensive discussion on skin color detection techniques, the reader is also referred to [35].

The color-based algorithm that we have used is similar to the one in [17]. Skin segmentation in the Hue-Saturation-Value (HSV) color space, which has been a popular choice among researchers 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 HSV 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 [17]). More specifically:

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

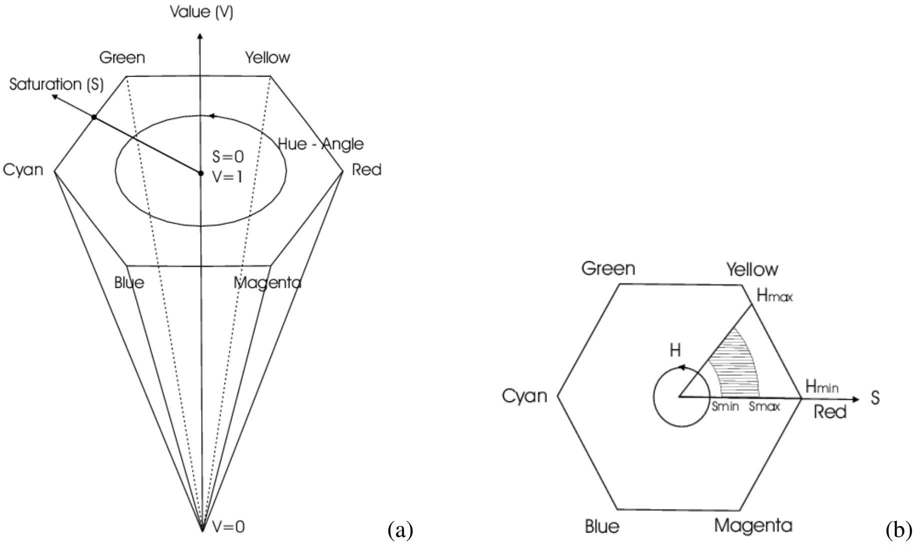


Figure 1: *(a) Hue-Saturation-Value color space, (b) Thresholding in the Hue-Saturation space*

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 $f(h)g(s) = 1$. Since the HSV color space has a hexcone shape, as illustrated in Figure 1-a, this is equivalent to cutting a sector out of the hexagon, as seen in Figure 1-b [17]. 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 by means of a number of opening and closing operations. 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. In [17], the best-fit ellipse was computed using moments. Our algorithm uses the general conic-fitting method presented in [36], with additional constraints to fit an ellipse to scattered data. Additional decision criteria are incorporated to ensure that invalid ellipses will not be fit. These criteria refer to the orientation of the ellipse, the ratio of the ellipse axes and the area occupied by the ellipse. The thresholds for the criteria have been determined by experimentation and 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. However, it is important to note that the same threshold values have been used in all the experiments presented within this paper.

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 candidate regions that survive this process are then subjected to a facial feature extraction process. The first step of this process is to calculate the first order derivative with respect to the vertical axis of the input image I, by first converting to grayscale and then applying an extended Sobel operator. The kernel used is:

$$\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

The resulting image J is then thresholded to produce a binary image B, according to:

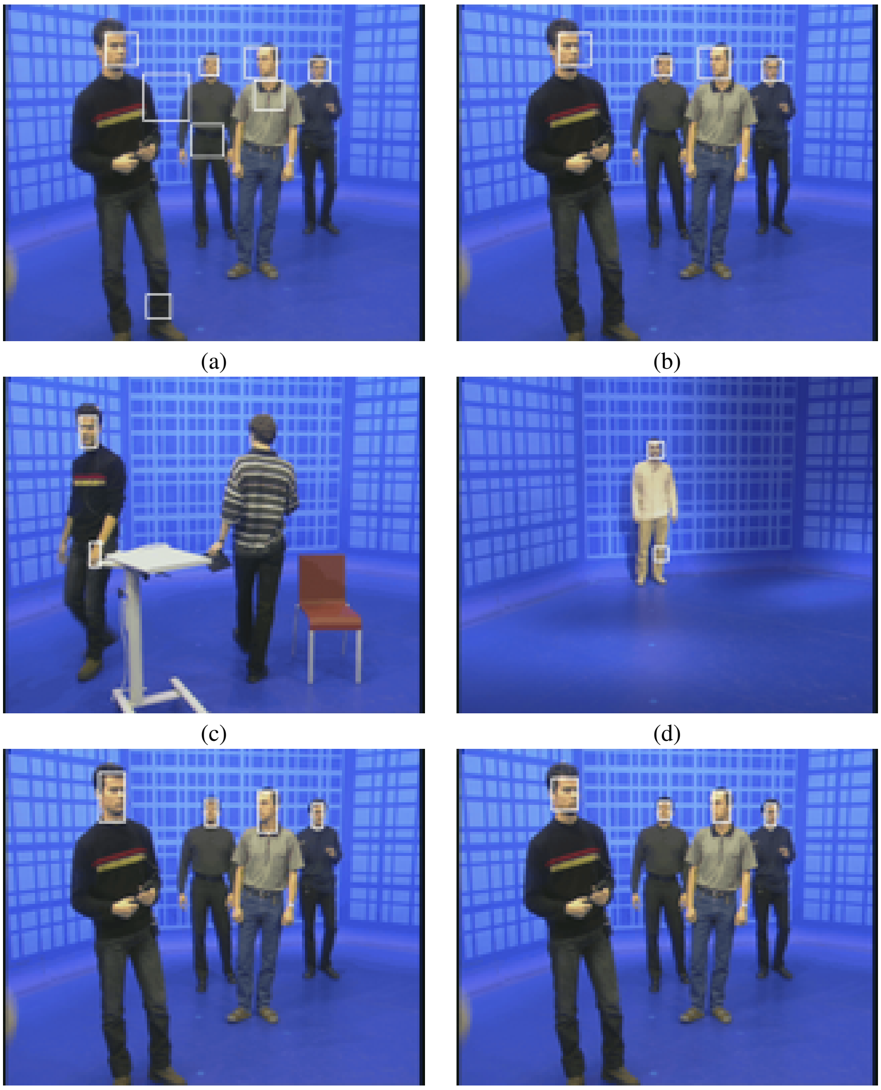
$$B(i,j) = \begin{cases} 1 & , \quad J(i,j) > \overline{J(i,j)} \\ 0 & , \quad \text{otherwise} \end{cases}$$

where $\overline{J(i,j)}$ denotes the average grayscale value of all image pixels.

The goal is to ensure that the existence of strong edges around the eyes will eliminate any falsely detected regions. It appears that this process achieves good results and fails only in rare cases. For instance, if the subject wears clothes with colors similar to the color of the human skin, folds in the clothes can potentially confuse the detector, as illustrated in Figures 2-c and -d. It should also be noted that the color-based detection algorithm will often detect face regions which include skin-like areas irrelevant to the subsequent tracking process (i.e. the neck), as can be seen in Figure 2-e. These areas will be eliminated by applying the second face detection algorithm, which is presented in Section 2.2 and fusing the results of the two detectors, as illustrated in Section 2.3.

2.2. Face detection based on Harr-like features

In [18], a real-time frontal face detection framework was proposed, based on simple features that are reminiscent of Harr basis functions. These features were extended in [37] to further reduce the number of false alarms. Although this is a fast and efficient frontal face detector with very good published results



(a) (b) (c) (d) (e) (f)

Figure 2: *Face detection in complex backgrounds. (a) False detections produced by the detector in [37], (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.*

on test datasets (namely the MIT+CMU set), exposure to real-world conditions can drastically deteriorate its performance. An example of false detections on images that contain complex or misleading background is illustrated in Figure 2-a. Another interesting fact, which is also illustrated in the same figure, is that when the face to be detected is not "up-frontal" , the detector will usually include a portion of the background in the corresponding bounding box. Such results, however, will cause problems if they are subsequently used as input to the tracking module. To overcome the first problem, namely the false detections, the algorithm is modified so as to include a color-based thresholding step. This step is identical to the initial skin-like segmentation step of the color-based detection algorithm, as specified by (1) and (2). However, it is not applied to the whole input image but to each face region detected by the original algorithm. Based on the fact that a face, whether frontal, profile or in any intermediate pose or orientation should contain a large portion of skin-like pixels, thresholding on the number of skin-like pixels is also employed. A candidate face region will be a valid face only if $S/T > \frac{1}{4}$, where S denotes the number of skin-like pixels according to the above thresholding procedure and T the total number of pixels contained in the candidate region. The choice of a small threshold value is related to the fact that the detector in question produces results that contain portions of the background. This value eliminates any false detections associated with the background, while maintaining all correctly detected faces, as can easily be seen in Figure 2-b.

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

The proposed system incorporates a combination of the two detectors presented above, in order to handle as many different detection scenarios as possible. Using this dual approach, 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 [37], modified by incorporating the color-based thresholding step described earlier. The color-based face detection scheme described in Section 2.1 is responsible for detecting faces in different poses and orientations, as well as for supplementing the results of the frontal face

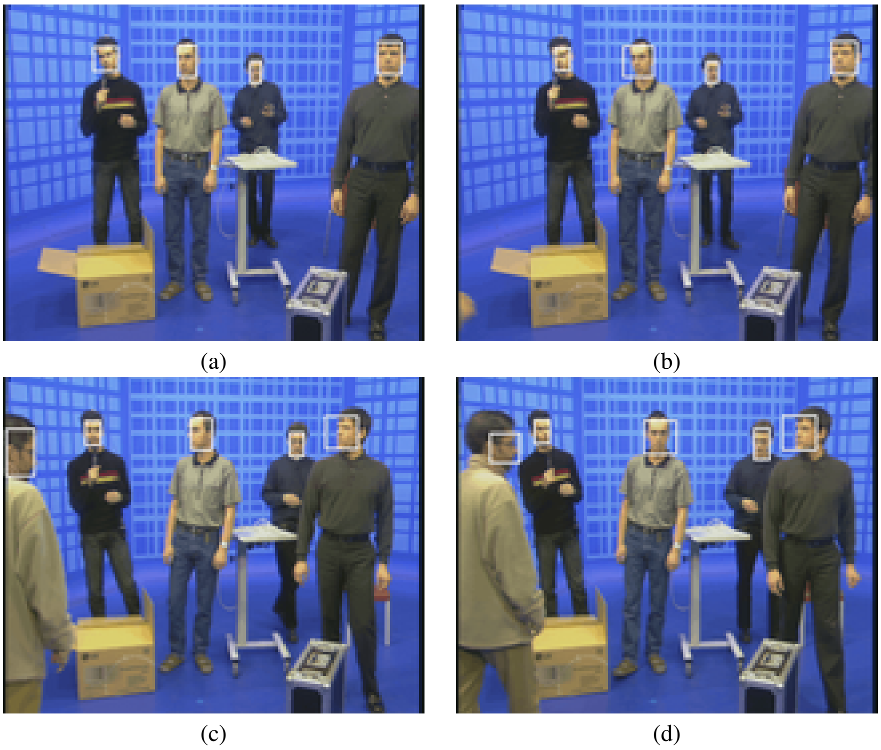


Figure 3: *Correct detections produced by the fusion of two detectors in frame (a) 10, (b) 15, (c) 45 and (d) 58 of a video sequence with complex background and additional foreground items.*

detector.

The combined algorithm proceeds as follows. Both algorithms are applied to the input image. The first detector, presented in Section 2.1 correctly detects frontal faces. However, portions of the background are included in the resulting bounding boxes, as illustrated in Figure 2-a. The second detector, described in Section 2.2 also detects frontal faces, including skin-like areas irrelevant to the subsequent tracking process (i.e. the neck), as can be seen in Figure 2-e. 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 faces that the other one has failed to do so. These additional faces are also accepted. More specifically, the color-based algorithm detects some frontal faces that the first detector can not handle, mainly because of restrictions in the minimum size of detected faces. 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 the first and second detectors that contained background pixels have been corrected. Results are very good, as illustrated in Figure 3. A schematic description of the overall detection module is depicted in Figure 4-b.

3. REGION BASED FEATURE TRACKING

In this Section, the algorithm used for tracking faces (or other regions of interest) is presented. The algorithm 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 [19], [20]. The displacement $\mathbf{d} = [d_x \quad d_y]^T$ 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 [19], [20]:

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

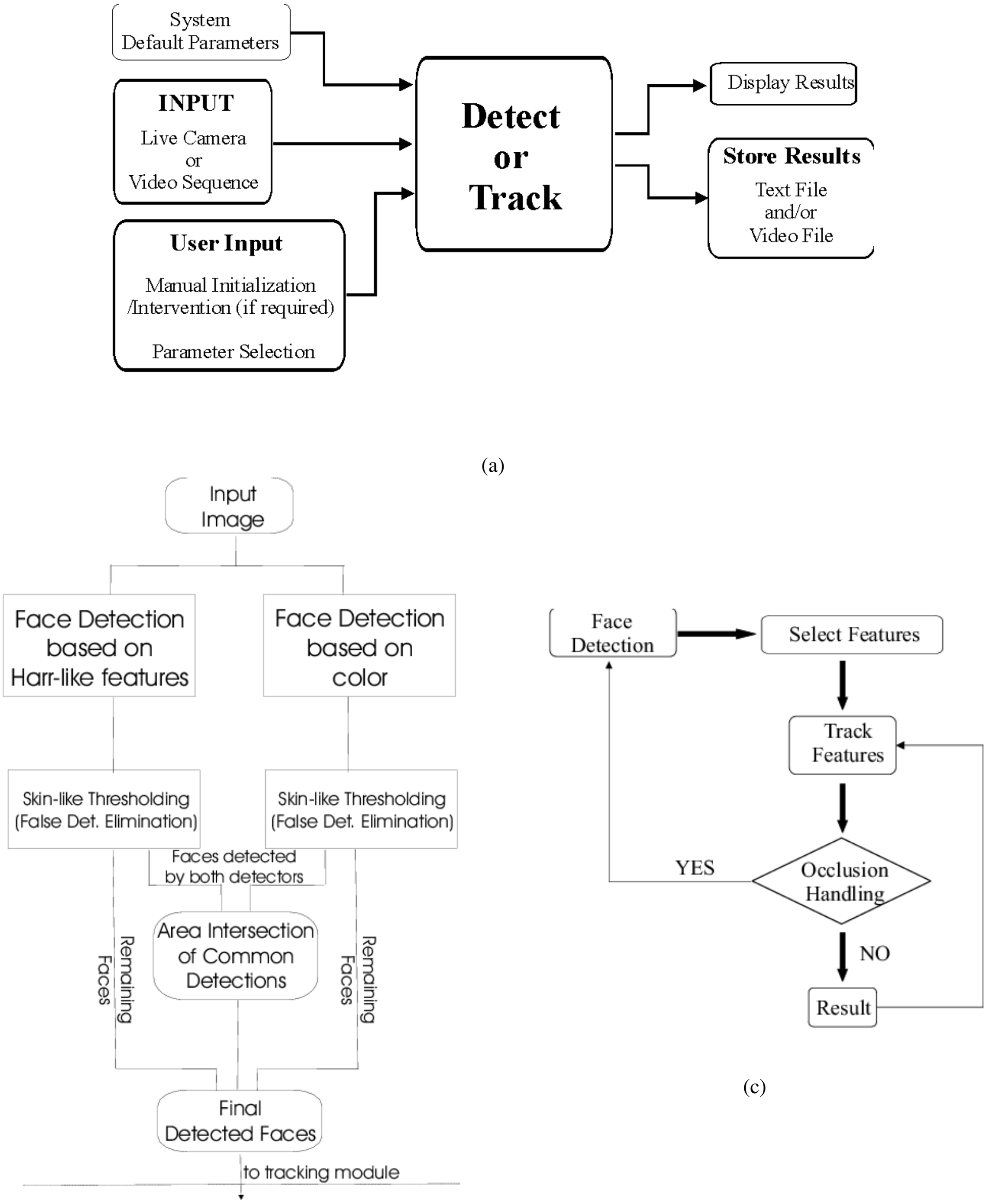
$$\mathbf{e} = 2\int\int_W [I(\mathbf{x}) - J(\mathbf{x})]\mathbf{g}(\mathbf{x}) w(\mathbf{x}) d\mathbf{x} \tag{5}$$

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 [38]. 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 3-b to -d.

Feature generation is based on the algorithm used for point feature tracking [19], [20], 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



(a)

(b)

(c)

Figure 4: *Schematic Diagrams: (a) Overall system, (b) Detection module and (c) Tracking module*

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 [19], [20].

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 (i.e. existing features are kept, while new features are generated inside the region until the specified number of features is reached). 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 4-c.

4. TRACKING USING A CALIBRATED CAMERA

In order to be able to extract the 3-D coordinates of the tracked object of interest (e.g. a face), camera calibration is necessary. The latter is performed using the method described in [39]. A planar chessboard pattern is observed by the camera in a number of different orientations. The method yields better results when an average of 20-25 different orientations are captured. Different orientations can be obtained by either moving the camera or the pattern, without explicit knowledge of the actual motion. Radial lens distortion is also accounted for. One of the advantages of the method is the fact that a simple pattern and an easy to perform calibration procedure are used. Additionally, this method is more robust than self-calibration methods [40], [41].

In the case that the camera used is calibrated with the method described above, the 2-D (image) coordinates of the center of gravity of all tracked features within a region of interest are used to obtain the 3-D (world) coordinates of the tracked object(s) in each frame.

The calibration process essentially provides for each pixel a line connecting it to the camera center of projection, where the point projected on this pixel can lie (i.e. *projection line*). In order to exactly localize this point, i.e. calculate the world coordinates of this point using its 2-D image coordinates, an additional constraint among the world coordinates must be introduced [42]. In our case, we assume that the center of gravity of the object being tracked, is constrained to move on a plane, the equation of which is a priori known. Thus, the position of the point in the 3-D space is defined as the intersection of the projection line with the plane. The accuracy of each of the calculated 3-D coordinates depends on the angle between the projection line and the corresponding axis. In fact, the error of the calculated 3-D coordinates is inversely proportional to the angle. In other words, the deviation of the calculated 3-D coordinates from the actual ones increases as this angle approaches 0^o.

5. OVERALL SYSTEM DESCRIPTION

In this Section, the main features of the system are illustrated. The system is parameterized, which means that it can be fine-tuned for the environment it is supposed to operate on. The overall system diagram is depicted in Figure 4 (a).

Tracking can be performed in two different modes. The first one is a semi-automatic mode, where the detection process can be either automatic or manual, while the tracking stage is fully automatic. However, the user may intervene and manually correct the tracking (or the detection) results. The second one is fully automatic in both initialization and tracking, in the sense that the user need not intervene at any point while processing takes place.

In the first mode, if manual initialization is selected, user intervention is required to initialize the regions to be tracked in the first frame of the video sequence. Features are generated within each of the regions according to the algorithm in [19], [20]. These are the features used in the automatic tracking stage. Manual intervention can also take place under other circumstances, such as:

- initialization of the tracking algorithm for new faces entering the scene
- re-initialization if any of the tracked faces is lost
- correction of erroneous tracking results

Correction of erroneous results includes stopping the tracking of wrongly detected objects as well as correcting the tracked region, so as not to contain portions of the background. It is obvious that, using the manual initialization option, the system can be used to track any object(s) of interest, other than faces, in a video sequence.

In the second mode of operation, the user does not interact with the system during the processing of the video sequence. Both the detection and the tracking stages are performed automatically by the system, using the corresponding algorithms described earlier.

As already mentioned, the system can operate both on live input from a camera connected to a PC or on pre-recorded video sequences. If calibration data are available, they can be fed into the tracking system. The system provides for storing the tracking results, namely, video files depicting the bounding boxes of the objects being tracked, overlaid on the original video, and/or a text file containing the 2-D/3-D coordinates of the objects being tracked.

A tracking system should definitely cater for manual setup that would allow it to perform optimally in a number of different environments. Hence, a number of user-specified parameters in the system presented in this work can be properly adjusted to fine-tune its performance. It should be stressed, though, that the default parameters used are the ones that would allow the system to operate efficiently in a variety of different environments. Adjustments can, under certain circumstances, produce better results.

As far as detection is concerned, the following parameters can be set for the color-based detection algorithm (used also in the combined algorithm):

- the minimum and maximum hue/saturation thresholds used to initially segment the image
- the level of pre-processing, i.e. the number of times morphological operations will be successively applied to the initially segmented regions
- minimum and maximum acceptable values of the ratio associated with the axes of the ellipse
- minimum and maximum acceptable values of the angle between the major axis of the ellipses that have been fit and the horizontal axis
- percentage of the pixels that have been classified as skin-like to the total number of pixels contained in an initially detected region.

The parameters for the tracking module consist of:

- the maximum number of features that will be generated (if possible) in each region (either manually initialized or automatically detected) by the feature selection process
- re-detection interval (in frames) in order to cater for new faces entering the scene
- automatic feature regeneration interval (in frames), in order to compensate for lost features
- the threshold of the ratio of the actual tracked features to the maximum number of features that will invoke automatic feature regeneration within the tracked regions.

Additionally, the algorithm that will be used for detection, i.e. color-based detection, detection as in [18] or the combined algorithm that "fuses" the two detectors, can be specified.

6. EXPERIMENTAL RESULTS

A number of different tracking performance evaluation methods have been proposed, as in [43]. The ideal case involves direct comparison of the output of a tracking system against reference or ground truth data. Video sequences with ground truth

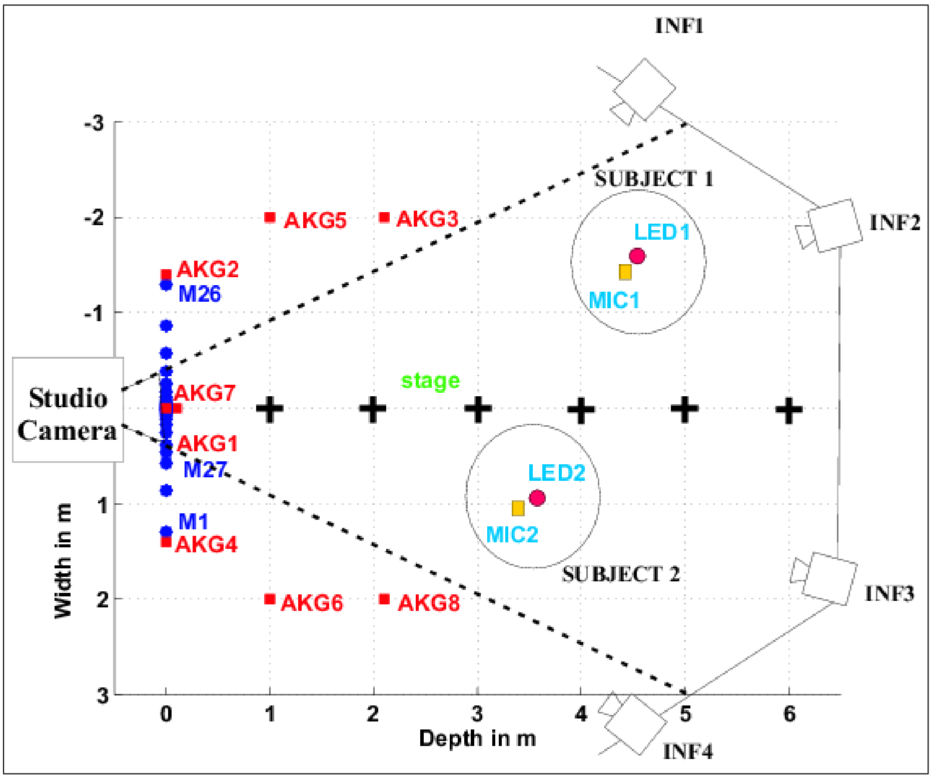


Figure 5: *The Virtual Studio.*

were available for the purpose of this study [44]. The sequences were obtained in the Virtual Studio of the Technical University of Ilmenau in Germany, as part of the CARROUSO ("Creating, Assessing and Rendering in Real Time of High Quality Audio-Visual Environments in MPEG-4 Context" [45]) European Research Project. The sequences included scenes shot with different subjects, lighting conditions, motion trajectories and occlusion conditions. Along with the video and audio data associated with the scenes, ground truth data were provided by means of the output of a 4 infrared camera system located on the studio ceiling. However, ground truth was available for a maximum of two subjects within the studio, namely those carrying special mobile infrared transmitters placed on their heads. Figure 5 illustrates the Virtual Studio. The dashed lines indicate the camera's field of view, INF1 . . . INF4 denote the infrared cameras and LED1, LED2 denote the mobile transmitters. AKG1. . . AKG8 denote fixed position microphones, while MIC1 and MIC2 correspond to two close-talking microphones, depending on whether one or two acoustic sources were present on stage. Visual aids are denoted by (+).

6.1. Face Detection

Both the color-based and the combined face detection algorithms have been tested on a representative sample of 1239 images taken from the above mentioned 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 Figures 2 and 3.

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%. A substantial 37.1% increase and a simultaneous 65% decrease in the detection and the false alarm rates respectively is achieved by the use of the introduced combined algorithm. Direct comparison with the feature-based detector, presented in Section 2.2 would not be accurate, 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 of the combined algorithm is indeed very high. This can become more evident if one considers the following facts: 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 separate detectors is employed. This is in contrast with previous works on multi-view detection, which either involve running multiple detectors on input images, each responsible for handling different views (poses-orientations), or applying a pose estimation stage prior to the detection stage [29], hence producing a substantial increase in the computational overhead. Additionally, results from other published methods refer to facial instances that correspond to a subset of all possible poses and orientations. Finally, the images used for testing these methods can not always be considered as real-world examples, since these images are usually acquired under specific conditions (e.g. constant lighting, uniform background etc.).

6.2. Tracking

The overall system has been tested on various video sequences with good results. Table 1 illustrates the default values of the parameters associated with the detection and tracking modules of the system, which were also used in these tests. The system 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. It should be noted, though, that 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.

Table 1: *System default parameter values. These values were used to produce the results presented in this paper.*

Parameter	Value
Hue	0(min) - 0.16(max)
Saturation	0.20(min) - 0.60(max)
Pre-processing level	2
Percentage of skin-like pixels	0.25
No. of contour points	40
Ellipse ratio	1.6(min) - 2.5(max)
Ellipse orientation (degrees)	45(min) - 135(max)
Ellipse area (pixels)	144
No. of features	150
No. of frames (re-detection)	100
No. of frames (regeneration)	10
Regeneration threshold (percent)	0.25

Test video sequences include the CARROUSO project video sequences [44]. In Figures 6, 7 and 8, the results of automatic face detection and 2-D tracking for three of the CARROUSO project video sequences are illustrated. Sample frames are taken at 50-frame intervals, except for the third sequence, where frames are sampled in such a way as to illustrate the ability of the system to recover from tracking failures. 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. Optimal lighting conditions refer to the fact that the studio lights were configured to produce a uniform lighting, with no dark spots or strong shadows introduced into the scene. The second sequence involves two subjects moving randomly, with optimal lighting conditions and at times occluding each other. In the third sequence, the two subjects are moving randomly, with sub-optimal lighting conditions and at times occluding each other. Sub-optimal lighting conditions refer to the fact that the studio lights were configured in such a way as to introduce dark spots or strong shadows into the scene, thus making both detection and tracking of the two subjects more difficult than in the second sequence.

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 6 (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 7, 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 7-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, as illustrated in Table 1. The two subjects are successfully tracked until frame 300, Figure 7-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 7-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 7-j to -n. The second subject leaves the field-of view of the camera, Figure 7-o, and re-enters later. The system again re-detects him, Figure 7-p, and successfully tracks both subjects for the remainder of the sequence.

Finally, the difficulties introduced by sub-optimal lighting conditions are illustrated clearly in Figure 8, where one can see that the tracking results are not as accurate as those of Figure 7. The system successfully detects the faces of the two actors initially and tracks them until frame 30, Figure 8-b, at which point occlusion is about to occur. This causes the system to lose track of the second subject (the shorter actor) and to re-detect him in subsequent frames, as illustrated in Figure 8-c. Both actors are then correctly tracked until frame 150, depicted in Figure 8-e. Occlusion occurs again, but the system later re-detects the occluded shorter subject, as seen in Figure 8-f. Tracking is subsequently performed without any problems, until frame 550, Figure 8-n. The occlusion that follows causes the system to re-detect only the first subject (the taller actor), while the second subject is detected later, when enough portion of its facial skin is visible to the camera, Figure 8-p. Both subjects are then successfully tracked for the remainder of the sequence.

In Figure 9, the coordinate system is illustrated, while in Figures 10 and 11, the results of 3-D tracking, i.e. the X and Z coordinates of the tracked object are plotted for the first two of the above three sequences, depicted in Figures 6 and 7, along with the ground truth associated with these sequences. However, it should be noted that the ground truth data provided do exhibit errors in the form of discontinuities and erratic trajectories, as illustrated in Figures 10 and 11. This is mainly due to the fact that even with 4 infrared cameras, the link between the mobile transmitter and any one of the cameras is at times lost. Additionally, the local movement of the head, where the infrared transmitters were located, can produce results that look erroneous. Even so, the ground truth data available can be used to provide valuable information about the performance of a tracking system. It is obvious, though, that in the video sequence segments where the ground truth data are unreliable, the error should be smaller than it actually is.

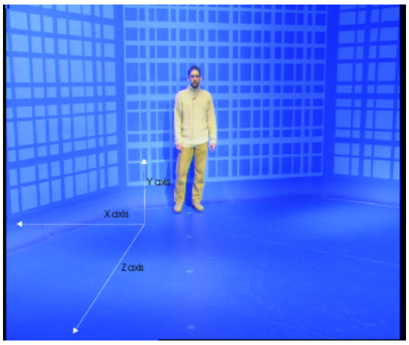


Figure 9: *The coordinate system for the CARROUSO project video sequences*

Calculation of 3-D coordinates for faces was proven to be inaccurate, due to the fact that the angle between the projection line and the Z-axis was very small. For this reason, the track-

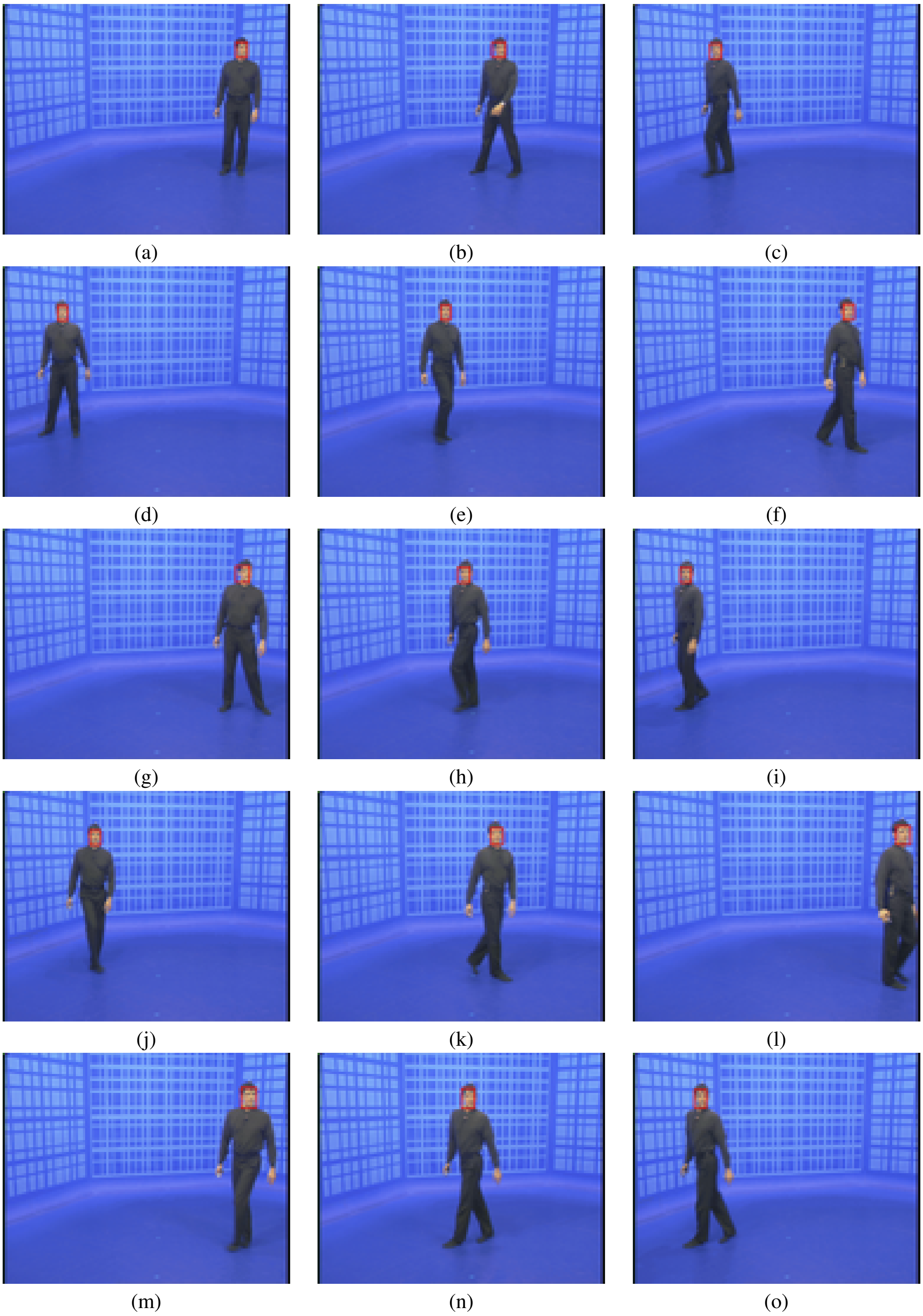
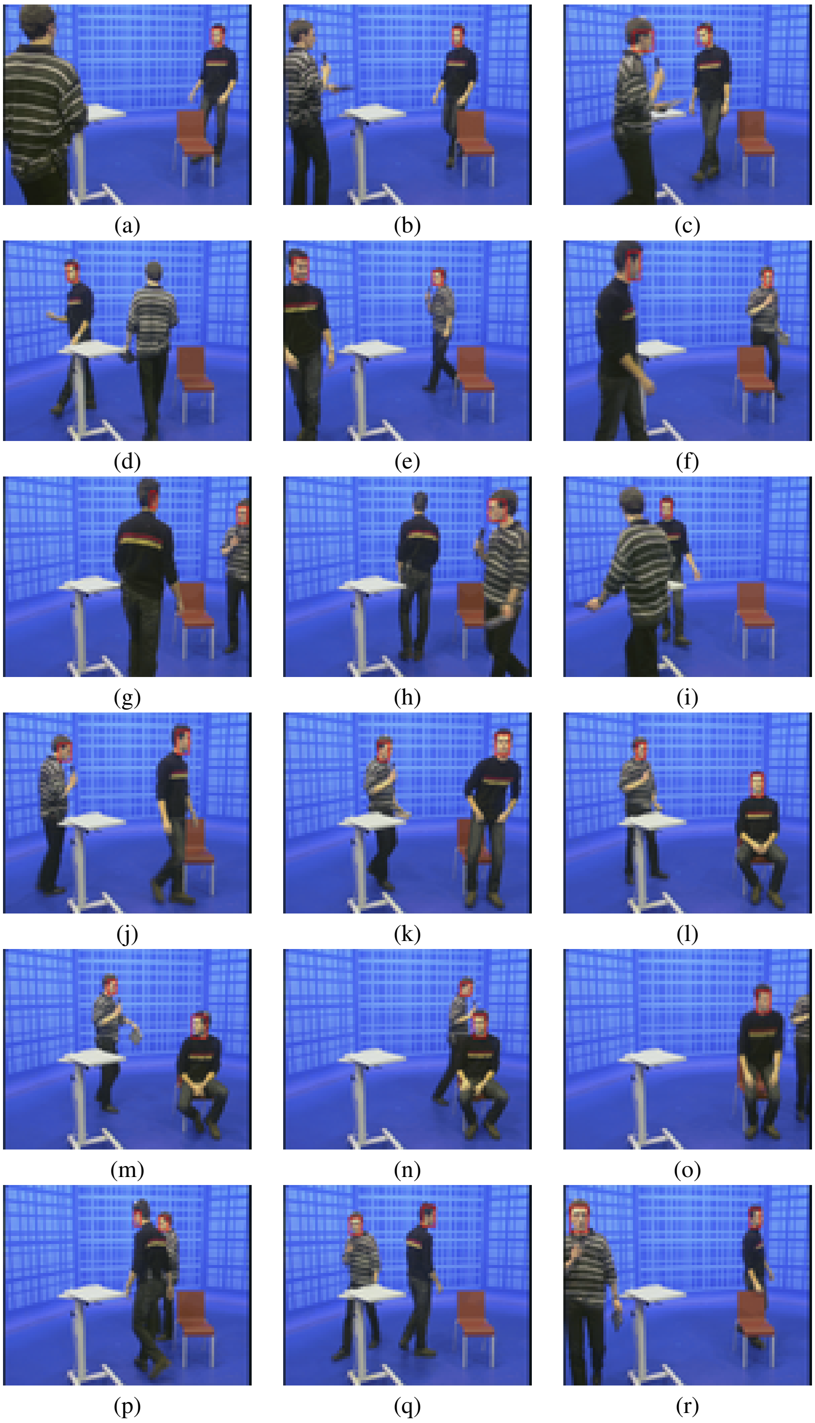


Figure 6: *2-D tracking results on the first test video sequence, 700 frames, sample frames displayed at 50-frame intervals (order: top-to-bottom, left-to-right).*

ing process was initialized manually. Instead of the face(s), both feet of the subject(s) were tracked. The center of gravity of the feet was used to calculate the 3-D coordinates. Since the feet center of gravity lies on the same vertical axis with the mobile transmitter (located on the head of the subjects), direct comparison between the X, Z tracking data (derived for the feet) and the X, Z ground truth data (provided for the head) was possible. As referred in Section 4, in order to calculate the 3-D coordinates, the point in question was assumed to lie constantly in the same plane. More specifically, it was assumed that the center of gravity of the feet was always located at a constant height, i.e. the value of the Y-coordinate of the point was constant. Since the subjects moved on a horizontal floor with no stairs or ramps, this assumption was indeed a valid one.

Figures 12 and 13 depict the absolute error between the tracking data and the ground truth data for the two video se-



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

(p) (q) (r)

Figure 7: *2-D tracking results on the second test video sequence, 850 frames, sample frames displayed at 50-frame intervals (order: top-to-bottom, left-to-right).*

quences of Figures 6 and 7. The erroneous peaks in the error plots of the calculated 3-D coordinates are associated with the instantaneous glitches of the infrared tracking system. Some examples of such inaccuracies are illustrated in Figures 10 and 11. Erroneous peaks in the error plots are also identified to indicate unreliable ground truth data. The X-coordinate absolute error for the first sequence ranges from 0 to approximately 0.4 meters, if the erroneous peaks are not taken into account. The absolute error for the Z-coordinate ranges from 0 to 0.6 meters. One source of error is associated with the fact that system-provided and ground truth 3-D coordinates refer to different points on the subject (feet and head respectively). Moreover, the assumption of constant height is at times violated, therefore producing deviations in the calculated 3-D coordinates. The corresponding errors for the second sequence are generally higher, possibly due to the unrestricted movement of the subject(s), ranging from 0 to 0.6 meters and 0 to 0.7 meters for the X and the Z coordinates respectively. It is obvious, however, that the results obtained are very satisfactory for a number of applications.

Additionally, a number of single-subject indoor and outdoor

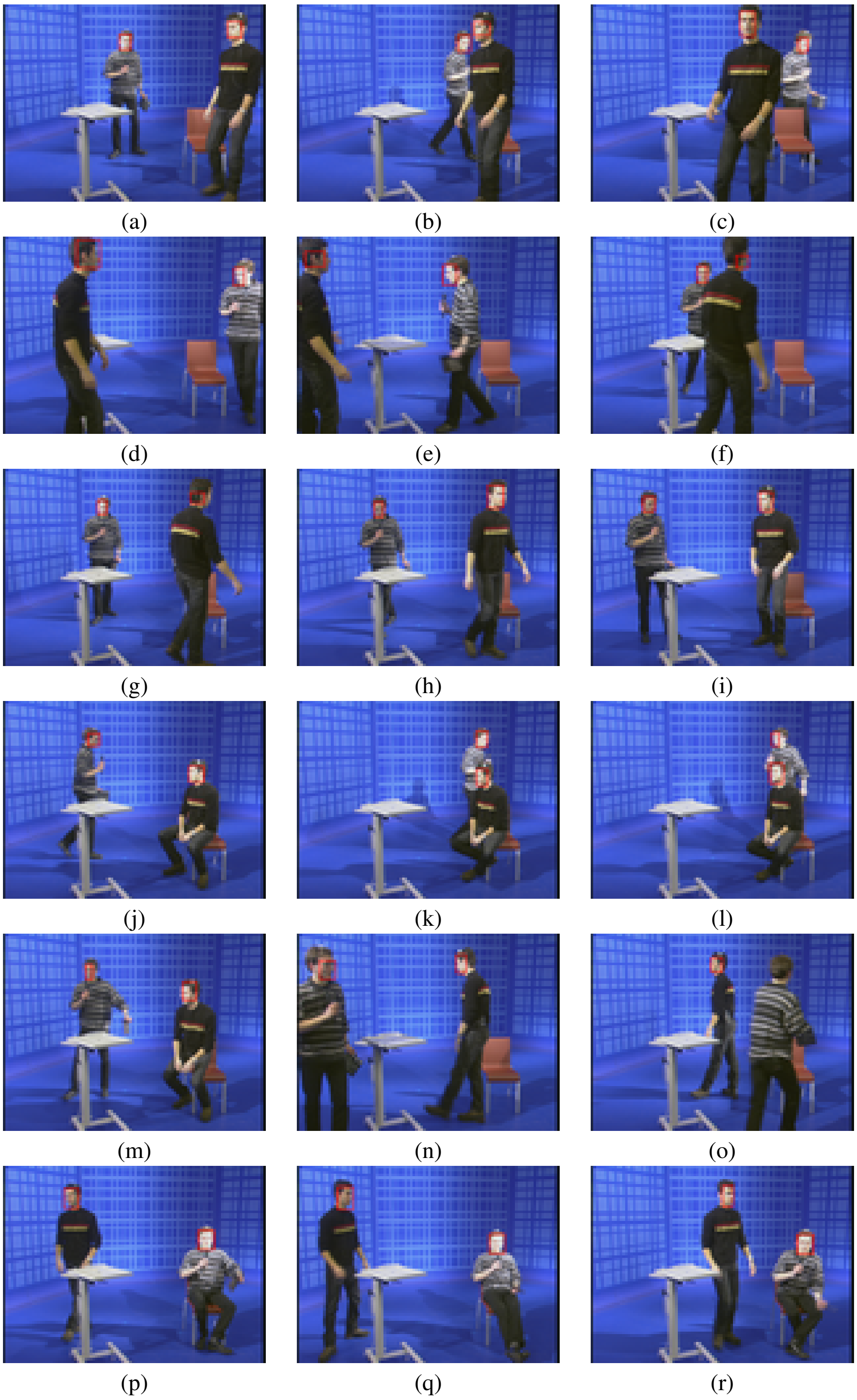
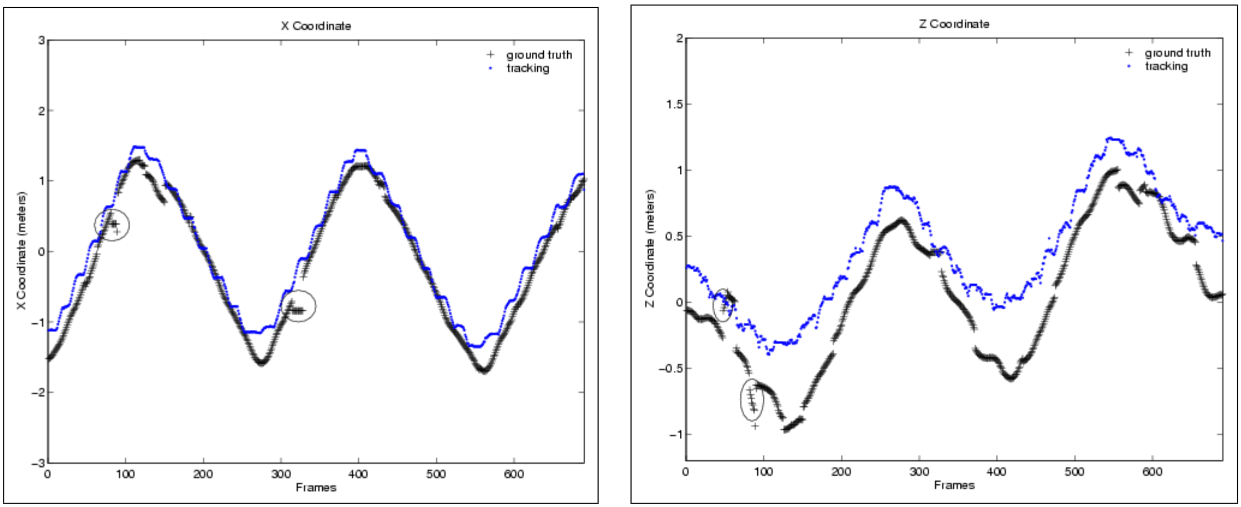


Figure 8: *2-D tracking results on the third test video sequence with sub-optimal lighting conditions, 750 frames, sample frames correspond to frame 0, 30, 62, 100, 150, 200, 250, 276, 300 and at 50-frame intervals afterwards (order: top-to-bottom, left-to-right).*

video sequences were available for testing. The results of automatic face detection and 2-D tracking for one of these video sequences are illustrated in Figure 14. In the sequence, a female subject is moving almost parallel to the camera, staying within the field-of-view of the camera at all times, with outdoor lighting conditions. Visual inspection of Figure 14 shows the correct localization of the face throughout the sequence.

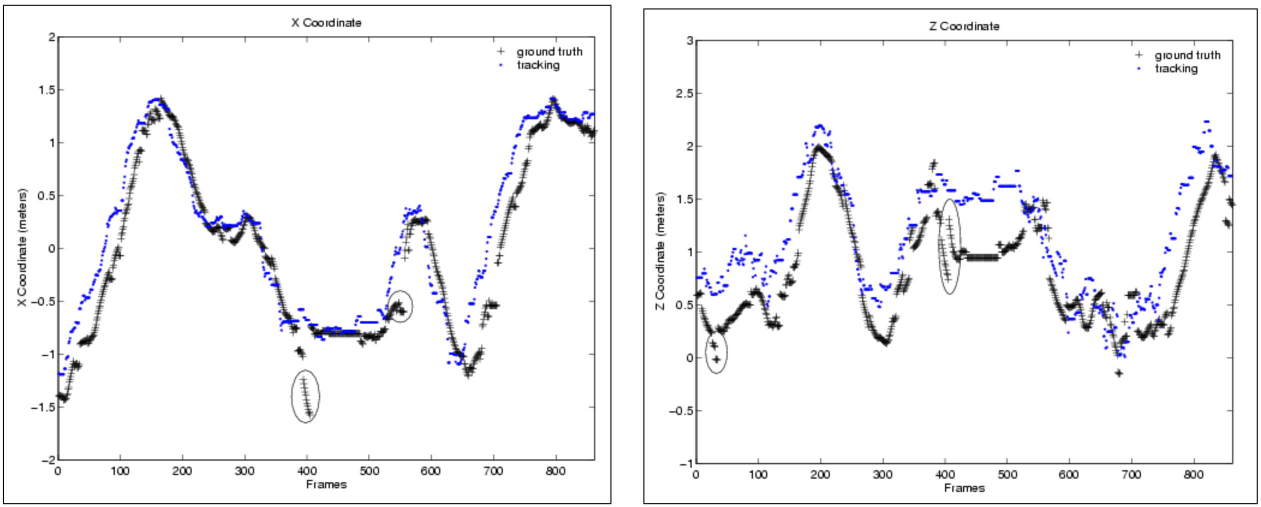
7. CONCLUSION

In this paper, a complete system for tracking people in 2-D, as well as calculating their 3-D coordinates using a calibrated camera 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,



(a) (b)

Figure 10: *3-D tracking results produced by the system, with respect to the (a) X coordinate and (b) Z coordinate for the sequence (first 690 frames) in Figure 6. Black crosses correspond to ground truth data associated with the sequence. Ellipses indicate unreliable segments of ground truth data values.*



(a) (b)

Figure 11: *3-D tracking results produced by the system, with respect to the (a) X coordinate and (b) Z coordinate for the sequence in Figure 7. Black crosses correspond to ground truth data associated with the sequence. Ellipses indicate unreliable segments of ground truth data values.*

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. The system can also accommodate calibrated tracking and can hence provide 3-D coordinates of any tracked object(s) of interest. It has been tested on a variety of video sequences, including a database of studio video sequences, for which 3-D ground truth data originating from a 4-camera infrared tracking system exist. It has been shown to perform reliably, especially when compared to expensive commercial tracking systems. Finally, fine-tuning for adaptation to different environments has been provided by means of user-specified parameters both for detection and tracking.

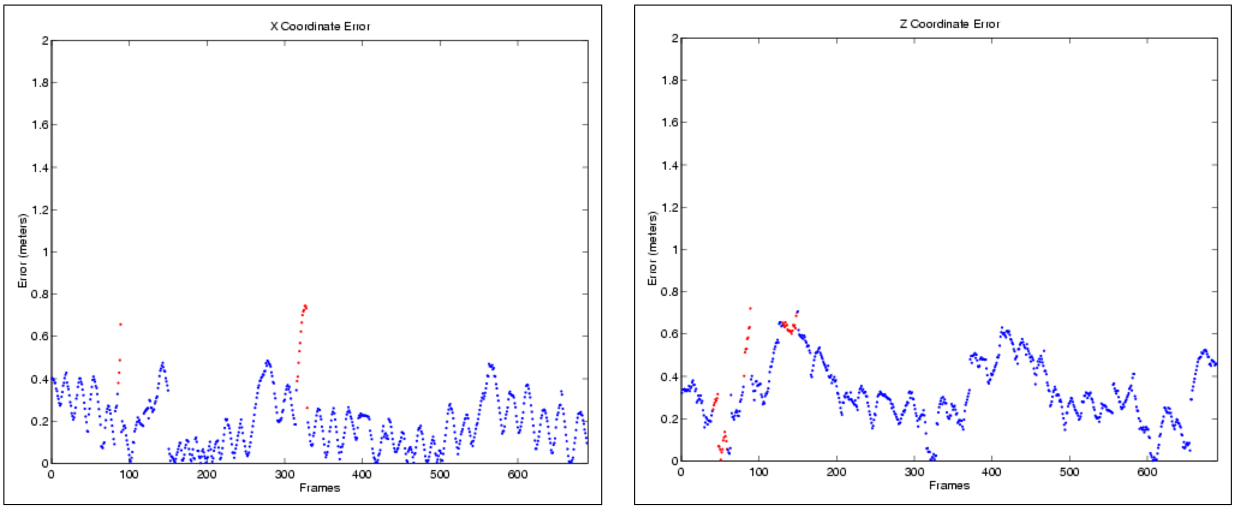
8. ACKNOWLEDGEMENTS

This work has been conducted in conjunction with the “SIMILAR” European Network of Excellence on Multimodal Interfaces of the IST Programme of the European Union (http://www.similar.cc).

9. REFERENCES

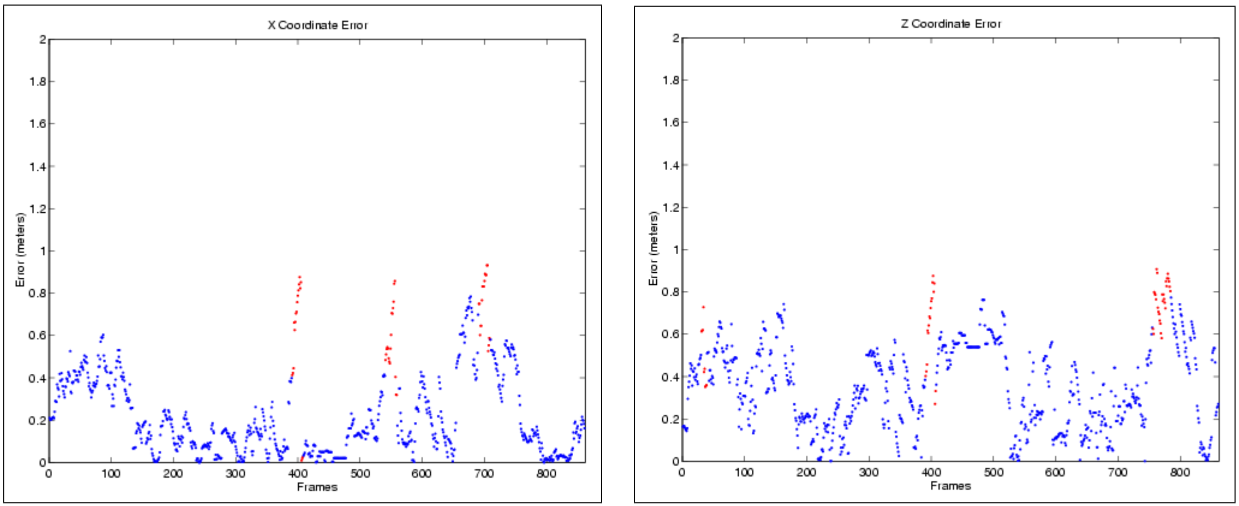
[1] I. Haritaoglu, D. Harwood, and L. S. Davis, “Ghost: a human body part labeling system using silhouettes”, in *Fourteenth International Conference on Pattern Recognition (ICPR98)*, vol. 1, (Vienna, Austria), pp. 77–82, August 1998. 31

[2] J. Han and B. Bhanu, “Detecting moving humans using color and infrared video”, in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent*



(a) (b)

Figure 12: *Absolute error of the system, with respect to ground truth for the (a) X coordinate and (b) Z coordinate, for the sequence in Figure 10. Red dots indicate unreliable error values due to unreliable ground truth data.*



(a) (b)

Figure 13: *Absolute error of the system, with respect to ground truth for the (a) X coordinate and (b) Z coordinate, for the sequence in Figure 11. Red dots indicate unreliable error values due to unreliable ground truth data.*

Systems (MFI2003), (Tokyo, Japan), pp. 228–233, July 2003. 31

[3] A. Wu, M. Shah, and N. D. V. Lobo, “A virtual 3D blackboard: 3D finger tracking using a single camera”, in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition (AFGR2000)*, (Grenoble, France), pp. 536–543, March 2000. 31

[4] Z. Duric, F. Li, Y. Sun, and H. Wechsler, “Using normal flow for detection and tracking of limbs in color images”, in *Sixteenth International Conference on Pattern Recognition (ICPR2002)*, vol. 4, (Quebec, Canada), pp. 268–271, August 2002. 31

[5] M.-H. Yang, D. J. Kriegman, and N. Ahuja, “Detecting Faces in Images: A survey”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002. 31

[6] E. Hjelmas and B. K. Low, “Face Detection: A survey”, *Computer Vision and Image Understanding*, vol. 83, pp. 236–274, 2001. 31

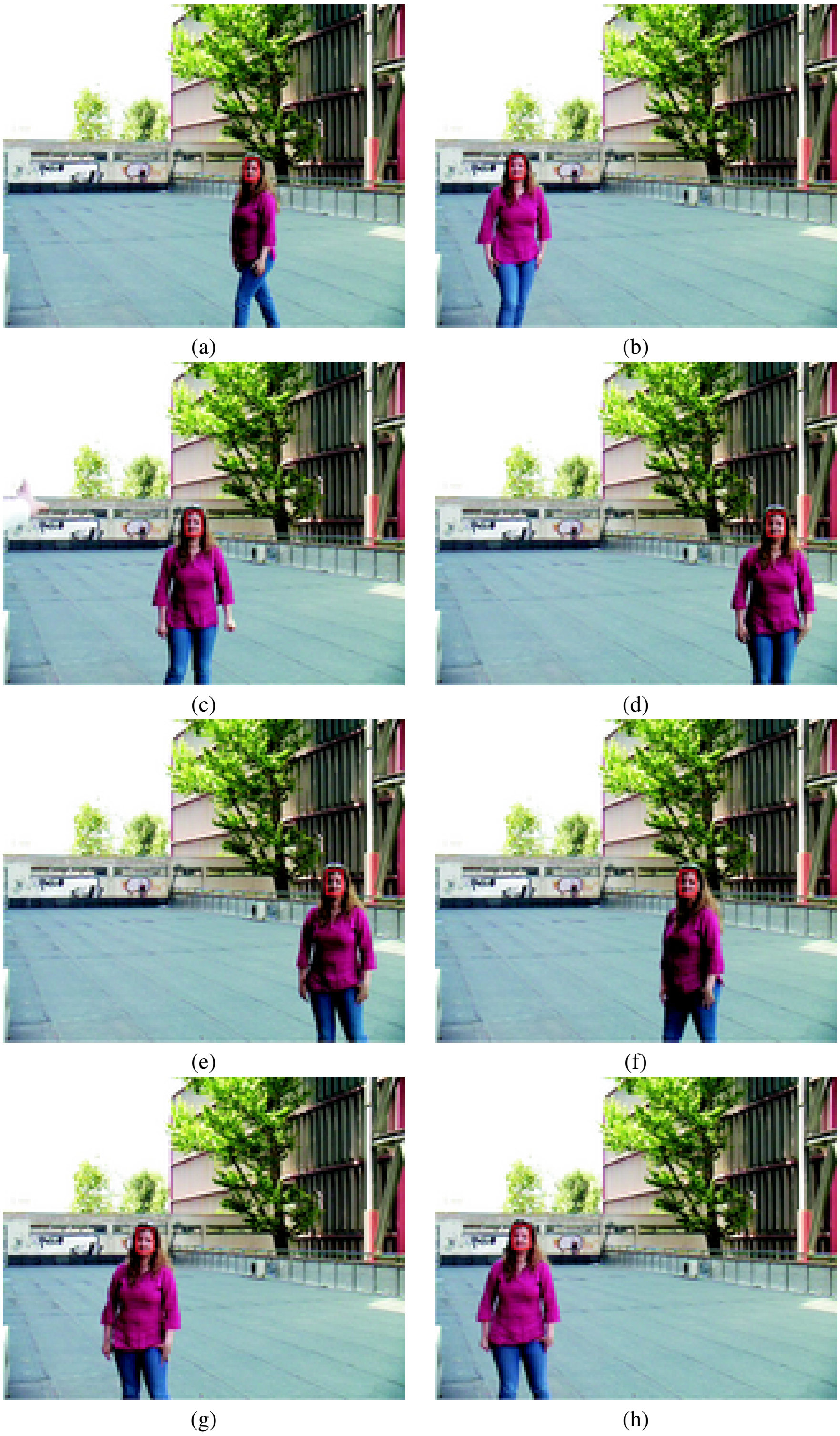
[7] G. Stamou, M. Krinidis, E. Loutas, N. Nikolaidis, and I. Pitas, “2D and 3D Motion Tracking in Digital Video”, in *Handbook of Image and Video Processing* (A. C. Bovik, ed.), Academic Press, 2005. 31

[8] T. B. Moeslund, A. Hilton, and V. Krüger, “A Survey of Advances in Vision-Based Human Motion Capture and Analysis”, *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–127, 2006. 31

[9] D. M. Gavrila, “The Visual Analysis of Human Movement: A Survey”, *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999. 31

[10] J. K. Aggarwal and Q. Cai, “Human Motion Analysis: A Review”, *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999. 31

[11] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “PFinder: Real-Time Tracking of the Human Body”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997. 31



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 14: *2-D tracking results on an outdoor test video sequence, 640 frames, sample frames displayed at 80-frame intervals (order: top-to-bottom, left-to-right).*

[12] O. Bernier, M. Collobert, R. Feraud, V. Lemaire, J. E. Viallet, and D. Collobert, "MULTRAK: A system for automatic multiperson localization and tracking in real-time", in *Fifth IEEE International Conference on Image Processing (ICIP98)*, vol. 1, (Chicago, United States), pp. 136–140, October 1998. 31

[13] A. Colmenarez, B. Frey, and T. S. Huang, "Detection and tracking of faces and facial features", in *Sixth IEEE International Conference on Image Processing (ICIP99)*, vol. 1, (Kobe, Japan), pp. 657–661, October 1999. 31

[14] L. L. Yang and M. A. Robertson, "Multiple-face tracking system for general region-of-interest video coding", in *Seventh IEEE International Conference on Image Processing (ICIP2000)*, vol. 1, (Vancouver, Canada), pp. 347–350,

September 2000. 31

[15] I. Haritaoglu, D. Harwood, and L. S. David, "W4: Real-Time Surveillance of People and Their Activities", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000. 32

[16] M. Krinidis, N. Nikolaidis, and I. Pitas, "2D Feature-Point Selection and Tracking Using 3-D Physics-Based Deformable Surfaces", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 876–888, July 2007. 32

[17] 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*, vol. 7, no. 1, pp. 124–137, 1997. 32, 33

[18] P. Viola and M. J. Jones, "Robust Real-time Object Detection", Tech. Rep. 01, Cambridge Research Laboratory, 2001. 32, 33, 37

[19] C. Tomasi and T. Kanade, "Shape and Motion from Image Streams: a Factorization Method - Part 3 Detection and Tracking of Point Features", Tech. Rep. 91-132, Computer Science Department, Carnegie Mellon University, 1991. 32, 35, 36, 37

[20] J. Shi and C. Tomasi, "Good Features to Track", in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR94)*, (Seattle, United States), pp. 593–600, June 1994. 32, 35, 36, 37

[21] C. Terrillon, M. David, and S. Akamatsu, "Automatic Detection of Human Faces in Natural scene Images by Use of a Skin Color Model and Invariant Moments", in *Third IEEE International Conference on Automatic Face and Gesture Recognition (AFGR98)*, (Nara, Japan), pp. 112–117, April 1998. 32

[22] A. Saber and A. Tekalp, "Frontal-View Face Detection and Facial Feature Extraction Using Color, Shape and Symmetry Based Cost Functions", *Pattern Recognition Letters*, vol. 17, no. 8, pp. 669–680, 1998. 32

[23] S. Tsekeridou and I. Pitas, "Facial Feature Extraction in Frontal Views using Biometric Analogies", in *IX European Signal Processing Conference (EUSIPCO98)*, vol. 1, (Rhodes, Greece), pp. 315–318, September 1998. 32

[24] H. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, "Multimodal System for Locating Heads and Faces", in *Second IEEE International Conference on Automatic Face and Gesture Recognition (AFGR97)*, (Killington, VT), pp. 41–46, October 1996. 32

[25] K. Yow and R. Cipolla, "Locating Human Faces in Photographs", *Image and Vision Computing*, vol. 15, no. 9, pp. 713–735, 1996. 32

[26] V. Govindaraju, "Feature-Based Human Face Detection", *International Journal of Computer Vision*, vol. 19, no. 2, pp. 129–146, 1996. 32

[27] A. Samal and P. Iyengar, "Human Face Detection Using Silhouettes", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 9, no. 6, pp. 845–867, 1995. 32

[28] J. Miao, B. Yin, K. Wang., L. Shen, and X. Chen, "A Hierarchical Multiscale and Multiangle System for Human Face Detection in a Complex Background Using Gravity-Center Template", *International Journal of Pattern Recognition*, vol. 32, no. 7, pp. 1237–1248, 1999. 32

[29] M. J. Jones and P. Viola, "Fast Multi-view Face Detection", Tech. Rep. 96, Mitsubishi Electric Research Laboratories, 2003. 32, 38

[30] H. Rowley, S. Baluja, and T. Kanade, "Rotation Invariant Neural Network-Based Face Detection", in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR98)*, (Santa Barbara, CA, United States), pp. 38–44, June 1998. 32

[31] H. Schneiderman and T. Kanade, "Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition", in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR98)*, (Santa Barbara, CA, United States), pp. 45–51, June 1998. 32

[32] K. Mikolajczyk, R. Choudhury, and C. Schmid, "Face detection in a video sequence - a temporal approach", in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR2001)*, vol. 2, (Kauai, Hawaii), pp. 96–101, December 2001. 32

[33] 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)*, (Corfu, Greece), pp. 58–63, September 1999. 32

[34] 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, Proceedings of SPIE* (M. Hunt, ed.), vol. 4301, (San Jose California, USA), pp. 102–113, January 2001. 32

[35] V. Vezhnevets, V. S. V, 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. 32

[36] A. Fitzgibbon and R. Fisher, "A Buyer's Guide to Conic Fitting", in *Fifth British Machine Vision Conference (BMVC99)*, (Birmingham, UK), pp. 513–522, 1995. 33

[37] R. Lienhart and J. Maydt, "An Extended Set of Haar-Like Features for Rapid Object Detection", in *IEEE International Conference on Image Processing (ICIP02)*, (Rochester, New York, USA), pp. 900–903, September 2002. 33, 34

[38] E. Loutas, K. Diamantaras, and I. Pitas, "Occlusion resistant object tracking", in *IEEE International Conference on Image Processing (ICIP01)*, vol. 2, (Thessaloniki, Greece), pp. 65–68, October 2001. 35

[39] Z. Zhang, "Flexible Camera Calibration by Viewing a Plane from Unknown Orientations", in *Seventh IEEE International Conference on Computer Vision (ICCV99)*, vol. 1, (Corfu, Greece), pp. 667–673, September 1999. 37

[40] S. J. Maybank and O. D. Faugeras, "A theory of self-calibration of a moving camera", *The International Journal of Computer Vision*, vol. 8, no. 2, pp. 123–152, 1992. 37

[41] Q.-T. Luong and O. Faugeras, "Self-calibration of a moving camera from point correspondences and fundamental matrices", *The International Journal of Computer Vision*, vol. 22, no. 3, pp. 261–289, 1997. 37

[42] E. Trucco and A. Verri, *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998. 37

[43] S. Pingali and J. Segen, "Performance Evaluation of People Tracking Systems", in *Third IEEE Workshop on Applications of Computer Vision (WACV96)*, (Sarasota, Florida, USA), pp. 33–38, December 1996. 37

[44] M. Krinidis, G. Stamou, H. Teutsch, S. Spors, N. Nikolaidis, R. Rabenstein, and I. Pitas, "An Audio-Visual Database For Evaluating Person Tracking Algorithms", in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, (Philadelphia), March 2005. 38, 39

[45] "Commission of the European Communities, IST project CARROUSO (Creating, Assessing and Rendering in Real Time of High Quality Audio-Visual Environments in MPEG-4 Context)". http://www.emt.iis.fraunhofer.de/projects/carrouso/. 38