Monocular tracking of the human arm in 3D

Luis Goncalves†, Enrico Di Bernardo†‡, Enrico Ursella‡ and Pietro Perona†‡

† California Institute of Technology, 116-81, Pasadena, CA 91125, USA
‡ Università di Padova, Italy
{luis,dibe,ursella,perona}@systems.caltech.edu

Abstract

We address the problem of estimating the position and motion of a human arm in 3D without any constraints on its behavior and without the use of special markers. We model the arm as two truncated right-circular cones connected with spherical joints. We propose to use a recursive estimator for arm position, and to provide the estimator with error signals obtained by comparing the projected estimated arm position with that of the actual arm in the image. The system is demonstrated and tested on a real image sequence.

1 Introduction and motivation

Observing the human body in motion is key to a large number of activities and applications:

Security – In museums, factories and other locations that are either dangerous or sensitive it is crucial to detect the presence of humans and monitor/classify their behavior based upon their gait and gestures.

Animation – The entertainment industry makes increasing use of actor-to-cartoon animations where the motion of cartoon figures and rendered models is obtained by tracking the motion of a real person.

Virtual reality – The motion of the user of a virtual reality system is necessary to adjust display parameters and animations.

Human-machine interfaces – The motion of the human body may be used as a convenient interface between man and machine. For example the hand could be used as a 3D mouse.

Biomechanics – Reconstructing the 3D motion of human limbs is used for clinical evaluation of orthopedic patients and for training of both professional and amateur athletes.

Signaling – In airports, at sea, and in other high-noise environments the arms and torso are used for signaling.

Camera control – Active camera control based on the motion of humans can be used for sport events, conferences, and shows, thus replacing human operators. It may also be used to make human operators more effective in security monitoring.

Traffic monitoring – Pedestrians are often a component of street traffic. They need to be detected and their behavior understood (e. g. intention to cross at a traffic light, gesture signaling for emergency help) in order to help avoid collisions and dangerous situations, and in order to detect accidents immediately.

Customer monitoring – Data on the behavioral pattern of exploration and purchasing of store customers is extremely valuable to advertising companies, producers and sales management.

Current techniques for tracking the human body involve a large variety of methods. *Security, traffic monitoring, signaling, and customer monitoring* are typically implemented using human observers that survey the scene either directly or via a multiple camera closed circuit TV system. For *animation and biomechanics* multiple camera systems and manual tracking of features across image sequences is used. For *virtual reality* an assortment of gloves, suits, joysticks and inductive coils is used. For *human-machine* interfaces we have joysticks, mice and keyboards.

All of these methods require either employing dedicated human operators or using ad-hoc sensors. This results in a number of limitations:

1. *Practicality* – the user needs to wear markers or other ad-hoc equipment which may be impractical, uncomfortable, constrain the user to a limited work space, be difficult to transport;

2. *Cost* – computational and sensory hardware and human operator time.

0-8186-7042-8/95 $4.00

3. *Timeliness* – The data may not be available in real-time, but only after a lag required to process a batch of images, allow communication between human operators etc.

If tracking the human body could be made automatic and non-invasive, and therefore cheaper, more practical and faster, not only the applications listed above could be better performed, but also a number of new applications would be feasible.

1.1 Automatic human motion estimation

Previous work on human motion estimation can be coarsely grouped into three types :

- gesture classification [5, 6]
- systems which track or classify periodic motions with 1 degree of freedom [12, 9, 10]
- estimation of 3D unconstrained motion; of the hand from a monocular view [11]; of the body, with the use of multiple cameras and special markers [2]

We are interested in estimating 3D unconstrained motion. The most accurate system at the moment is ELITE [2], which is able to estimate position with an accuracy of 0.1% of the workspace diameter. However, to achieve such accuracy, it is necessary to use a system composed of special-purpose markers, infrared lighting, and 4-8 cameras that need to be accurately calibrated. As in [11], we explore the opposite side of the spectrum of approaches to the problem: how accurately can one track in 3d the human body with the simplest, cheapest, and most convenient setup: a single grayscale camera and no special markers.

In this paper we study the more constrained problem of estimating the motion of the human arm. This is a good starting point because arm estimation is very useful in numerous applications (eg. human-machine interaction, security); it's also easily extendible to the whole body since leg structure is very similar to that of the arm and so only torso and head tracking remain to be done.

In the next section we present some theoretical considerations on the accuracy achievable in depth reconstruction from a monocular view. Afterwards, we describe our arm model, the estimation method, and an experiment with a 1132 frame sequence. Finally, we discuss the success of the method and the directions for future research.

2 The accuracy achievable from a monocular view

We will not attempt to give a full analysis, but rather to provide some intuition. From a monocular view, the depth of an isolated point is impossible to recover; it is information on relative structure which allows depth to be determined. The simplest structure to study is that of a line segment. In our analysis, we keep one endpoint of the segment fixed in space and calculate the dependence of the depth estimation of the other endpoint with respect to deviations in its image plane coordinates. In order to facilitate comparison with our experiment, we calculate image coordinates using the camera parameters of the camera in the experiment. Furthermore, we place the fixed endpoint of the line segment along the optical axis at the same depth as the location of shoulder during the experiment, and the length of the line segment is the length of the upper-arm of the subject performing the experiment.

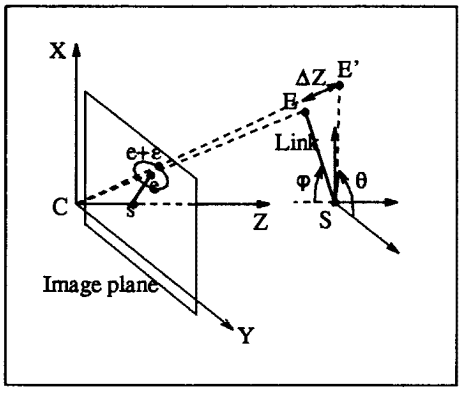

Figure 1: ***Estimating depth sensitivity:*** *What is the change in depth (ΔZ) for a given image coordinate error (ϵ) given the angle ϕ from the optical axis?*

Fig.1 shows the coordinate system used in calculating the sensitivity. The angle ϕ is the angle between the segment and the optical axis. Fig.2(top) shows the error of the endpoint's depth estimate when it's image coordinates are disturbed by 1 pixel (in the worst direction). The qualitative nature of the results agree with one's initial intuition; when the line segment is pointing towards the camera, the depth error is quite small, and when the segment endpoints are equidistant from the camera the depth error increases quickly to infinity. From this

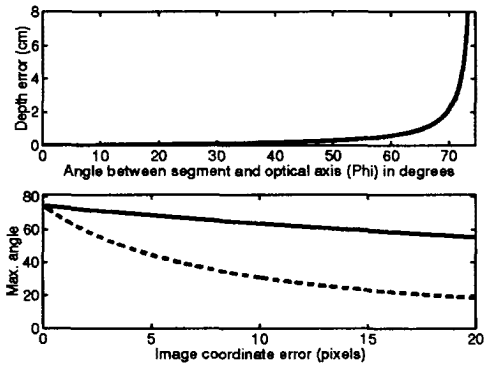

Figure 2: ***Sensitivity Analysis:*** *(Top):Typical error in depth estimate of the free endpoint of a line segment when there is 1 pixel error in it's image plane coordinates. (Bottom) An example of the maximum permissible angle between the segment and the optical axis as a function of image plane coordinate error, so that relative error is less than 7.5% (solid), 1.0% (dashed).*

plot we calculate as a function of image coordinate error the maximum angle of deviation from pointing straight at the camera permissible so that the relative depth error is less than 7.5% and less than 1%(Fig.2(bottom)). It can be seen that as the accuracy required increases, the range of acceptable positions decreases significantly.

Of course, since our estimation method (described below) does not use measurements of endpoint position in the image plane, a direct application of these results is not possible, however, it gives us some idea of how sensitive the depth estimation is with respect to the information in the image. One can improve upon pure length measurements by using both the dynamics of the system being estimated (ie, the existence of a smooth trajectory, velocity), and more of the structure of the object (ie, even if one arm segment is in an ill-conditioned position, the other may not be, and so the position of the arm can still be estimated accurately).

3 The estimation system

We develop a method for arm tracking inspired by Dickmann's work on lane following [4]. The basic idea is that rather than extracting features from the image, to make direct comparisons between what is in the image and what is expected. The current estimate of arm position is used to predict the appearance of the arm in the image, and the difference between the predicted and actual images is then used as an error measurement for a recursive estimator.

3.1 The arm model

In order to generate the predicted image, we need to "render" a 3D model of the arm from the camera's point of view. We choose a simple 3D model (Fig.3) in which the upper and lower arm are modeled as truncated right-circular cones, and the shoulder and elbow joints are modeled as spherical joints. Clearly, a real arm is not conical, however, we hope that except for cases of extremely muscular arms the approximation is sufficiently accurate. Furthermore, the elbow joint is not spherical but planar, but for a conical limb model (as well as for a stick model) rotation along the limb axis is unobservable, and thus the motions of models with spherical and planar elbow joints are indistinguishable. Assigning two degrees of freedom to each limb or one DOF to the forearm and three to the upper arm becomes a matter of aesthetic preference.

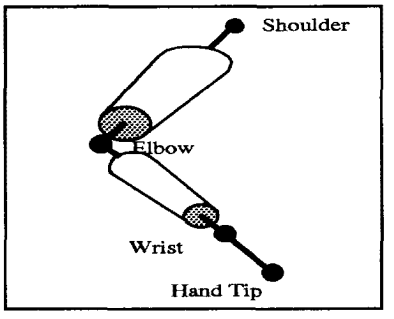

Figure 3: ***The arm model:*** *Limbs are modeled as truncated right-circular cones. The elbow and shoulder joints are modeled as spherical joints, and the hand tip is assumed to be along the forearm axis.*

The strongest simplification of our arm model concerns the modeling of the hand. At this initial stage, in order to keep the model as simple as possible and to have as few degrees of freedom as possible (only four positional DOF with two spherical joints), we do not model (or render) the shape of the hand, but simply assume it to extend along the axis of the forearm, with the fingertip of the middle finger a constant distance from the wrist joint.

This strong simplification will eventually be eliminated, in order to be able to track arms in which the hand moves with respect to the forearm, but it provided us with a reasonable starting point to test the concept.

Our model thus requires 7 parameters to describe it's shape: the longitudinal lengths of the hand, forearm and upper arm (3), and the diameters of the two limb segments at each end (4). Each of these parameters must be measured with at least 1 cm accuracy. Furthermore, we assume that the 3D position of the shoulder is known. There is a natural hierarchy to the segmentation of the human body, and shoulder position is determined by tracking the torso. Since we attempt to track only the arm, we assume the shoulder position is known.

3.2 The Recursive Estimator

Our recursive estimator is an Extended Kalman Filter with implicit measurements [8, 3, 7]. The state of the filter is the four spherical joint angles of the arm model. The dynamics on the state is simply a random walk in the spherical coordinates. The covariance matrix of the random walk is calculated at each step such that it remains constant in a Euclidean coordinate system for the hand tip and elbow positions. This calculation compensates for the fact that the sensitivity of the hand and elbow positions with respect to the spherical coordinates varies from one point in state space to another. The implicit measurements are a non-linear function of the current state estimate and the current image. The key to the success of the method is the ability to obtain a linearization of the measurement equation. This involves calculating the jacobian of the measurements with respect to the state (and the image), which in turn involves knowing a precise camera calibration. Although a straightforward calculation, it is rather tedious, and we refer the details to [1].

There are very many choices for possible image error measurements. The ultimate measurement (technically impossible for the time being) would be to use a biomechanically accurate model of the arm, along with knowledge of texture, reflectance function and lighting conditions, to render a prediction of the image with photographic quality, and then measure the difference of intensity values between predicted and real image. The measurement currently in use is considerably simpler (Fig.5). We image the arm against a dark background (Figs.7,8) to have high contrast boundaries for most part of

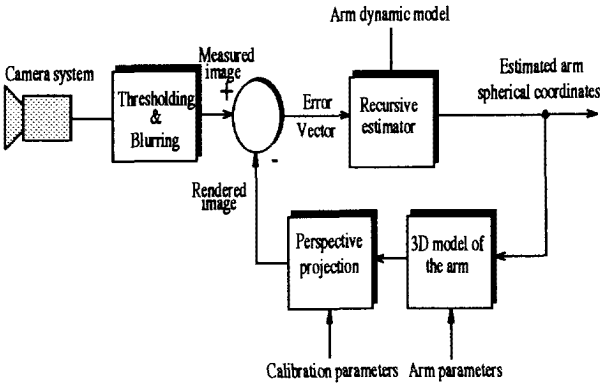

Figure 4: ***The estimation system:*** *Real and rendered arm views are compared to provide an error signal to a recursive estimator.*

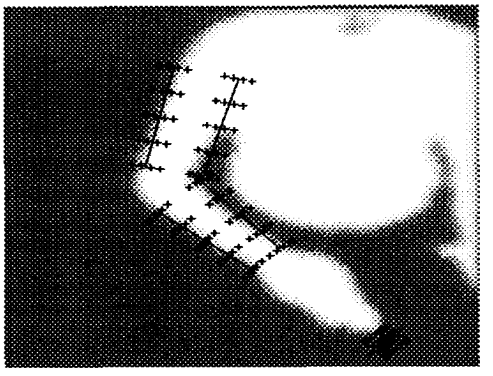

Figure 5: ***The measurements for the recursive estimator:*** *The (+) indicate locations on the thresholded and blurred image where intensity values are compared with those from the predicted arm position (outlined by lines).*

the arm. The image is first thresholded, and then blurred with a 2 dimensional Gaussian filter of specified width. Then the difference between the real image and the predicted image is calculated at 20 points on both sides of each (predicted) limbs' contours and the predicted hand tip position (100 measurements in all). If the predicted and the real image fall exactly on each other (and in the absence of measurement noise and modeling error), all these differences are zero, otherwise, the deviations from this ideal value constitutes the innovation process used by the Kalman filter for the update of the state

estimate (Fig.6).

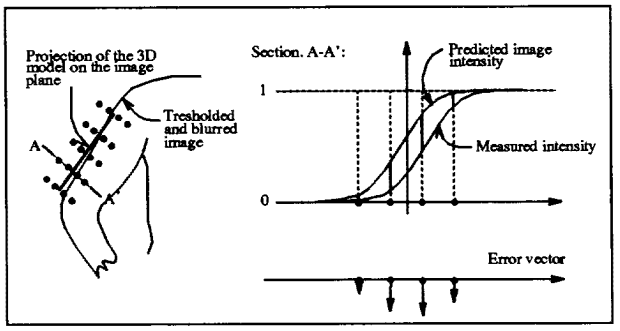

Figure 6: ***Detail of the measurement process:*** *Transverse to the predicted arm contour, the acquired thresholded and blurred image is sampled at 4 points. The difference between the expected and measured intensity values generates an error vector.*

Kalman filtering theory assumes that the measurement noise is Gaussian with zero mean. For our measurements, it is very difficult to determine the noise statistics, and surely it is not Gaussian. Arguably we may assume it is zero mean, and since the thresholded values are bounded, we can calculate an upper bound on the noise variance.

4 Experimental Results

4.1 Description of the experiment

The method was tested on a real sequence of images in which a human subject moved his fingertip along a rectangular pattern drawn on a table. Before commencing the motion, a checkerboard pattern (Fig.7) was placed on the table in order to calibrate the camera and to obtain the ground truth of the motion. In order for the real arm to match well with our simple arm model, during the motion the subject maintained his hand rigidly extended along the axis of the forearm, and attempted to maintain his shoulder position fixed in space. The sequence consists of 1 1/4 turns along the rectangular path, and is 1132 frames long (acquired at 30 frames/sec). The initial fit of the model to the first frame was done manually, and the tracking was performed twice; on every frame (30 Hz rate), and on every 10th frame (3 Hz rate) to explore the sensitivity of the method with respect to frame rate.

4.2 Experiment Results

Fig.8 shows the initial position of the hand (with the initial estimate) and the true and tracked tra-

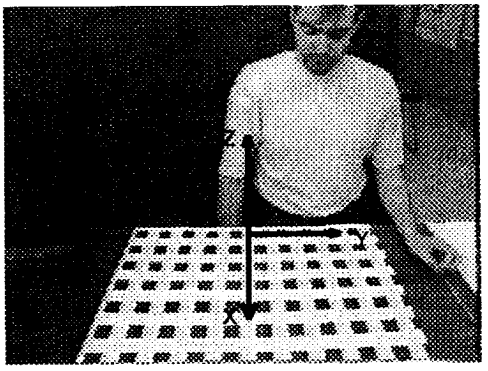

Figure 7: ***The calibration image:*** *A checkerboard pattern is used to determine the camera parameters and the transformation between the camera reference frame and the ambient (checkerboard) reference frame (necessary to determine the ground truth).*

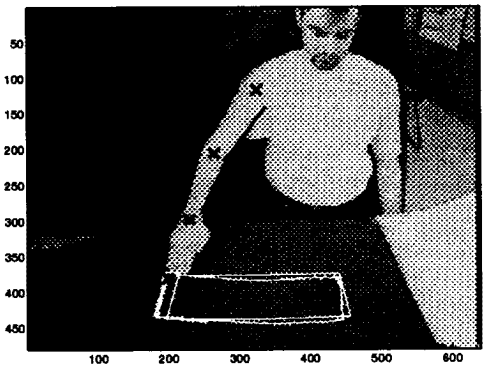

Figure 8: ***The tracked trajectory:*** *The initial arm position is shown with the initial estimate. (X) mark estimated shoulder, elbow, wrist and fingertip positions, solid lines outline estimated arm. Ground truth trajectory as well as estimated trajectory are shown. Shift of estimated trajectory is due to the fingertip not lying on the forearm axis.*

jectory as they appear from the camera point of view. The tracked trajectory appears shifted to the right (in the subject's reference frame) with respect to the real trajectory. The reason for this shift is that the subject's middle finger fingertip was not exactly aligned with the forearm axis, but instead was approximately 2 cm to the left of it. Since the

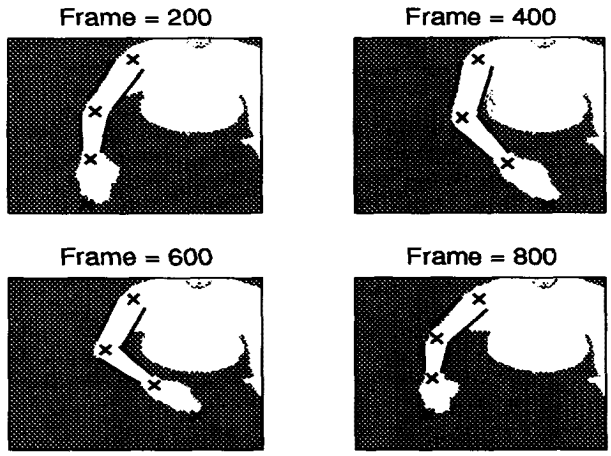

Figure 9: ***The tracked arm:*** *The estimated arm position projected onto the thresholded image for 4 frames of the sequence.*

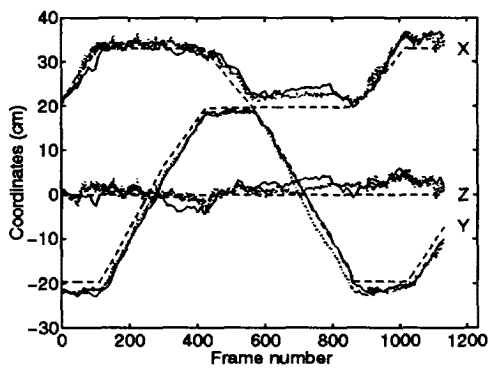

Figure 10: ***Position versus time:*** *The tracking error when every frame was tracked (dotted) and when every 10th frame was tracked (solid) as compared to ground truth (dashed). The errors of the two tracks are comparable and under 5 cm in all coordinates.*

motion was performed by tracing out the rectangular path with the middle finger fingertip, the tracked trajectory appears shifted to the right. Fig.9 shows the projection of the estimated arm position on the thresholded images at four different positions during the motion.

Fig.10 compares the tracked and the true trajectory in the ambient reference frame as a function

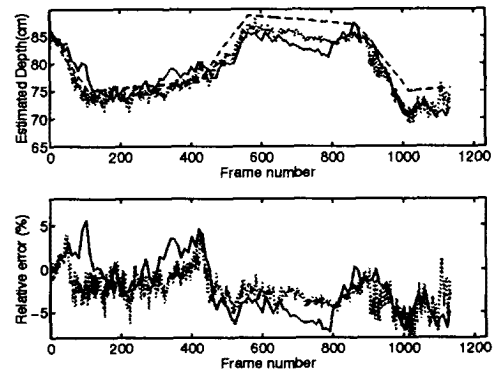

Figure 11: ***Depth error versus time:*** *Top plot shows tracking error in the direction of the camera optical axis. Dotted - every frame tracked, Solid - every 10th frame tracked, Dashed - ground truth. Maximum error is less than 5 cm. Bottom: relative depth error is less than 7.5%.*

of time. In this reference frame the Z component should always be zero (since the fingertip is always on the surface of the table), and it is estimated as such with a maximum error of 5 cm. The X and Y components are also estimated with an error of less than 5 cm at both tracking rates. Fig.11 shows that the depth error (error in the direction of the camera optical axis) is also less than 5 cm at both tracking rates (a 7.5% relative error).

In performing the experiment, we found that the tracking was very sensitive to the arm parameters. A considerable amount of time was spent trying to adapt the 7 arm parameters and the shoulder position in order to obtain these results. With the original untuned parameters, the errors were between two to three times as large. It is possible that more tuning can reduce the error further, however, the accuracy of the tracking is limited by the fact that there is unmodeled movement of the shoulder (assumed stationary) and of the hand with respect to the forearm axis.

5 Conclusions and Future Work

We have demonstrated that the human arm may be tracked accurately in 3D using a single camera and a simple 7 parameter model.

Our promising results suggest several directions for future work:

In order to eliminate the assumption of constant shoulder position, we need to develop a kinematic

model of the head, neck, and shoulder region of the body. A new model for the hand is also necessary to allow for hand movement with respect the forearm. It is also necessary to develop simple and structured methods for automatically determining the arm parameters.

In order to continue our study on the accuracy of 3D position attainable from a monocular view, it would be interesting to experiment with lenses of different focal length. In particular, a wide angle lens will produce a larger perspective deformation which should aid in tracking since there will be larger variation in apparent arm width depending on arm position. Also, it is extremely straightforward to extend our method to receive input from more than one camera, and this will allow us to make a direct comparison between monocular and stereoscopic tracking.

Currently our method is limited by the fact that we assume the arm is moving in front of a dark background. This is a justifiable initial simplification since it allowed us to test the general principles of the method without having to worry about developing a reliable and high quality measurement. To increase the practicality of the system a measurement which can work in more general scenes must be developed. There are very many possible choices, the only requirement is that for whatever property of the image we decide to measure we can also predict what the measurement would be based on the estimated state, so that an error signal can be obtained.

We have still not dealt with the problem of occlusions. Intuitively, a proposed solution can be to use the current estimate of the arm position to compute where the occlusions occur in the image plane, and make image measurements only at unoccluded locations.

Acknowledgments

This work is supported in part by the California Institute of Technology; a fellowship from the "Ing.A.Gini" foundation; the Office of Naval Research grant ONR NOOO14-93-1-0990; an NSF National Young Investigator Award; the Center for Neuromorphic Systems Engineering as a part of the National Science Foundation Engineering Research Center Program; and by the California Trade and Commerce Agency, Office of Strategic Technology.

References

[1] E. Di Bernardo, L. Goncalves, and P. Perona. State-guided image feedback for arm motion estimation. Technical report, California Institute of Technology, April 1995.

[2] N.A. Borghese, M. Di Rienzo, G. Ferrigno, and A. Pedotti. Elite: A goal oriented vision system for moving objects detection. *Robotica*, 9:275–282, 1991.

[3] R.S. Bucy. Non-linear filtering theory. *IEEE Trans. A.C. AC-10, 198*, 1965.

[4] E. D. Dickmanns and V. Graefe. Applications of dynamic monocular machine vision. *Machine Vision and Applications*, 1:241–261, 1988.

[5] W. T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. Technical Report 94-03a, Mitsubishi Electric Research Labs., 201 Broadway, Cambridge, MA 02139, 1994.

[6] W. T. Freeman and C. Weissman. Television control by hand gestures. Technical Report 94-24, Mitsubishi Electric Research Labs., 201 Broadway, Cambridge, MA 02139, 1994.

[7] A.H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

[8] R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. of the ASME-Journal of basic engineering.*, 35-45, 1960.

[9] S.A. Niyogi and E.H. Adelson. Analyzing gait with spatiotemporal surfaces. *Proceedings of the Workshop on Motion of non-rigid and articulated objects*, pages 64–69, 1994.

[10] R. Polana and R.C. Nelson. Recognizing activies. In *Proceedings of ICPR*, 1994.

[11] J.M. Rehg and T. Kanade. Digiteyes: Vision-based hand tracking for human-computer interaction. In *Proceedings of the workshop on Motion of Non-Rigid and Articulated Bodies*, pages 16–24, November 1994.

[12] K. Rohr. Incremental recognition of pedestrians from image sequences. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 8–13, New York City, June, 1993.