View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

Markov-based Silhouette Extraction for Three–Dimensional Body Tracking in Presence of Cluttered Background

Ali Shahrokni† Vincent Lepetit† Tom Drummond‡
Pascal Fua†*

† Computer Vision Laboratory, EPFL
CH-1015 Lausanne, Switzerland

‡ Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ

Abstract

We propose a novel method to detect human body contours in presence of clutter and complex texture. Contours are extracted using a novel Markov-based approach which learns a texture along a given scanline in order to detect texture crossings. In contrast to conventional silhouette detection algorithms based on gradient, our texture boundary detection method allows extraction of silhouettes of textured and non-textured objects under difficult conditions such as having a cluttered/moving background. We demonstrate on demanding examples of monocular body tracking that our proposed method yields better results than gradient-based techniques.

1. Introduction

A number of promising silhouette-based approaches to body tracking and human pose estimation have been proposed recently [1, 5, 6, 9, 13]. However most of them rely on the fact that silhouettes can be extracted using relatively simple algorithms such as background subtraction [4] or standard edge- and gradient-based techniques [7, 2, 1]. However, in practice, this rarely is the case and these silhouette extraction methods can be very brittle. They tend to fail in the presence of highly textured objects and clutter, which produce too many irrelevant edges. In such situations, it is advantageous to detect texture boundaries instead. However, because texture segmentation techniques such as those based on graph cuts [3] require computing statistics over image patches, they are more useful for detection in a single image than for tracking.

In this paper, we overcome these limitations by extending an earlier texture based approach [11] that was initially designed to detect the outlines of projected polygonal objects to finding the projected contours of 3–D articulated models that are usable for human body tracking. To this end, we have clarified the formulation and validated it using

*This work was supported in part by the Swiss National Science Foundation.

Figure 1: Left: Scanning a line through model sample point p for which a 1^{st} order statistical model associated is used to find the texture crossing point c. Right: An example of outlines of a person wearing loose clothes and standing against another person in low quality video obtained using our method.

extensive simulation analysis. This improved approach allows us to obtain good articulated body tracking results under difficult conditions due to textured or loose clothing and cluttered or moving background, which demonstrates its applicability under everyday life situations.

Our method is inspired by the use of Markov processes for texture description and segmentation. However, our requirements differ from those of classical texture segmentation methods in the sense that we wish to find the optimal pose parameters rather than arbitrary region boundaries. Given an approximate body pose and corresponding model outlines in the image such as those shown in Fig. 1-left, we model textures as Markovian processes that generate pixel sequences along scan lines perpendicular to these projected outlines. Finding the true outlines then becomes equivalent to detecting the boundary of the two texture processes responsible for the pixel intensities inside and outside of the object, which can be done robustly. To further increase reliability, we can incorporate a prior statistical model of the known target, in this case the texture inside a human body, which can be updated over time.

In the following sections, we first discuss related work on tracking using silhouettes, then we present and validate our approach to boundary detection using simulations and synthetic data. We will then demonstrate its effectiveness on real-data by incorporating it into a monocular body tracking algorithm that can take advantage of both silhouette data and 2–D correspondences between frames. In particular, we will show that using our approach to silhouette detection results in substantial improvement over standard edge-detection when the background is cluttered and changes over time. Our contribution is therefore not the tracking algorithm itself, which we use as a testbed, but the approach to silhouette extraction that we advocate and that yields the performance increase we observe.

2. Related Work

All methods for inferring 3–D pose of articulated models such as whole bodies [1, 7, 5, 6, 9, 8] or hands [12, 2, 14] using occluding contours depend on reliably extracting the outlines of the subject. For example, Kakadiaris and Metaxas [8] establish direct correspondences between points on the occluding contours to points on the surface of the model using projective geometry. Due to projective singularity along the line of sight

corresponding to points on the occluding contours, the success of this approach is highly correlated to how well the outlines are extracted. This is equally true for the methods such as [1, 2] that retrieve a registered pose based on how well it matches the outlines. Athitsos and Sclaroff [2] propose a method that can generate a ranked list of plausible 3–D hand configurations that best match an input image. Probabilistic line matching using Euclidean embedding of Chamfer distances from a set of reference images is used for image data base indexing. Their method yields good results in presence of clutter. However this requires a rich database with a high number of reference images for embedding of the distance transform with enough accuracy. Unfortunately, construction of such a database is not always practical, especially for a whole body scenario.

Using edges obtained by gradient based methods for example as in [5, 7, 9, 6] is appealing due to its simplicity and speed, but its application is restricted to the cases where the contrast is sufficient. Furthermore, these methods tend to fail in the presence of highly textured objects and clutter, which produce too many irrelevant edges. Background subtraction [4] techniques also require a good estimation of the background image which is not always possible especially if it is not static. In such situations, applying texture segmentation methods to tracking of body outlines becomes an attractive alternative because they work reliably well whether complex texture and clutter exists or not. However because texture segmentation techniques such as methods based on graph cuts [3] require computing statistics over image patches, they tend to be computationally intensive and have therefore not been felt to be suitable for tracking. By contrast, the approach presented in this paper does borrow ideas from the texture segmentation literature [10]. Hidden Markov random fields appear naturally in problems such as image segmentation, where an unknown class assignment has to be estimated from the observations at each pixel. Recently we have proposed a novel texture-based linesearch method [11] for texture boundary detection which can be used for real-time tracking of polygonal objects. This method has the advantage of texture segmentation methods, as it assumes a Markovian model of texture on the target and background areas in spite of relying solely on gradient or edge information while, unlike other texture segmentation algorithms it is fast and adapted to tracking applications thanks to its model-based structure and line search approach.

3 Robust Silhouette Detection

In this section, we describe how silhouette points on the body outlines are detected. To the best of our knowledge this is the first time that statistical models used for texture segmentation are adapted and applied to the problem of articulated body tracking. Our method unlike conventional segmentation methods, scans lines for texture boundary points and is fast as well as robust and reliable for any kind of object and background texture and complexity.

The process of texture boundary detection starts by obtaining and sampling the silhouettes of the projection of the predicted articulated model pose. Assuming that each body part in the chain of articulations is modeled by a quadric, as will be discussed in section 4 the projection of these quadrics is a set of conics on the image plane. The latter is used to generate a discrete set of sample points and the associated normal directions which correspond to the pose in the previous frame. This information serves to estimate

a new set of silhouette points for the current frame by employing the method described below.

3.1 Markov model and texture boundary detection

The silhouette points are found by detecting a change in the statistics of the texture along a scanline normal to the conic boundary, as shown in Fig. 1-left. We start by formalising the search criteria and derive the method we use to evaluate them.

A texture is modeled as a statistical process which generates a sequence of pixels. The problem is then cast as follows: A sequence of n pixel intensities, $S_1^n = (s_1, s_2, ...s_n)$, is assumed to have been generated by two distinct texture processes each operating on either side of an unknown change point. Thus the observed data is considered to have been produced by the following process: First a changepoint c is selected uniformly at random from the range [1-n]. Then the pixels to the left of the changepoint (the sequence S_1^c) are produced by a texture process T_1 and the pixels to the right (S_{c+1}^n) are produced by process T_2. The task is then to recover c from S_1^n. If both T_1 and T_2 are known then this corresponds to finding the c that maximises:

$$P(B_c|T_1, T_2) = P(S_1^c|T_1)P(S_{c+1}^n|T_2) \tag{1}$$

with B_c being the event of having a texture boundary at point c. $P(S_1^c|T_1)$ is given by

$$P(S_1^c|T_1) = P(s_1|T_1)\prod_{i=2}^{c} P(s_i|s_{i-1}, T_1).$$

If, however, one of the textures say T_1 is unknown, then the term $P(S_1^c|T_1)$ must be replaced by the integral over all possible texture processes:

$$P(S_1^c) = \int P(S_1^c|T)P(T)\ \mathrm{d}T \tag{2}$$

While it may be tempting to approximate this by considering only the most probable T to have generated S_1^c, this yields a poor approximation for small data sets, such as are exhibited in this problem. In [11] we have shown that the integral can be solved in closed form for reasonable choices of the prior $P(T)$ (e.g. uniform). This gives

$$P(S_1^c) = P(s_1)\prod_{i=2}^{c} P(s_i|S_1^{i-1})$$

where, for a uniform prior over T,

$$P(s_i|S_1^{i-1}) = \frac{1 + \text{num}\ (S_{i-1}^i \in S_1^{i-1})}{i + I - 1} \tag{3}$$

where I is the number of states in the Markov processes or bins in the intensity histogram, and num $(S_{i-1}^i \in S_1^{i-1})$ is number of times the substring S_{i-1}^i occurs in the string S_1^{i-1}.

Therefore, the change point can be detected by maximisation of $P(B_c|T_1, T_2)$. The conditional probability terms in Eq. 1 are given by a graylevel transition matrix W, that allows us to compute the terms num $(S_{i-1}^i \in S_1^{i-1})$, while the prior is assumed to have uniform distribution.

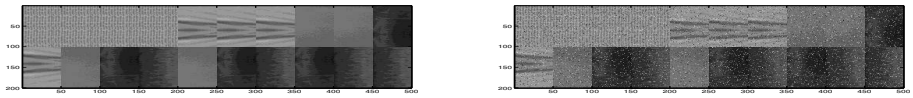

Figure 2: Texture patterns used to generate test sequences. We take the scanlines to be vertical arrays of pixels on the image. (left) original and (right) noisy images.

If the Markovian processes on both sides of a scanline are assumed unknown their statistics can be calculated using the convenient closed form solution of the integral of Eq. 2, as given by Eq. 3. On the other hand, we can assume that the texture Markov process T_2 which is in our case the texture of different body parts of the human subject is known. Therefore the transition matrix corresponding to T_2 is built once in the beginning of the tracking process and used throughout tracking. On the other hand the statistics for the Markov process T_1 are assumed unknown and are calculated using Eq. 3. This method is robust and reliable under versatile conditions where there can be cluttered patterns or non-textured objects.

As a direct extension, we can also use *scanstripes* instead of scanlines to search for texture crossings. Scanstripes are a set of parallel scanlines that are scanned simultaneously to count the number of occurrences of substring S_{i-1}^{i} in string S_1^{i-1} on each scanline independently. This helps the Eq. 3 converge faster and adds robustness to noise as long as the isotropic model is valid for the texture. This is further verified in the following subsection.

3.2 Validation of the silhouette extraction algorithm

Our algorithm was tested on the sequences generated using random textures collected from the web and stitched next to each other to form a sequence of two textures to be separated by our algorithm. The original tested textures are also shown in Fig. 2 with and without added noise to test the performance of the algorithm in presence of noise.

Fig. 3 shows two sample test sequences (without noise) and the log probability of the changepoint $P(B_c|T_1,T_2)$ for all the pixels along a sequence. The test sequences are composed of two separate textures, with the true boundary in the middle of each sequence. When applicable the transitions matrix is precalculated for the left half of the scanline using a region with similar texture. Fig. 3-right shows a case where the scanline search with unknown transition matrices has failed. This is due to the uniformity of the texture on the right side which introduces a peak in the texture crossing probability at the point where this uniformity is perturbed. Nevertheless, it can be seen that using scanstripes and precalculating the transition matrix adds robustness and prevents this kind of mistake.

The global behaviour of the algorithm was also tested using 1000 randomly selected sequences from different textures with varying lengths from 20 to 200 pixels. Fig. 4 shows the histogram of the absolute difference between the global maximum index of the log of the probability $P(B_c|T_1,T_2)$ and the true index of the boundary for different methods with different lengths used for the scanlines and scanstripes (a set of 5 scanlines). In Fig. 4 we can see two rows corresponding to noise free (first row) and noisy test images (second row). The leftmost histogram in each row corresponds to the results using known target model (transition matrix), the second histogram is obtained using known target model, and the rightmost one in each row shows the results using scanstripes instead of scanlines.

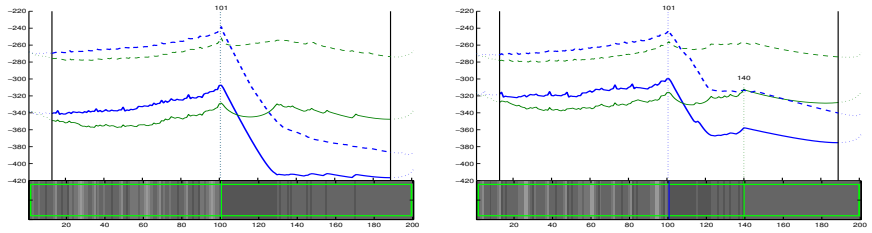

Figure 3: Log probability of the changepoint on the gray patterns shown below the graphs and where each gray level appears as a vertical bar. Probabilities are obtained using known statistics (solid bold line), unknown statistics (thin solid line), scanstripes instead of scanlines with known statistics (bold dotted line), and scanstripes instead of scanlines with unknown statistics (dashed thin line). Left: Probability maximisation gives the correct boundary index for the texture strings. We observe that the curves obtained with unknown statistics are relatively flat compared to those derived using learnt models. This shows the advantage of using known statistics. Right: Scanline search with unknown statistics fails to find the correct boundary. The scanstripes results are computed using additional scanlines not shown here.

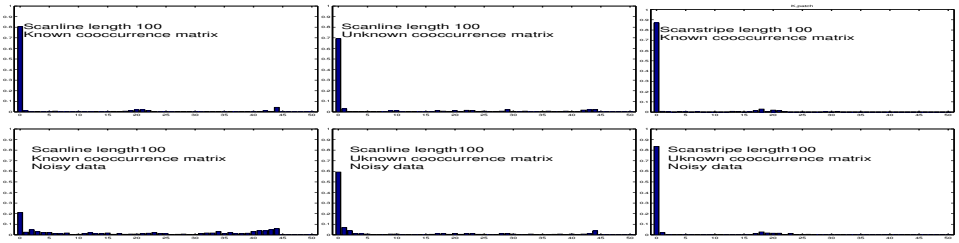

Figure 4: Histograms of the absolute difference between the detected boundary indices and the true one for 1000 tries on different test sequences with varying lengths from 20 to 200 pixels. First row is the results on noise free data while the second row corresponds to noisy textures.

These results show that the algorithm can successfully separate the textures with a high reliability factor (also in presence of noise) which increases using scanstripes instead of scanlines. It must be mentioned that assuming known statistics makes the detection more vulnerable to noise as the model is not adapted to the test sequence during the probability calculation. On the other hand when we use unknown statistics, the model adapts itself to the transition matrix of the noisy data. Due to this fact, the presented histograms for the scanstripes on the noise free image (rightmost histogram on the first row) are obtained with known statistical model, while for the noisy case (rightmost on the second row) we have assumed unknown models which adapts better. Nevertheless, these histograms support the fact that the combination of this algorithm with robust estimation techniques for tracking yields satisfactory results even in presence of noise. The length of scanline also affects the shape of the histograms. Due to lack of space we only mention that increasing the scan length introduces a more prominent peak around zero while reducing the outliers. Our simulations show that the scanlines should be at least 20 pixels long for effective detection.

4. Monocular Body Motion Tracking

Here we present a testbed framework with a multiple cue objective function in order to use our silhouette extraction method for 3–D human motion tracking. Monocular vision is challenging because it involves tackling such difficult issues as ambiguities associated with articulated motion seen from a single camera, very high dimensional search spaces, partial self-occlusions, and poor quality of image features in the absence of markers. We start with a system that uses feature points to track and we add to it contour tracking and ambiguity detection. For the latter we chose a method similar to the one described in [12].

In our implementation, each body part in the chain of articulations is modeled by a quadric (ellipsoid, cylinder or truncated cone) represented by a 4×4 matrix Q that describes both its shape and its position in space, which is a function of body poses defined in the hypothesis set. Finding the pose in the next frame is then carried out by minimising the cost associated to each hypothesis set element. The cost is a combined energy function consisting of the following terms:

$$\begin{array}{lcl} \mathscr{F} & = & \sum_i F_i^{point} + \sum_i F_i^{silhouette} \\ F_i^{point} & = & w_i^{point} \, \| \widehat{p}_i - \overline{p}_i \|^2 \\ F_i^{silhouette} & = & w_i^{silhouette} \mathrm{Dist}_{s_i}^2 \end{array} \tag{4}$$

We describe these terms in detail below.

Regions inside target (term F^{point}) Robust image point matching is used to generate point correspondences which serve as observations for projection error minimisation in the subsequent frames. In order to increase accuracy of the correspondences, we incorporate affine 2–D image motion estimation for matching in the subsequent frames. The set of matched points is further refined by robust fundamental matrix computation and enforcement of epipolar geometry constraints on each body part separately. This provides the first term in the energy function given by Eq. 4, where $\overline{p}_i$ is a matched point in frame t, and $\widehat{p}_i$ is the estimated projected point using model pose corresponding to frame t.

Target's outline (term $F^{silhouette}$) The detected body silhouette points obtained using the method explained in Section 3 provide distances Dist_{s_i} of the cast ray passing through a given silhouette image point $s_i = (u', v')$ from the associated quadric Q on the model in 3–D space. These distances are added to the energy function to be minimised. These terms along with the terms corresponding to the image point observations are robustly weighted using M-estimator method to exclude outliers (due to noise or local incompatibility of the learned texture) are normalised to have balanced impact on the energy function. We chose to use the algorithm with known transition matrices for each body part. Fig. 5 shows the detected texture boundaries for the sampled quadric silhouettes. The search starts from samples on the projected model outline in the previous position. This starting point is not necessarily close to the real outline. This is due to the fact that the model is an approximation of the real subject and moreover its pose comes from the previous frame. Our method is nevertheless able to detect the texture boundaries in an reliable way.

In practice not all possible joint angle values are acceptable and we constrain them to remain within an acceptable range. For each such constraint, we add a penalty term which is nonzero if the parameter goes beyond the limits. Another plausibility factor is the smoothness of motion. We suppose that the human motion is reasonably steady in a short

(a) (b) (c) (d)

Figure 5: Examples of the detected silhouette points on the body (a, b, c). White circles are the detected texture boundaries corresponding to the body outlines. (d) illustrates the training phase for Markov texture modeling, the white patches show the areas used to compute the model for each body part. This is done by rendering the initial pose given by manual initialisation and obtaining the mask of its projection on the corresponding frame.

frame interval. That implies that body parts do not tend to change direction abruptly. This gives the last term in the energy function of Eq. 4 where θ_i^t is the joint angle i at frame t. In our implementation observations from three adjacent frames are used to minimise this objective function for each hypothesis. The resulting pose are iteratively used in the same manner by sliding the optimisation window over the sequence.

5 Tracking Results

Here we present some results obtained using our proposed method on some monocular sequences. The statistical model of the texture of the subject is constructed in the first frame in terms of a transition matrix for each body part by considering the pixels which lie under the projection of the body pose on the corresponding frame which comes from manual initialisation. This is usually enough to approximate the statistics of the model of the subject provided that the subject's appearance does not change dramatically through out the sequence. The model statistics can be easily updated once in a while in the cases where this condition is not guaranteed. The patches used for this training phase are shown in Fig. 5(d).

As an example of our experiments with real scenes and human motions, next we consider a sequence of walking toward the camera which is inherently difficult due to considerable change in target's dimensions. The first row of Fig. 6 shows some shots of the original 100 frame sequence with superimposed tracking results. The results are further verified by resynthesising the scene as seen by an arbitrary camera. The second row shows the resynthesised view of the tracking results from frontal top, which evidently show a forward motion of the model towards the camera. We have further tested the stability of the results by augmenting the frames obtained by a second camera that was used only for verification and not during tracking as shown in the last row. We compared our algorithm for silhouette detection with a gradient-based method and Fig. 8 shows part of the body and the silhouettes extracted using texture-based (left) and gradient-based (right).

The second example is tracking of a person wearing a highly textured top where as there is almost no texture on the arms. Moreover the background is highly cluttered and contains a moving person. Fig. 7 shows some frames of tracking results with the

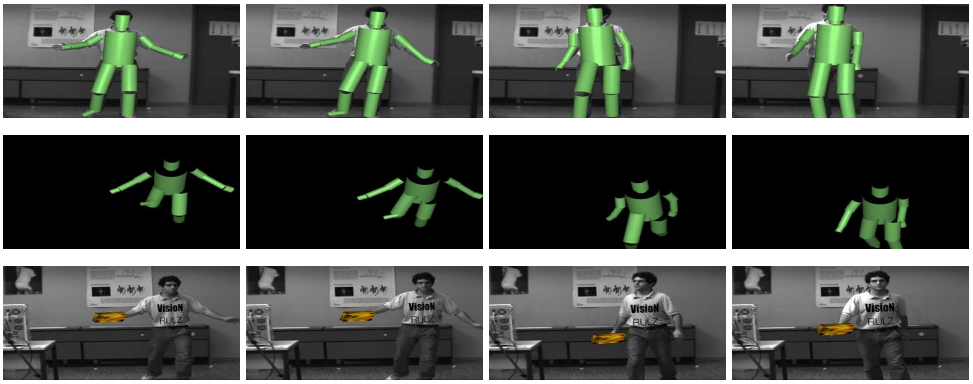

Figure 6: Tracking results for some 100 frames of a walking sequence towards the camera.The first row shows superimposed model on several frames used for tracking. The second row shows the resynthesised view of the tracking results seen from frontal top. The last row validates the tracked 3–D pose by augmenting the frames obtained by a second camera that was not used for tracking.

Figure 7: Tracking under extreme conditions. The subject is wearing a highly textured top whereas her arms lack texture. The background is highly cluttered and contains a moving person.

superposed model. The extracted silhouettes (see Fig. 8-left) are less clean compared to previous results (due to shadows around the shoulders and high clutter and fluctuations in the complexity of texture, moving background, etc.) but the results are far more credible than the ones obtained using a gradient-based method as shown in the right side of Fig. 8. In terms of processing time, the algorithm is quite fast and it takes a small fraction of a second to extract the silhouette points from a given frame on a standard PC.

6 Conclusion

In this paper we demonstrated the applicability of a texture-based approach to finding silhouettes in the context of monocular human body tracking. We have shown that it allows us to obtain good results under difficult conditions due to highly textured body, lack of texture on body parts, loose clothing and cluttered and moving background.

The silhouettes are estimated by projecting the model's current configuration into the current frame, computing first order statistics along lines perpendicular to those contours and finding texture boundaries between object and background. While conventional gradient-based techniques are heavily used for outline detection, they usually fail under everyday life situations. Our proposed algorithm provides a reliable alternative under

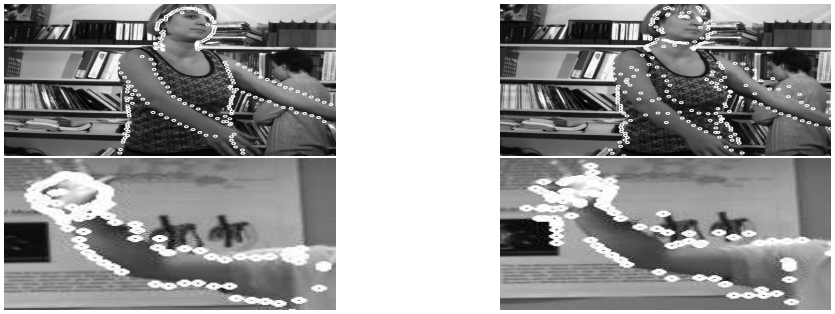

Figure 8: Examples of silhouettes obtained using texture-based proposed method (left) compared with those obtained by a gradient-based method (right).

these difficult conditions. Moreover it remains fast and can potentially lead to real-time applications.

References

[1] Ankur Agarwal and Bill Triggs. 3d human pose from silhouettes by relevance vector regression. In *Conference on Computer Vision and Pattern Recognition*, 2004. to appear.

[2] Vassilis Athitsos and Stan Sclaroff. Estimating 3d hand pose from a cluttered image. In *Conference on Computer Vision and Pattern Recognition*, pages 432–439, Madison, WI, June 2003.

[3] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive gmmrf model. In *European Conference on Computer Vision*, volume 1, pages 428–441, 2004.

[4] J.W. Davis and A.F. Bobick. A robust human-silhouette extraction technique for interactive virtual environments. In *Workshop on Modelling and Motion Capture Techniques for Virtual Environments*, pages 12–25, Geneva, Switzerland, November 1998.

[5] J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *Conference on Computer Vision and Pattern Recognition*, pages 2126–2133, Hilton Head Island, SC, 2000.

[6] T. Drummond and R. Cipolla. Real-time visual tracking of complex structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):932–946, july 2002.

[7] D.M. Gavrila and L. Davis. 3d model-based tracking of humans in action : A multi-view approach. In *Conference on Computer Vision and Pattern Recognition*, pages 73–80, San Francisco, CA, June 1996.

[8] I.A. Kakadiaris and D. Metaxas. Model based estimation of 3d human motion with occlusion based on active multi-viewpoint selection. In *Conference on Computer Vision and Pattern Recognition*, page 81, San Francisco, CA, June 1996.

[9] Anurag Mittal, Liang Zhao, and Larry S. Davis. Human body pose estimation using silhouette shape analysis. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03)*, page 263, Miami, Florida, 2003.

[10] E. Ozyildiz. Adaptive texture and color segmentation for tracking moving objects. Master's thesis, Pennsylvania State University, 1999.

[11] A. Shahrokni, T. Drummond, and P. Fua. Texture Boundary Detection for Real-Time Tracking. In *European Conference on Computer Vision*, pages Vol II: 566–577, Prague, Czech Republic, May 2004.

[12] N. Shimada, Y Shitirai, Y. Kuono, and J. Miura. Hand Gesture Estimation and Model Refinement using Monocular Camera – Ambiguity Limitation by Inequality Constraints. In *Automated Face and Gesture Recognition*, pages 268–273, Nara, Japan, 1998.

[13] C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular 3D Human Tracking. In *Conference on Computer Vision and Pattern Recognition*, volume I, page 69, Madison, WI, June 2003.

[14] A. Thayananthan, B. Stenger, P.H.S. Torr, and R. Cipolla. Tracking Articulated Hand Motion using a Kinematic Prior. In *British Machine Vision Conference*, pages 589–598, Norwich, UK, 2003.