Human Body Part Labeling and Tracking Using Graph Matching Theory

Nicolas Thome
n.thome@foxstream.fr
Sas Foxstream*

Djamel Merad
Djamel.Merad@univ-lyon2.fr

Serge Miguet
Serge.Miguet@univ-lyon2.fr

Laboratoire d'InfoRmatique en Images et Systèmes d'information
Université Lumière Lyon 2
5 Avenue, Pierre Mendès France
69576 Bron Cedex, FRANCE

Abstract

Properly labeling human body parts in video sequences is essential for robust tracking and motion interpretation frameworks. We propose to perform this task by using Graph Matching. The silhouette skeleton is computed and decomposed into a set of segments corresponding to the different limbs. A Graph capturing the topology of the segments is generated and matched against a 3D model of the human skeleton. The limb identification is carried out for each node of the graph, potentially leading to the absence of correspondence. The method captures the minimal information about the skeleton shape. No assumption about the viewpoint, the human pose, the geometry or the appearance of the limbs is done during the matching process, making the approach applicable to every configuration. Some correspondences that might be ambiguous only relying on topology are enforced by tracking each graph node over time. Several results present the efficiency of the labeling, particularly its robustness to limb detection errors that are likely to occur in real situations because of occlusions or low level system failures. Finally the relevance of the labeling in an overall tracking system is described.

1. Introduction

Tracking humans in video sequences has been extensively studied among the computer vision community. The methods may be classified into the following groups: Motion-based, Model-Based, Appearance-Based and Feature-Based. In [13], an hybrid approach is proposed. Each detected object is indeed tracked with a region-based strategy. As clear evidence for single object association is determined, an articulated appearance model is generated and dynamically updated. It provides a discriminative feature that is used to perform recognition in difficult situations such as occlusions. The appearance model generation requires a partition of the human silhouette in order to localize the limbs. The main limitation in [13] corresponds to the assumption about the pose for the body part labeling. People are indeed supposed to be in an upright standing posture. We propose here an approach dedicated to locate and identify visible body parts in the image, that is independent on the viewpoint and the human pose.

2. State Of the Art

There has been plenty of works on the subject of detecting and tracking human limbs. For an exhaustive review, the reader can refer to [1]. Approaches using 3D models try to find correspondences by minimizing some image-to-model criterion [10]. They suffer from being computationally expensive and often only consist in tracking by requiring a manual initialization. Approaches using 2D models aim at locating articulated structures in the image. They may be decomposed into top-down and bottom-up [6, 7] strategies. The former ones try to find the image location best corresponding to a given template encoding the model properties. To be efficient and general, the search must be performed at different levels of scale and orientation. Moreover, they often rely on appearance, which is too restrictive for our purpose. Bottom-up approaches process in an opposite manner. in a first time, body parts candidates are located in the image, based on low level image features. Then, only a subset of the candidates is kept, discarding outliers.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

This is performed by using a human model dedicated to enforce a consistent assembly of the different limbs. Other approaches first segment the human silhouette before analyzing its shape properties to extract the body parts [2, 4] . Our approach belongs to this class. In [2], Haritaoglu et al. perform the labeling by first determining the pose among a set of predefined ones. However, this preprocessing scheme is inevitably prone to fail in some cases, decreasing the overall system performances. The approach proposed by Mori and Malik in [4] identically retrieves the human pose before performing the labeling. Among a pre-stored set of exemplar 2D views for which key points are manually identified, they first determine for the test shape TS the best match in the shape context meaning, before transferring the key points to TS. In the proposed method, we label body parts from the silhouette without any assumption about the pose or the viewpoint. We want our method to only rely on shape, without any reference to other features such as geometrical ones stored in the configuration. In addition, the posture may be determine as a body parts identification result.

3. Approach Overview

Our algorithm processes as follows. First, we identify in the image several segment sets, corresponding to the body parts to identify (see Section 3.1). Then, a graph encoding the silhouette shape is generated and matched against a 3D model of the human skeleton (section 3.2). Finally, each matched graph node is tracked over time (Section 3.3).

Our main contributions for the purpose of body parts labeling are :

- Encoding in the graph structure the silhouette properties in the most compact form. In particular, we ignore all features depending on the viewpoint, the human pose or the geometry and we rely only on shape and topology.
- Using an efficient graph matching strategy to find the best correspondences between the image and model graphs.
- In ambiguous correspondences using tracking to enforce a coherent solution. Thus additional features are only integrated when strictly required.

3.1. Limbs Detection

The body part labeling is performed on the detected silhouettes corresponding to single humans in each frame. These image regions are localized thanks to a motion segmentation algorithm followed by a region-based association.

3.1.1 Silhouette Extraction and Tracking

The first step of the system consists in applying a motion segmentation algorithm, leading to a binary map where moving and static pixels are labeled. This is achieved by modeling the background for each pixel by a mixture of Gaussians (first in introduced in [11]). In addition, we make it possible to not assign the "moving" label to shadow pixels by using a color space invariant in luminance. Finally a connected component is applied to get sufficiently large regions where motion occurs. Then, a simple region-based tracking strategy is developed to match regions detected in one frame and in the subsequent ones. In particular, it enables us to detect regions corresponding to single humans. These regions are robustly tracked over time by using an articulated appearance model that constitutes a feature used to perform matching in difficult situations. For further details about this part of the approach, the reader is referred to [13].

3.1.2 Skeleton computing

For each "single human" detected region, the limbs are considered as parts of the silhouette skeleton. The first step in order to detect visible segments corresponding to body parts in the image consists thus in determining the skeleton points. Whatever the strategy used, the main difficulty related to the skeleton computation corresponds to its sensitivity to noise. To overcome this shortcoming, we first smooth the silhouette. This is achieved by computing the Fourier Descriptor of its outer contours.

The Fourier Descriptors provide a discriminative signature of the contour of an object ([15, 5]). The $A(k)$ coefficients are determined by the computation of the DFT (Discrete Fourier Transform) for the N contours points considered as complex numbers $X(i)$:

$$A(k) = \frac{1}{N} \sum_{i=0}^{N-1} X(i) e^{-j2\pi ki/N} \qquad (1)$$

The coefficients $A(k)$ represent the discrete contour of a shape in the Fourier (frequency) domain. The general shape of the object is represented by the lower frequency descriptors, whereas high frequency coefficients capture details of the object. Thus, smoothing the silhouette is carried out by only keeping a subset of the lowest frequency descriptors. At this stage, the skeleton is determined by computing the Delaunay triangulation of the smoothed reconstructed silhouette. This approach is the most adapted to our purpose for the following reasons. First, the computation is fast and accurate. Moreover, the Delaunay triangle structure is isomorph to a graph by containing neighborhood information. This clearly facilitates the graph generation process (see 3.2).

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

3.1.3 Getting A Set of Segments

The skeleton point sequences is then poligonalized. This step consists in identifying a set of N points P_i, $i \in [1; N]$ and the link between them, representing the $N-1$ segments. We point out that each skeleton point corresponds to the center of the circumscribed circle to each Delaunay triangle. To each link between P_i and P_j ($i \neq j$) is associated a quantity M_r corresponding to the mean radius of the segment along the skeleton points and computed by the following way : $M_r = \frac{1}{M} \sum_{i=1}^{M} r_i$, where M is the number of skeleton points between P_i and P_j and r_i corresponds to the radius of the i^{th} point. We come back in section 3.2.2 on the use of M_r for the purpose of the graph generation.

Skeleton points may be classified depending on their neighborhood degree. Points having a single neighbor (S) correspond to end points. Points having more than two neighbors (M) define starting points for segments. Points having exactly two neighbors (C) corresponds to points on a continuous curve between (M) and (S) points.

We can notice that (S) and (M) skeleton points must be end points of segments corresponding to body parts. In addition, some (C) points are also likely to belong to the P_i set. Thus, we split each (C) sequence into k segments, so that the mean curvature for the corresponding skeleton points is 0 (in practice under a given small threshold). Figure 1 illustrates the skeleton computation. Figure 1a) represents an extracted silhouette, figure 1b) the skeleton computation (with (S) points in yellow, (M) points in green), figure 1c) the first set of segments after the poligonalization and in 1d) the last set after removing small edges.

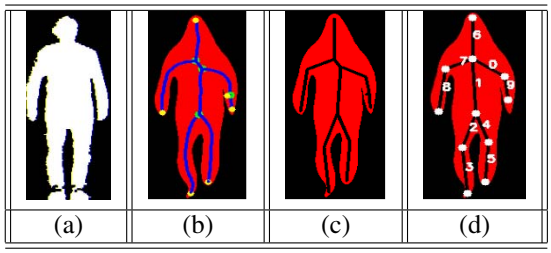

Figure 1. Skeleton Computing

3.2. Body Parts Labeling

3.2.1 Chosen Representation

At this stage, a set of points linked by edges has been extracted from the silhouette. They correspond to candidates for the different body parts. We propose to match this processed 2D skeleton to a 3D model of human segments. The figure 2a) presents the structure of the chosen model. The strength of the skeleton representation for the purpose of limb labeling relies on the fact that its topology is independent of the dimension. Thus, the connexity between points or edges contain the same information in the 2D image plane and in the 3D world, and trying to find a matching is meaningful. Moreover, it is easy to transform the skeleton segment sets into a graph. It enables us to take advantage of the results on the graph matching theory, which has been extensively studied in the last decades and proved its efficiency in the computer vision community.

3.2.2 Graph Generating

From the skeleton segment sets we build a Directed Acyclic Graph (DAG), capturing the topology of the silhouette. As we consider only the outer contour, the graph is actually always a tree. For a maximum of clarity, we will always refer to the graph by using the vocabulary of nodes and arcs, as we will use vertices and edges for the set of segments in section 3.1.3. There are two major questions to answer for generating the graph. First, we have to choose whether the graph nodes correspond to the edges or to the vertices of the segments set. Second, we must choose a root for the graph.

We make the choice of using segment vertices as nodes and decide to root the model tree in node 0 (blue), corresponding to the bottom of the torso. The justification for these decisions follows. Once defined in the model, the root has to be identified in the image at each time step. Thus, the root choice is influenced by the possibility to extract image features that enable us to robustly localize it. For that reason, we choose to detect the blue edge $(0-1)$ (as shown in figure 2) in the image, because it is robustly identified as the one having the largest mean radius M_r as defined in section 3.1.3. Intuitively, the torso edge must be the limb whose mean distance to the silhouette contour is maximum. Now, the reason why we do not use edges as graph nodes is simple : rooted at the torso, it is not possible to distinguish arms and legs from the graph structure (the neighborhood is encoded in the same manner). We thus use the vertices as nodes to compute the graph. The vertex corresponding to the node 0 has finally to be identified in the image. To perform this task, we generate the two possible kinds of graphs from the two image vertices of the root segment. We then compute the topological signature (see 3.2.2.1) of the roots in the two graphs, and find the one being the closest to the graph model in figure 2b). We weight this topological $D_T(G_i, G_M)$ distance with another one related to the feature vector $D_F(G_i, G_M)$ attached to each node during tracking (see 3.3), to be robust to largely occluded graphs, defining the global distance as $D(G_i, G_M) = w_T * D_T(G_i, G_M) + w_F * D_F(G_i, G_M)$.

3.2.2.1 Encoding Graph Structure A powerful way to encode the structure of a DAG consists in turning to the domain of spectral graph theory (see [9]). Any directed graph can indeed be represented as an antisymmetric $0, 1, -1$ ad-

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

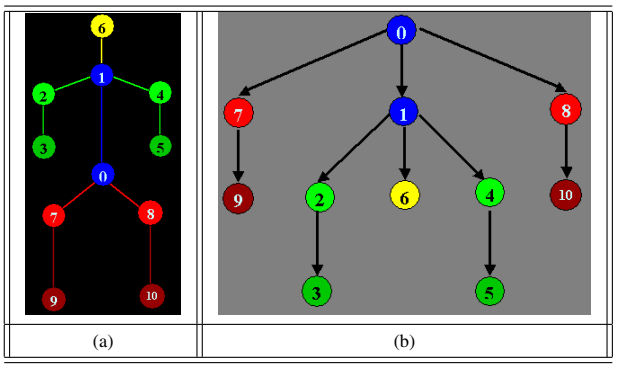

Figure 2. The 3D models

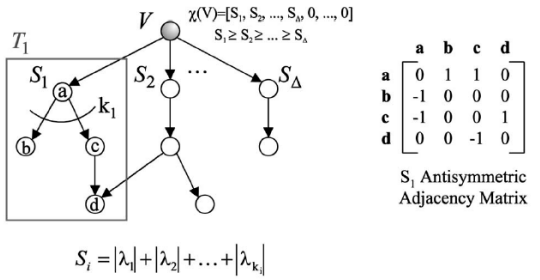

Figure 3. Topological Signature Computation

jacency matrix AG, with $1s$ ($-1s$) indicating a forward (backward) edge between adjacent nodes in the graph (and $0s$ on the diagonal). For a graph G with adjacency matrix AG, we define the spectrum $\Gamma(AG)$ as the set of magnitudes of its n eigenvalues.There are many advantages in using the spectrum representation for describing the DAG structure. $\Gamma(AG)$ encodes important structural properties of the graph, including its size and the degree distribution of its nodes. Moreover, results on spectral graph theory have established its stability against minor perturbation due to noise, occlusion, or node split/merge (see [8] for more details). As noted in [9, 8, 3], describing the graph structure consists in computing a Topological Signature Vector (TSV) $X(N_i)$ for each graph node (N_i). $X(N_i)$ corresponds actually to the sum of the eigenvalues magnitudes for each child of the node. In the context of graph matching the TSV of two different graphs must have the same size so that the distance computation (see 3.2.3) is meaningful. Thus each $X(N_i)$ is initialized as a N-sized vector, where $N = max(MaxDegree(G1), MaxDegree(G2))$. The computation precesses as follows : for each graph node N_i having k children C_j ($j \in [1;k]$), a subgraph whose root corresponds to each C_j is generated. Then the adjacency matrix AG is determined and the spectrum $\Gamma(AG)$ is computed. The eigenvalue sums correspond to the j^{th} element of the TSV for N_i. Finally the TSV is sorted in decreasing order. The figure 3 illustrates the computation of the TSV.

3.2.2.2 Initialization As N, the dimension of the TSV is for the majority of the nodes larger that the number of their children, the TSV is padded with zeros in [9, 3]. However, as noted in [8], this solution does not make it possible to discriminate terminal nodes T from nodes having an unspecified number of children that are terminals BT. The reason for this is that the 0 eigenvalue magnitude sum of a leaf node is indistinguishable from the padded 0 of the TSV. To overcome this shortcoming, it is suggested in [8] to add as extra dimension the eigenvalue magnitude of the root node subgraph. We propose here an alternative approach consisting in initializing all the TSV's with -1 values. The number of 0 values for a BT node makes it possible to determine its number of terminal node. Thus, this representation captures the same information as the eigenvalue magnitude (because we are close to the leaves of the graph) and the TSV dimension is not increased.

3.2.3 Graph Matching

Algorithms for matching two graphs G_I and G_M corresponding to skeletons can be found in [14, 9]. The aim is to find a path among the two graphs for which each node correspondence is optimal. The roots having been matched (3.2.2), the algorithm process as follows. Two subgraphs S^1G_I and S^1G_M starting from the roots are generated. The best matched (I_1, M_1) as defined in 3.2.3.1 inside the subgraphs is determined. The process is then recursively applied to the subgraphs S^2G_I and S^2G_M whose roots are I_1 and M_1, respectively. This depth-first search strategy ends as soon as a leaf of one of the graph is reached. Then, the backward step of the algorithm starts.

3.2.3.1 Best Match Determination At each algorithm iteration, a matching Matrix $\Pi(S^iG_I, S^iG_M)$ between the two subgraphs S^iG_I and S^iG_M is computed. Each of its element corresponds to the euclidean distance between S^iG_I and S^iG_M TSV's, and we denote it as the matching weight. Thus, we seek for a minimal values in $\Pi(S^iG_I, S^iG_M)$. If several candidates are at the same score, we rank them in descending order depending of their cardinality. The cardinality for a node N_i corresponds to the number of nodes in the subgraph rooted in N_i. This condition makes it possible to support candidates whose subgraphs will be the largest and thus to facilitate the forward process. If multiple matching possibilities still appear, it points out the fact that the information encoded in the graph structure is not sufficient to conclude. Additional features are then required to take

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

a decision. Usual configurations where it is likely to occur and solutions proposed to overcome them are detailed in 3.2.3.2.

3.2.3.2 Ambiguities Resolving At each node matching step, ambiguities remain if there are k equivalent correspondences (N_I, N_M) (same weight and cardinality), with $k > 1$. For our body part labeling application, it occurs in two main kinds of situations. First, head and arms are indistinguishable from the graph if a single segment is extracted for the arms. Second, right/left ambiguities between arms and legs are usually present.

We propose to solve the first one by the following geometric reasoning. For each matching performed with the node of the head model (node 6 of figure 2), we compute for each matched node N_I (whose father is F_{NI}) in the image the projection P_R onto the root computed has $P_R = \overrightarrow{N_i - F_{NI}} \cdot \overrightarrow{N_1 - N_O}$. The head is supposed to be the node for which P_R is maximum. This assumes that the head and torso slopes are close to each other (what is a reasonable hypothesis whatever the configuration), while arms and torso slopes are not. Finally, the head is supposed to not have been detected in the image if the maximum P_R value is below a given threshold. The right/left ambiguities can not easily be resolved with static image features. To descriminate these cases, we use the temporal information stored in the tracking framework. The next section is dedicated to this study.

3.3. Body Parts Tracking

As the human silhouette is extracted initially, a feature vector corresponding to each image node is initialized. If we have to face equivalent correspondence pairs, the choice is done randomly. The feature vector associated with each node stores simple information of position and velocity. As long as the same human is tracked, each node N_k in the next frame F_{t+1} is tracked against all nodes N_i in the previous one F_t. The tracking is performed in the following simple way. All N_i are projected into the next frame assuming a constant speed motion model. Then a distance feature is computed between each pair of nodes N_i and N_k (located at P_i and P_k, respectively) as : $D_F(N_k, N_i) = ||P_i - P_k||$. The matching retrieves the correspondence whose distance is minimal. This tracking strategy is particularly useful and efficient to disambiguate the potential multiple matches resulting from left/right symmetry of the human body (see 3.2.3.2).

4. Results

We present here some results illustrating the efficiency of the proposed limb labeling strategy.

Figure 4 focuses on results corresponding to the graph matching part of the system. The coloring convention used is related to figure 2 : Head is drawn in yellow, torso in blue, arms in green and legs in red. Note that the distinction between the two potential segments in arms and legs are illustrated with a difference of intensity. These results prove the approach ability to manage unspecified viewpoints or human postures, leading to a proper labeling in each case. In figures 4a) and 4b), the body part identification is presented for a standing posture with a back view and a side view, respectively. We can notice that the head labeling is properly performed in both cases although arms are formed by a single segment. This is a result of the geometric reasoning presented in 3.2.3.2. Figures 4c), 4d) and 4e) show the results for sitting, falling and lengthened poses, respectively. It demonstrates the robustness of the root (torso) detection with the maximum mean radius criterion defined in 3.1.3. Note that the head is localized in (c) and (e) as being the segment whose slope is the closest to the root, whereas in (d) the graph topology is sufficient to conclude as the two segments for the arms are detected. Finally figure 4f) shows an example where someone is walking on the hands.

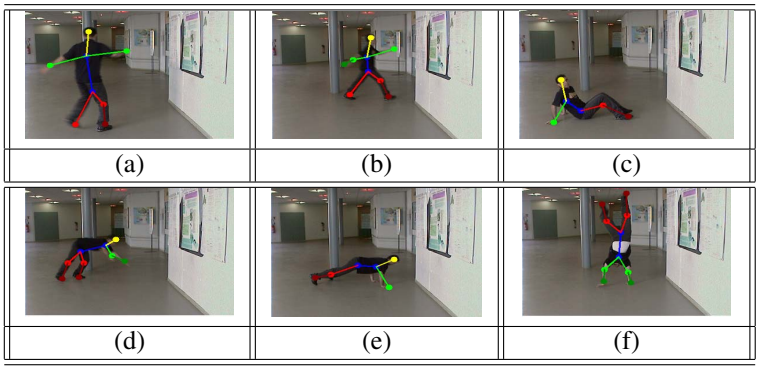

Figure 4. Labeling Results in Various Poses and Viewpoints.

Figure 5 presents results for which a significant number of silhouette image segments is missing (at least one branch starting from one of the two root nodes is removed). This can be due to various elements. There may be partial occlusions in the image (Figures 5a) and 5b)). There can also be auto occlusions of the limbs, making it impossible to detect them in the silhouette (Figures 5c), 5d) and 5e)). Finally, this can be due to the absence of inner contour extraction (Figure 5f)). Solutions for improving this are suggested in section 5.

Finally figure 6 shows the results of the tracking part of the system. At the top/left frame, the person is tracked and graph matching is performed. The symmetric indistinguishable left/right limbs are then initialized randomly. After that, each ambiguous configuration is checked against the tracked nodes, making it possible to enforce a unique co-

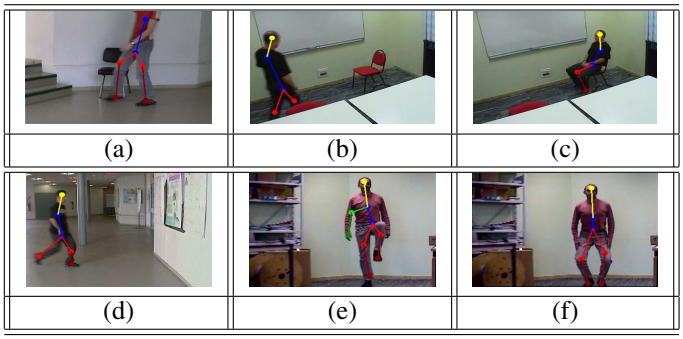

Figure 5. Robustness to limb occlusion.

herent labeling over time.

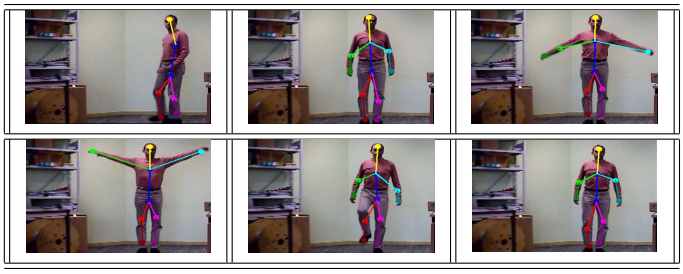

Figure 6. Tracking Results.

5. Conclusion And Future Works

We propose an original approach dedicated to limb labeling. After background subtraction, the silhouette shape is captured by generating a graph from the skeleton. A graph matching algorithm relying only on topology is used to identify the different body parts. Moreover, to enforce the matching in ambiguous conditions, each graph node is tracked over time. Applied in the context of an overall tracking system [13], this labeling strategy makes it possible to build and update the appearance model for unspecified postures and viewpoints. The direction for future work is as follows. Firstly, an easy improvement would be to make the method manage inner contours as well as outer ones. Secondly, tracking the limbs may be made more robust. At the current time, this step is essentially dedicated to provide additional information for enforcing the limb labeling. It could be interesting to use each body part appearance to improve the tracking performances. It includes occlusion detecting, and the ability to keep locating limbs during self-occlusions (arms and torso for example). Thirdly, as the body parts are labeled, it could be interesting to retrieve the 3D relative configuration of the image skeleton. A lot of works initiated in [12] and assuming an image body part labeling, have been performed in that purpose. This step could make it possible in the future to build a 3D appearance articulated human model, providing an efficient feature for matching that is invariant to viewpoint.

References

[1] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer vision and Image Understanding*, 75, 1999.

[2] I.Haritaoglu, D.Harwood, and L.Davis. Ghost: A human body part labeling system using silhouettes. *Fourteenth International Conference on Pattern Recognition, Brisbane*, 8 1998.

[3] D. Merad, J. Didier, and M. Scuturici. Tracking 3d free form object in video sequence. *Third Canadian Conference on Computer and Robot Vision, IAPR-CRV 2006, Quebec, June*, 2006.

[4] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV (3)*, pages 666–680, 2002.

[5] E. Peerson and K. Fu. Shape discrimination using fourier descriptor. *Trans. Syst., Man, Cybern., vol. SMC-7, no. 3, pp. 170179*, 1977.

[6] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego*, volume 1, pages 271–278, 2005.

[7] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *Proc. 10th Int'l. Conf. Computer Vision*, volume 1, pages 824–831, 2005.

[8] A. Shokoufandeh, D. Macrini, S. Dickinson, K. Siddiqi, and S. W. Zucker. Indexing hierarchical structures using graph spectra. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[9] K. Siddiqi, A. Shokoufandeh, S. Dickinson, and S. W. Zucker. Shock graphs and shape matching. *International Journal of Computer Vision, 30, 1-24,*, 1999.

[10] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3d body tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA*, 1:447–454, Dec 2001.

[11] C. Stauffer and W. Grimson. Learning patterns of activity using real-ime tracking. *IEEE Trans. on Patt. Anal. and Machine Intell. vol. 22 no. 8 pp. 747-757*, 2000.

[12] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding: CVIU*, 80(3):349–363, 2000.

[13] N. Thome and S. Miguet. A robust appearance model for tracking human motions. *AVSS*, September 2005.

[14] S. W.Reyner. An analysis of a good algorithm for the subtree problem. *SIAM Journal of Computing, 6, 4, 730-732*, 1977.

[15] T. Zahn and R. Roskies. Fourier descriptors for plane closed curves. *IEEE Trans. Comput., vol. C-21, no. 3, pp. 269281*, 1972.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE