Real-time tracking of highly articulated structures in the presence of noisy measurements

T. Drummond

Department of Engineering
University of Cambridge
Cambridge, UK CB2 1PZ

R. Cipolla

Department of Engineering
University of Cambridge
Cambridge, UK CB2 1PZ

Abstract

This paper presents a novel approach for model-based real-time tracking of highly articulated structures such as humans. This approach is based on an algorithm which efficiently propagates statistics of probability distributions through a kinematic chain to obtain maximum a posteriori estimates of the motion of the entire structure. This algorithm yields the least squares solution in linear time (in the number of components of the model) and can also be applied to non-Gaussian statistics using a simple but powerful trick. The resulting implementation runs in real-time on standard hardware without any pre-processing of the video data and can thus operate on live video. Results from experiments performed using this system are presented and discussed.

1. Introduction

Visual tracking of complex, highly articulated structures is an important technology for several domains. In particular there has been much interest in tracking human motion [9]. There are many applications including tasks such as surveillance, motion capture and human-computer interaction. The problem addressed by this paper is that of real-time tracking an articulated structure in the view of one or more cameras. The system is model-based [16, 12, 7, 4] and uses a CAD model that comprises piecewise rigid components with curved surfaces and known kinematic constraints.

Several tasks such as human-computer interaction require real-time performance. Attaining this is difficult and there are relatively few examples of such systems. An exception is Pfinder [17] which does so by limiting its processing to tracking coloured blobs in the image plane. Black and Jepson [1] also use an image-based approach with multiple eigenspaces to recognise hand gestures. Cham and Rehg [3] suggest that real-time performance is not possible in three dimensions and use a two-dimensional scaled prismatic model. Hel-Or and Werman [8] use a fully articulated three-dimensional model but use an $O(N^3)$ algorithm to obtain a least squares solution for the pose. Here we present an $O(N)$ algorithm which propagates statistics of probability distributions along the kinematic chain to obtain the maximum a posteriori solution for the pose in real-time. Where these statistics are Gaussian, the same least squares solution is obtained. Further the algorithm can be adapted to operate iteratively with robust (non-Gaussian) statistics. Delamarre and Faugeras [4] also use an $O(N)$ algorithm and give timings which are essentially real-time, but make use of a powerful (but comparatively expensive) pre-processing technique (Geodesic active contours) which yields impressive results.

In this work, we represent the pose of each component of the model separately as an element of the group of rigid body motions in three dimensions, SE(3). This is a redundant representation and requires that the articulation constraints be explicitly represented, however it has the advantages that it provides a symmetric representation and is not reliant upon accurate localisation of some key component to identify the pose of the remainder of the model. This is by contrast to [13, 2] which use a minimal parameterisation based on a tree structure. We make use of the exponential map [2, 15] connecting the Lie algebra of SE(3) with the group to represent motions, and use the adjoint representation of the group to transform quantities in the algebra between different coordinate frames.

To obtain robust real-time performance in the presence of noisy measurements, it is important to use a strong statistical framework. MacCormick and Blake [14] use a refinement of the Condensation algorithm [11] which partitions the search space in order to reduce computational complexity. This was extended by Deutscher et. al. [5] who used coupled monte-carlo techniques to track highly articulated human motion although the computational cost was prohibitive for real-time application. Cham and Rehg [3] also use a particle based method, but capture the structure of local modes of the posterior distribution in each particle and thus need fewer particles. The approach employed here is to use an iterative re-weighted least-squares technique which allows the use of the fast O(N) algorithm with non-Gaussian statistics.

For applications that deal with existing imagery, typically only a single view is available and several systems have been developed which can handle this [3, 2, 13, 1]. The system described here can also operate using a single view, although the performance is greatly improved when more views are available. We use an edge-based approach (as [7]). In our case the model is rendered first and matching edges are then sought in the image rather than processing the image to detect edges first and then matching those to the model.

Currently, we only attempt to solve the motion problem (where is each component at each time step) by contrast to [12] who use data to refine the model in addition to computing pose.

2. Geometric representation

2.1. Pose

The system presented here represents the pose of an articulated structure with a matrix which describes the transformation from the coordinates of each model component to each camera. These matrices form the group of Euclidean transformations on three-dimensional space, SE(3) and have the form:

$$E = \begin{bmatrix} R & t \\ 000 & 1 \end{bmatrix} \tag{1}$$

where R is an orthogonal 3×3 matrix with determinant 1 and t is an arbitrary 3-vector. For convenience, the bottom row of these matrices (being constant) will be omitted for the remainder of the paper. The internal camera parameters for each camera are also stored and thus the projection matrix for each component of the model into each camera can be computed.

2.2. Structure

In order to construct models of structures with curved surfaces, a representation based on intersections of pairs of quadrics is used. This is a convenient representation as it permits the use of standard structures such as truncated cones, cylinders and spheres as well as truncated ellipsoids and hyperbolic surfaces.

These structures can be rapidly rendered by computing the normalised image conic C for each quadric Q in the view of a normalised camera with a Euclidean projection matrix E. Each quadric can be transformed into camera coordinates to give

$$Q' = E^{-T} Q E^{-1} \tag{2}$$

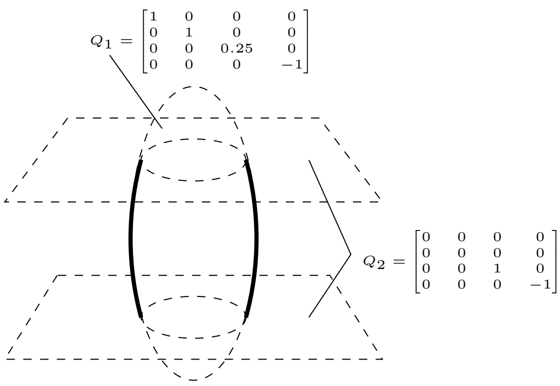

Figure 1: Q_1 is the quadric to be rendered and is clipped by the (in this case degenerate) quadric Q_2. Only the sample points on the conic image of Q_1 that correspond to points inside Q_2 are rendered (shown in bold).

If the transformed quadric is written as

$$Q' = \begin{bmatrix} A & \mathbf{b} \\ \mathbf{b}^T & c \end{bmatrix} \tag{3}$$

then the conic can be computed as

$$C = Ac - \mathbf{b}\mathbf{b}^T \tag{4}$$

The conic C can then be rendered by mapping a unit circle (parameterised by θ) to C and generating a series of sample points (to be used in tracking, see Section 3) from a discrete set of θs. These are evenly separated with spacing Δp by setting $\Delta\theta = \frac{\Delta p}{dp/d\theta}$. Each sample point generated by this procedure is then checked for visibility (in front of the camera, within the intersecting quadric and not occluded by any other pair of intersected quadrics).

The articulated structures considered in this paper are constructed from these primitives. The structures comprise piecewise rigid components and each component is made up of a number of these clipped quadrics. The quadrics in each component share a common coordinate frame and are all rendered using the Euclidean projection matrix for the component.

The kinematic chain is represented by specifying a parent for each component and listing the constraints that apply between the two. These are described in more detail in Section 4. Because two components may share the same parent, the kinematic chain may form a tree structure, however articulated cycles are not permitted.

2.3. Motion

The task of this system is to compute the Euclidean matrices E for each component into each camera at each time step.

This problem can be reduced to computing a Euclidean motion matrix M such that

$$E_t = E_{t-1} M_t \tag{5}$$

The matrices M give the transformation from the coordinate frame of a component at time t to the frame at time $t-1$ (coarse manual initialisation is given for the first frame). Because the motion is represented in the coordinate frame of the moving component, the motion matrices are the same for all stationary cameras viewing the scene and thus only need to be computed once per component per time frame regardless of the number of views. The matrices M can be represented by points $\boldsymbol{\alpha}$ in the Lie algebra of SE(3) by means of the exponential map [2].

$$M = \exp\left(\sum_{i=1}^{6} \alpha_i G_i\right) \tag{6}$$

where G_i are a set of generator matrices which form a basis for the algebra. Because the time steps are small (40ms for PAL video), the motion matrices are close to the identity and thus $\boldsymbol{\alpha}$ should be close to the origin. The tracking task is then reduced to a six dimensional search for $\boldsymbol{\alpha}$ for each component at each time frame.

It should be noted that if a component contains just a single quadric, it will appear in the image as a single conic having just five degrees of freedom. Thus there will be (in general) a one parameter family of solutions for $\boldsymbol{\alpha}$ for this component considered in isolation and this can only be constrained by considering the articulation constraints.

3. Statistical framework

The statistical formulation used in this approach defines the probability of observing an image given a particular pose of the model in terms of the presence of edges in the image close to those rendered in the model. These edges are detected by performing a 1-dimensional search from each sample point in a direction normal to the rendered edge. These measurements are assumed independent and the probability of the image given a pose is the product of the probabilities of each measurement.

3.1. Gaussian statistics

If these measurements were normally distributed then the maximum a posteriori pose would be given by a least-squares solution. Such a solution can be computed very efficiently from the Jacobian of partial derivatives of each distance measurement with respect to each motion parameter α_i. If the distance measurements are small, an accurate approximation to the Jacobian J can be obtained by considering the edge-normal component of the motion of the sample points with variation of α_i which can be easily

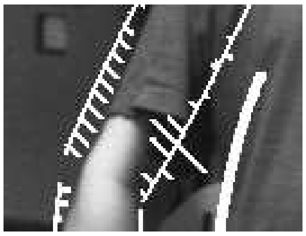

Figure 2: Noisy edge measurements. The solid bar indicates failure to detect an edge.

computed in closed form. This can be used to obtain a maximum a posteriori estimate of the motion parameters μ for each component

$$\boldsymbol{\mu} = C^{-1} J^T \mathbf{m} \qquad \text{where} \quad C = [J^T J] \tag{7}$$

and $\mathbf{m}$ is the vector of distance measurements. The sum-squared residual measurements S for a solution $\boldsymbol{\alpha}$ is given by

$$S = (\boldsymbol{\alpha} - \boldsymbol{\mu}) C (\boldsymbol{\alpha} - \boldsymbol{\mu}) + ||(\mathbf{m} - J\boldsymbol{\mu})||^2. \tag{8}$$

This gives a probability distribution over the Lie algebra (with two arbitrary constants a and b) of the form

$$p(\boldsymbol{\alpha}) = a \exp\left(-b(\boldsymbol{\alpha} - \boldsymbol{\mu}) C (\boldsymbol{\alpha} - \boldsymbol{\mu})\right) \tag{9}$$

3.2. Non-Gaussian statistics

In practice, the measurement distribution is found to be non-Gaussian by virtue of containing many more samples in the tails of the distribution (see Figure 2). Such distributions can be handled within the least squares framework by introducing a re-weighting function [10]. This function is evaluated for each distance measurement m in $\mathbf{m}$ and is used to scale m and the corresponding row of J. If the re-weighting function is $w(m)$, then the least squares procedure iteratively converges on a solution given by the probability distribution

$$p(m) = \exp\left(-\int_0^m m' w(m') dm'\right) \tag{10}$$

This approach has the advantage that distributions other than Gaussian can be modelled, convergence is fast and each iteration requires just a linear solution. The real power of this approach becomes apparent when complex systems such as kinematic chains are considered since it is relatively easy to propagate Gaussian statistics through the chain and these can then be wrapped with a re-weighting function to obtain an iterative solution to the desired statistics. The re-weighting function used in this work is $w(m) = 1/(c + |m|)$. This results in a distribution that behaves as a Gaussian for $m \ll c$ and a Laplacian for $m \gg c$.

4. Highly articulated kinematic chains

Pose statistics are now considered for kinematic chains. Equation (9) gives independent probability distributions for the motion parameters for each component of the model. The model is not free to move arbitrarily, however, since the articulation constraints must be respected. This means that the desired solution is one which maximises the product of the probabilities for the motion of each component and also satisfies the constraints. This can be computed efficiently (in linear time in the number of components) by propagating the statistics through the model. This is done by using the constraints present between each adjacent pair of links in the kinematic chain to obtain the maximum a posteriori motion for the entire chain. This motion will provide the least squares solution for the pose, subject to the kinematic constraints.

This is achieved in two stages, shown in Algorithm 1. First the joint distribution of each component's motion and that of its parent in the kinematic chain is considered. The motion of the child component can be marginalised by allowing it to take its modal value conditional upon the motion of the parent. By doing this, the statistics of the parent component can be modified to incorporate those of the child. This process is repeated up the kinematic chain until the component at the top of the chain carries the propagated statistics for the entire chain. Second, the maximum a posteriori pose of each component is assigned, starting at the top of the chain and propagating back down the chain.

4.1. Propagating statistics

It has been shown [6] that constraints corresponding to slides, hinges and ball joints can be linearised and are homogeneous on the values of the motion parameters of the two components $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ and thus have a simple form:

$$D_{12}\boldsymbol{\alpha_1} + D_{21}\boldsymbol{\alpha_2} = \mathbf{0}. \tag{11}$$

Within $\boldsymbol{\alpha_1}$-$\boldsymbol{\alpha_2}$ space there is an embedded manifold of motions which respect the physical articulation constraints. Equation (11) corresponds to forcing the motion to lie in the tangent space to this manifold at the origin (through which the manifold must pass). For a ball joint positioned at x, y, z in the coordinate frame of component 1, the constraint matrix D_{12} is:

$$D_{12} = \begin{bmatrix} 1 & 0 & 0 & 0 & -z & y \\ 0 & 1 & 0 & z & 0 & -x \\ 0 & 0 & 1 & -y & x & 0 \end{bmatrix} \tag{12}$$

The motion $\boldsymbol{\alpha_2}$ is represented in component 2's coordinate frame which is different to that of component 1. If these coordinate frames were the same then $D_{21} = -D_{12}$. Thus if α_2 is transformed into component 1's coordinate frame using the adjoint representation of the transformation between

Algorithm 1 Statistical propagation on a kinematic chain

> **for** $i = 1 \cdots n-1$ **do**
>   Let p = i's parent
>   Marginalise $\boldsymbol{\alpha_i}$ and update $\boldsymbol{\mu_p}$ and C_p using (21), (22)
> **end for**
> assign $\boldsymbol{\alpha_n} = \boldsymbol{\mu_n}$
> **for** $i = n-1 \cdots 1$ **do**
>   assign $\boldsymbol{\alpha_i}$ using (19)
> **end for**

these coordinate frames the constraint in (11) becomes

$$D_{12}\boldsymbol{\alpha_1} - D_{12}\,\mathrm{Ad}(E_1^{-1}E_2)\boldsymbol{\alpha_2} = 0 \tag{13}$$

(where E_1 and E_2 are the Euclidean projection matrices of components 1 and 2). Thus

$$D_{21} = -D_{12}\,\mathrm{Ad}(E_1^{-1}E_2). \tag{14}$$

Given $\boldsymbol{\alpha_2}$ it is possible to obtain the value of $\boldsymbol{\alpha_1}$ which satisfies the constraints (11) and minimises the sum-squared residual (8). This can be computed by introducing Lagrange multipliers $\boldsymbol{\lambda}$ for the constraints and solving

$$2C_1(\boldsymbol{\alpha_1} - \boldsymbol{\mu_1}) + D_{12}^T\boldsymbol{\lambda} = \mathbf{0} \tag{15}$$

$$\text{giving} \quad \boldsymbol{\alpha_1} = \boldsymbol{\mu_1} - \frac{1}{2}C_1^{-1}D_{12}^T\boldsymbol{\lambda} \tag{16}$$

$$\text{So} \quad D_{21}\boldsymbol{\alpha_2} + D_{12}\boldsymbol{\mu_1} - \frac{1}{2}D_{12}C_1^{-1}D_{12}^T\boldsymbol{\lambda} = 0 \tag{17}$$

$$\text{and} \quad \boldsymbol{\lambda} = 2\left[D_{12}C_1^{-1}D_{12}^T\right]^{-1}(D_{21}\boldsymbol{\alpha_2} + D_{12}\boldsymbol{\mu_1}) \tag{18}$$

$$\boldsymbol{\alpha_1} = \boldsymbol{\mu_1} - C_1^{-1}D_{12}^T\left[D_{12}C_1^{-1}D_{12}^T\right]^{-1}(D_{21}\boldsymbol{\alpha_2} + D_{12}\boldsymbol{\mu_1}) \tag{19}$$

The total sum-squared error for components 1 and 2 can be written as

$$S_{1,2} = (\boldsymbol{\alpha_1} - \boldsymbol{\mu_1})C_1(\boldsymbol{\alpha_1} - \boldsymbol{\mu_1}) + (\boldsymbol{\alpha_2} - \boldsymbol{\mu_2})C_2(\boldsymbol{\alpha_2} - \boldsymbol{\mu_2}) \tag{20}$$

(discarding the constant terms). By substituting the optimal value for $\boldsymbol{\alpha_1}$ from (19), $S_{1,2}$ can computed as a function of $\boldsymbol{\alpha_2}$. This gives new values for $\boldsymbol{\mu_2}$ and C_2 which take into account the optimal pose of component 1.

$$\boldsymbol{\mu_2'} = \boldsymbol{\mu_2} - C_2^{-1}D_{21}^T\left[D_{12}C_1^{-1}D_{12}^T + D_{21}C_2^{-1}D_{21}^T\right]^{-1}(D_{12}\boldsymbol{\mu_1} + D_{21}\boldsymbol{\mu_2}) \tag{21}$$

$$C_2' = C_2 + D_{21}^T\left[D_{12}C_1^{-1}D_{12}^T\right]^{-1}D_{21} \tag{22}$$

This propagates the statistics of component 1 through the articulated joint into component 2. The process can be repeated, propagating the new statistics for component 2 into the next component up the chain until the statistics for all components have been propagated into the root of the chain.

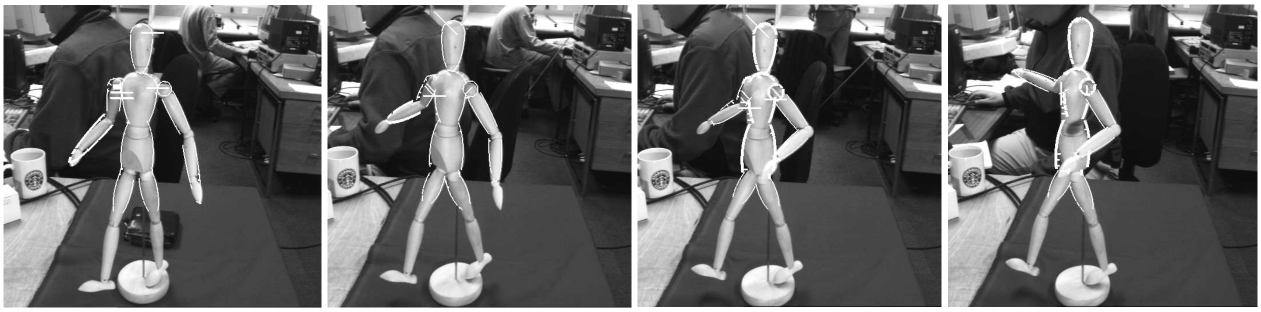

Figure 3: Tracking the wooden mannequin in a single view. Solid lines represent failure to find an edge on the search path

At this point, the final value of $\boldsymbol{\mu}$ for the root of the chain represents the maximum a posteriori value for the motion parameters $\boldsymbol{\alpha}$ of this component, taking into account the measurements of the entire articulated structure. Equation (19) can then be used to propagate the maximum a posteriori estimate for the pose of the entire structure back down the kinematic chain so that the maximum a posteriori pose of the entire chain is obtained.

As mentioned previously, this entire process is wrapped within a iterative re-weighting scheme so that the iterative behaviour of the system is equivalent to propagating the correct non-Gaussian statistics along the kinematic chain. Although the use of non-Gaussian statistics results in an iterative algorithm (with each iteration involving execution of Algorithm 1), this system only performs one iteration per video frame. Where the motion is large, a single iteration is found to provide a sufficiently good solution to permit tracking to continue and where it is small, the algorithm converges very rapidly.

4.2. Coercing the constraints

Because the manifold corresponding to poses of the model which respect the articulation constraint is curved and the process outlined above is only correct to first order, the resultant pose will contain a quadratic error which violates the constraints. Thus at each stage, it is necessary to remove this error and ensure that the constraints are met exactly. Since the error is very small, it is possible to achieve this very simply by running down the kinematic chain and coercing the constraint in a non-symmetric manner. This is done by computing the logarithm of the matrix describing the transformation between each pair of components and projecting its coefficients into the right null space of the constraint matrix D_{12}.

$$\text{Find} \quad \mathbf{a} \quad \text{s.t.} \quad \exp\left(\sum_i a_i G_i\right) = E_2^{-1} E_1 \tag{23}$$

$$\mathbf{a}' = \mathbf{a} - D_{12}^T \left[D_{12} D_{12}^T\right]^{-1} D_{12}\, \mathbf{a} \tag{24}$$

The Euclidean projection matrix E_1 can then be rebuilt to exactly satisfy the constraint by

$$E1 = E2 * \exp\left(\sum_i a_i' G_i\right) \tag{25}$$

5. Experimental Results

Two humanoid models have been used for experiments tracking a wooden mannequin and a human. The models are similar and both have 18 degrees of freedom. The model of the mannequin is quite accurate and fits well, while the model of the human is only approximate.

5.1. Mannequin

An experiment was performed to track the mannequin in the view of a single camera. Because of difficulty in moving the mannequin by hand without causing catastrophic occlusion, a series of incremental movements were performed with the tracker turned off. After each movement, the tracker was operated for approximately a quarter of a second (about 6 frames) before the next movement was made. Results from this tracking sequence are shown in Figure 3. The single camera tracking system runs in real-time at PAL video frame rate (25Hz).

5.2. Human

For the human tracking experiment, views from three synchronised cameras were used. Single view tracking was found to be much less effective on this task due to a relatively poor fit between the human and the model. Figure 4 shows a number of frames from a tracking sequence in this experiment.

6. Summary and Conclusions

A novel method for fast tracking of highly complex articulated structures has been presented. The algorithm runs in real-time at 25 Hz for a single view and 10Hz for three views. The maximum tolerable speed of motion that this algorithm can handle is limited however, particularly in the human case due to inaccuracies in the CAD model. Future work will consider methods for extending the framework to compute the structural parameters of the model.

Figure 4: Tracking a human in three concurrent views (cameras distributed vertically).

References

[1] M. J. Black and A. D. Jepson. Eigen tracking: Robust matching and tracking of articulated objects using a view based representation. In *Proceedings of ECCV'96*, volume 1, pages 329–342, 1996.

[2] C Bregler and J Malik. Tracking people with twists and exponential maps. In *Proceedings of CVPR'98*, pages 8–15, 1998.

[3] T-J Cham and J M Rehg. A multiple hypothesis approach to figure tracking. In *Proceedings of CVPR'99*, volume 2, pages 239–245, 1999.

[4] Q. Delamarre and O. Faugeras. 3D articulated models and multi-view tracking with silhouttes. In *Proceedings of ICCV'99*, volume 2, pages 716–721, 1999.

[5] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proceedings of CVPR2000*, volume 2, pages 126–133, 2000.

[6] T. Drummond and R. Cipolla. Real-time tracking of multiple articulated structures in multiple views. In *Proceedings of ECCV2000*, volume 2, pages 20–36, 2000.

[7] D M Gavrila and L S Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *Proceedings of CVPR'96*, pages 73–80, 1996.

[8] Y. Hel-Or and M. Werman. Constraint fusion for recognition and localisation of articulated objects. *International Journal of Computer Vision*, 19(1):5–28, 1996.

[9] D. Hogg. A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[10] P. J. Huber. *Robust Statistics*. Wiley series in probability and mathematical statistics. Wiley, 1981.

[11] M. Isard and A. Blake. CONDENSATION - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[12] I. A. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *Proceedings of CVPR 1996*, pages 81–87, 1996.

[13] D. G. Lowe. Fitting parameterised 3-D models to images. *IEEE T-PAMI*, 13(5):441–450, 1991.

[14] J. MacCormick and A. Blake. Partitioned sampling, articulated objects, and interface quality hand tracking. In *Proceedings of ECCV2000*, volume 2, pages 3–19, 2000.

[15] R.M. Murray and S.S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

[16] J. M. Regh and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proceedings of ICCV'95*, pages 612–617, 1995.

[17] C. Wren, A. Azerbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE T-PAMI*, 19(7):780–785, 1997.