AUTOMATIC HUMAN MOTION ANALYSIS AND ACTION RECOGNITION IN ATHLETICS VIDEOS

Costas Panagiotakis, Ilias Grinias, and Georgios Tziritas

Department of Computer Science, University of Crete
P.O. Box 2208,71409, Heraklion, Greece
phone: + (30) 2810 393517, fax: + (30) 2810 393501, email: {cpanag, grinias, tziritas}@csd.uoc.gr

ABSTRACT

We present an unsupervised, automatic human motion analysis and action recognition scheme tested on athletics videos. First, four major human points are recognized and tracked using human silhouettes that are computed by a robust camera estimation and object localization method. Statistical analysis of the tracking points motion obtains a temporal segmentation on running and jump stage. The method is tested on athletics videos of pole vault, high jump, triple jump and long jump recognizing them using robust and independent from the camera motion and the athlete performance features. The experimental results indicate the good performance of the proposed scheme, even in sequences with complicated content and motion.

1. INTRODUCTION

Human motion analysis using computer vision techniques has many applications in many areas, such as analysis of athletic events, surveillance, entertainment, user interfaces, content-based image storage and retrieval. These systems attempt to detect, track and identify people and recognize their action given a number of predefined actions. Thus, there has been a significant number of recent papers on human tracking and activity recognition. We can classify these systems into different categories, according to the input data, the assumptions adopted, the method used and the output. Wang, Hu and Tan [10] emphasize on three major issues of human motion analysis systems, namely human detection, tracking and activity understanding. According to them, there are 2D, with or without explicit shape models, and 3D approaches.

First, we consider 2-D approaches. Wang et al. [11] propose a method to recognize and track a walker using 2D human model and both static and dynamic cues of body biometrics. Moreover, many systems use Shape-From-Silhouette methods to detect and track the human in 2D [6] or 3D space [2]. The silhouettes are easy to extract providing valuable information about the position and shape of the person. When the camera is static, background subtraction techniques can give high accuracy measures of human silhouettes. Otherwise, camera motion estimation methods [3] can locate the independently moving objects.

Several approaches have been proposed recently in the literature for detecting video actions and activities using 2D or 3D motion captured data. Bodbick and Davis [1] use temporal templates strategy. They interpret human motion in an image sequence by using motion-energy (MEI) and motion history images (MHI). Mori et al. [4] use 3D motion data and associate each action with a distinct feature detector and HMM, followed by hierarchical recognition. In [5], the action recognition is performed using a probabilistic context-
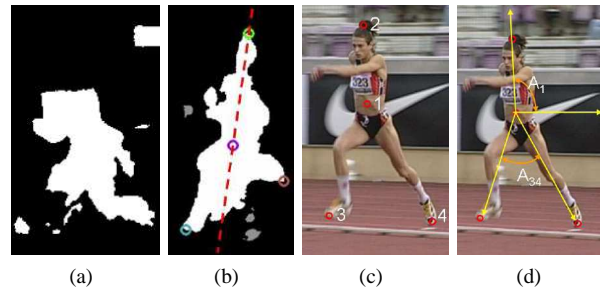

Fig. 1: (a) Low quality silhouette: low accuracy human boundary, the silhouette could be partitioned to several segments and several objects could be appeared. **(b)** Estimated human points: head center (green point), mass center (magenta point), left end of leg (blue point) and right end of leg (brown point). The human body major axis is shown as a red dashed line. **(c)** The four major human points. **(d)** The two characteristics angles: the human major axis angle (A_1) and the angle between legs (A_{34}).

free grammar (PCFG) based on an automatic keyframe selection process.

Most of them consider simple classes like running, walking and standing using as input video sequences from static camera and controlled environments. Thus, they obtain high accuracy measurements about human silhouettes and high performance results. A challenging problem appears when the camera is moving and the estimated human silhouettes are of low quality or extremely wrong (see Fig. 1(a)). In this work we focus on automatic human detection, tracking and action recognition under real and dynamic environments of athletic meetings. We suppose that the camera tracks the athlete and we the test algorithm in sports like pole vault, high jump, triple jump and long jump. Furthermore, our method works when other humans appear in the scene. The main contribution of the method is that it works automatically without any initialization or prior knowledge about camera motion and human parameters, providing also statistical results about athlete motion. Moreover, the proposed, robust and independent from the camera motion and the athlete performance features, obtain a high performance action recognition method.

1.1 System Overview

The proposed architecture consists of two main modules. First, four major human points are recognized and tracked using the precomputed human silhouettes. Silhouettes are computed using a general purpose algorithm for detecting and localizing the moving objects of videos. In general, the

basic steps of the algorithm are camera motion estimation, change detection and label propagation based on Bayesian statistics. The human major axis, the gait period and a temporal segmentation on running and jump stage are estimated by statistical analysis of the tracking points motion. On the second module, the action recognition task is performed using the above features. The input video could be from sports like pole vault, high jump, triple jump and long jump.

The rest of the paper is organized as follows. Section 2 presents the human motion analysis module. Section 3 describes the action recognition module. Finally, Sections 4 and 5 provide experimental results and the discussion, respectively.

2. HUMAN MOTION ANALYSIS

In this section, we describe the methods that detect and track the major human points. Moreover, we define useful action recognition features which are related to the human motion.

2.1 Moving Objects Localization

The overall method relies on athletes' silhouettes, extracted for each frame of the sport event video. Silhouettes are computed using the objects localization framework described in [8]. Moving objects detection is mainly based on change detection between successive video frames. The change detection problem is modelled by the mixture of two zero-mean Laplacian distributions, which correspond to pixel classes "static" and "mobile". An Expectation Maximization (EM) method is employed to fit the mixture model to the computed histogram of pixel inter-frame differences. The estimated by EM model parameters, are then used to label "static"/"mobile" pixels of high confidence to the class they belong. The remaining pixels are labeled by the Multi-label Fast Marching algorithm [9], using a propagation velocity which is based on Bayesian statistics of the mixture model and the local labeling information of neighboring pixels. Post processing procedures usually follow, in order to improve the localization of the detected objects. Thus, the final result of the localization framework is a map of the "mobile" objects of the scene.

Since the localization framework is applicable to videos where the camera remains static, the camera motion of sport event videos is robustly computed and appropriately subtracted, as it is described in [3]. The 2D motion field is given by equations:

$$u = \alpha_1 + \beta x + \gamma x^2 + \delta xy$$

$$v = \alpha_2 + \beta y + \gamma xy + \delta y^2$$

where β is the zoom factor, γ, δ are the quadratic parameters which are introduced in the 2D motion field by camera pan and tilt respectively and α_1, α_2, are the corresponding 2D translational parameters. This 2D model refers to a rotating camera with a possibly changing focal length, as exactly is the case in videos of many sport events. The motion estimation method is based on block matching, confidence measure computation and M-estimation and is fast and robust, leading to very good moving objects localization results, even in cases where the camera motion is very large and texture information is poor.

2.2 Major Human Points Estimation

In this step, four major human points, namely: the head center, the mass center, the left end of leg and the right end of leg (see Fig. 1(c)) are detected and tracked using as input human silhouettes extracted by the Moving Objects Localization method. The method is divided into two procedures: the detection procedure and the tracking procedure. We select to track the above points as they are visible in the whole sequence providing sufficient information for the activity recognition. Our purpose is to develop a robust algorithm on low quality human silhouettes (see Fig. 1). This method is an extension of [7], where three major human points (the head center, the mass center and the end of leg) are detected and tracked. The consistency and the balancing of the leg tracking of [7] method is improved by tracking both of the end of legs.

2.2.1 Detection

In this step, the four major human points are automatically detected. This procedure is executed just once, in the first silhouette frame of the sequence. The previous position of the four major human points is unknown, so the input of the method (human silhouette) should be of high quality, without many misclassified pixels. The algorithm named "Human Points' Detection" is executed as it is described below.

First, the mass center point (X_c, Y_c) is computed. This point is defined as the mass center of the foreground pixels F. Next, the human body major axis (see Fig. 1(b)) is computed. It is defined as the main axis of the best fit ellipse. This axis passes from the mass center point, that already has been estimated, so we have to compute just the axis orientation. The orientation Θ is defined by the three second order moments $m_{1,1}, m_{2,0}, m_{0,2}$ (Equation 1, 2).

$$m_{p,q} = \sum_{(x,y) \in F} (x - X_c)^p (y - Y_c)^q \qquad (1)$$

$$\Theta = \arctan\left(\frac{2m_{1,1}}{m_{2,0} - m_{0,2}}\right) \qquad (2)$$

It is assumed that the human stands vertically in the first frame, so the head is found above the mass center and the end of the leg is found under the mass center. The head point (H) is defined as the farthest major axis point from the mass center (C), that is found above the mass center. Then, the end of leg points search space is reduced to the silhouette boundary points S that are found under the mass center. This property can be expressed by the following constraint $\vec{CH} \cdot \vec{CL} < 0.1 \cdot |CH|^2, L \in S$. The first end of leg point (L_1) can be computed by getting the farthest foreground pixel from the C, that is found below the C. The next end of leg point should have the following properties: high distances from the mass center, the head point and the first end of leg point. Moreover, the triangle PCL_1 should be close to an isosceles triangle. The last two constraints are equal to the triangle PCL_1 area maximization. Therefore, the next end of leg point is computed by maximizing an appropriate function F_l, where $E(PCL_1)$ denotes the area of the triangle PCL_1: $F_l(P) = |PH| \cdot |PC| \cdot E(PCL_1)$. The function $F_l(P)$ is maximized when the above constraints are satisfied providing at the same time the L_2 point. Finally, it is trivial to distinguish the leg points L_1, L_2 to the left and right end of leg points using the human major axis.

2.2.2 Tracking

In this step, the four major human points are tracked. This procedure is executed in every frame of the sequence, apart from the first one, taking as input the position of the four major human points in the previous frame and the current silhouette image. Finally, the position of the four major human points in the current frame is estimated by the algorithm that is described below.

First, we reclassify the binary silhouette image pixels in order to reduce the number of wrong classified pixels. This is done using the following method. We compute the minimum distance of each foreground object from the previous position of the four human points multiplied by the percentage of the foreground pixels that belong to a line segment started on the mass center of the foreground object and ended on the specific major human point. If this distance is higher than a threshold then the foreground pixels will be classified to background class (gray pixels of Figure 1(b)). Next, we reclassify all the background pixels that belong to human silhouette holes to foreground class.

The four major human points can be detected by "Human Points' Detection" algorithm which has been described in the detection step. This method produces two pairs of solutions for the head point and the leg points, as it is unknown if the head point is found above or under the mass center. We choose the pair which is closer to the estimated pair of the previous frame.

2.3 Human Motion Parameters

In this section we describe the human motion parameters in which is based the temporal signal segmentation and the activity recognition. Using the estimated four human major points trajectories, we can compute features that are independent from the camera motion. First we introduce the human body axis which is given by the angle between the human major axis and the horizontal axis (A_1). If this angle is about 90°, the human is standing or running, while this angle changes a lot during the jump of the high jump.

Moreover, a very important angle about the human motion is the angle between the legs (A_{34}). This angle is related to the human pose and the camera position. However, from its trajectory, the gait period can be measured providing an estimation of the human speed. This angle is used to discriminate a triple jump sequence from a long jump sequence. Let $A_1(t)$, $A_{34}(t)$ denote the angles A_1, A_{34} respectively at frame t (see Fig. 1(d)). We avoid the possible 2π discontinuities of signals $A_1(t)$, $A_{34}(t)$ by adding an appropriate $2k\pi, k \in Z^*$ factor on $A(t)$ if $|A(t) - A(t-1)| > \pi$.

2.4 Temporal Signal Segmentation

A discrimination to running stage and jump stage provides a temporal action recognition in each sequence. In triple jump, long jump the running stage include the first two and half jumps and the half jump, respectively. This segmentation is based on the angle signal $A_1(t)$. We have supposed that the human major axis on the fist frame is vertical (the athlete is standing or running). Let t_1 be the time that the jump stage starts. It holds that $A_1(t)$ remains almost constant for the time period $[0, t_1]$. During the jump stage, this angle changes a lot, but over the time it can be approximated by a d-order polynomial, $d \leq 5$. We have used the following algorithm to compute the time t_1.
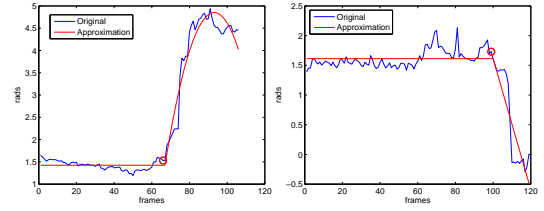

Fig. 2: Results of Temporal Signal Segmentation method for a high jump (left) and long jump (right) sequence. The original $A_1(t)$ signal and its approximation are plotted with blue and red line, respectively. The time t_1 is shown with red circle.
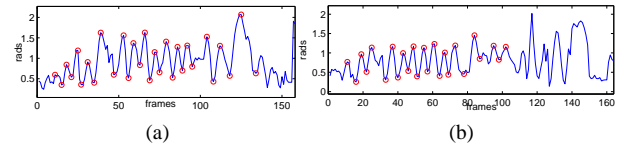

Fig. 3: The original $A_{34}(t)$ signal in a **(a)** triple jump and **(b)** long jump sequence. The times τ_k are shown with red circles.

Let h be a time variable. We approximate $A_1(t), t < h$, by a zero order polynomial and $A_1(t), t \geq h$ by a d-order polynomial in a least-squares sense, under the constraint that when $d \geq 1$ the approximated signal should be continuous. Let $E_1(h)$, $E_2(h,d)$ denote the errors of these approximations. Let h_d denote the time of the minimum error $(e_d = min_h(E_1(h) + E_2(h,d)))$. It holds that $e_d < e_{d-1}$ as the higher order polynomial will better approximate the curve. However, we have to select as t_1 the appropriate time h_d so that the numerical approximation of the second derivative of e_d ($e_{d+1} + e_{d-1} - 2 \cdot e_d$) is maximized. Results of this method are illustrated in Fig. 2, the order of the jump stage polynomial varies.

2.5 Gait Period Estimation

The gait period is a characteristic feature of the running stage. We can estimate it, by computing the mean period of angle signal $A_{34}(t)$ on a time window and a metric that measures the estimation robustness.

First, the times $(\tau_k, k \in \{1, \cdots, N\})$ of the local maximum and minimum of $A_{34}(t)$ are estimated in a short time window (see Fig. 3). If τ_k is a valid extremum (not noise), then the quantities $dp_k = |A_{34}(\tau_k) - A_{34}(\tau_{k-1})|$, $dn_k = |A_{34}(\tau_k) - A_{34}(\tau_{k+1})|$ should be about 1 rad. Thus, we introduce the reliability factor of the extremum τ_k $W(\tau_k) = dp_k^2 \cdot dn_k^2 \cdot e^{2 - dn_k^2 - dp_k^2}$. If $W(\tau_k)$ is close to one, then the measurement τ_k is probably valid. So, we have to use it on gait period estimation. Let $T_k = 2 \cdot (\tau_k - \tau_{k-1})$ be a measurement of the gait period using the τ_k, τ_{k-1}. The gait period (G) is determined by the weighted mean of T_k, $G = \frac{1}{\sum_{k=1}^{N-1} W(\tau_k)} \sum_{k=1}^{N-1} W(\tau_k) \cdot T_k$. The robustness of the G estimation can be measured by the mean (E_G) of the distribution $Pr(x = W(\tau_k)) = \frac{W(\tau_k)}{\sum_{i=1}^{N-1} W(\tau_i)}$, $E_G \in [0,1]$. If the mean is higher than 0.5, then the measurement G is probably valid.
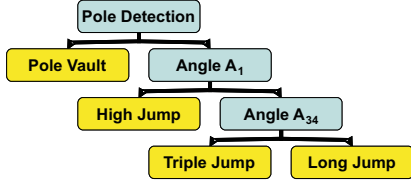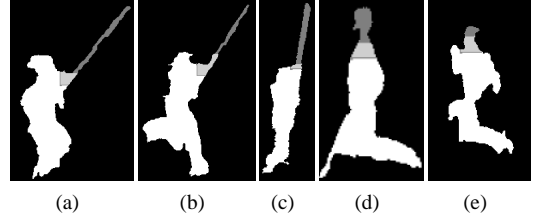
Fig. 4: The proposed classification schema.

Fig. 5: Results of the Pole Detection procedure. The light gray pixels denote those that ignored (last added) by the RG method and the gray pixels denote the estimated pole region. (a), (b), (c) The pole was successfully recognized, (a) $\varepsilon = 68.26$, (b) $\varepsilon = 44.39$, (c) $\varepsilon = 34.03$ (d) $\varepsilon = 8.12$, (c) $\varepsilon = 3.22$.

3. ACTION RECOGNITION

In this section, we present the action recognition method. First, the pole vault is recognized, since the pole can be easily detected using the silhouettes. Next, the high jump is recognized, based on the variance of the $A_1(t)$ signal. Finally, the most difficult discrimination between the triple jump and long jump is done using the angle signal A_{34}. The proposed schema is shown in Fig. 4.

3.1 Pole Vault Recognition

The pole vault is recognized first, since the pole can be easily detected using the athlete silhouette. We propose a silhouette analysis algorithm which is executed in each frame of the first half of the video sequence detecting the pole by its high eccentricity. If the pole is detected in at least two frames of a sequence, then the sequence is classified as pole vault. The method is described below.

First, the highest area object (O_1) is detected. Then, the end of pole point (P_e) is estimated. This point is defined as the farthest O_1 point from the mass center of O_1 object under the constraint that it is found above the mass center as the athlete is running. The pole pixels will be detected by a region growing method (RG) starting from P_e point. This method terminates when the area of region exceeds the 25% of the O_1 area or when the number of pixels (B) of the boundary between the region and O_1 exceeds a threshold (B_{max}). The threshold is a percentage (e.g. 20%) of the square root of the O_1 area approximating the O_1 mean width. However, the region will have been expanded in the athlete area. Therefore, we have to ignore the last pixels that RG adds, until the region eccentricity will be maximum (see Fig. 5).

Finally, the estimated pole region (O_2) is characterized as pole if its shape is like the pole's shape. We measure this similarity using the region eccentricity. A simple and low cost method for eccentricity estimation that works always successfully in our data is described hereafter. We compute the distance d between the farthest point (P_f) of O_2 from P_e and P_e itself. Then, the P_1 eccentricity (ε) can be estimated by the ratio $\varepsilon = \frac{\pi d^2}{O_2\ area}$. In the RG method P_f can be approximated by the last point that the method adds, so the eccentricity computation cost is $O(1)$. If ε is higher than a threshold (e.g. 20) and the region length is at least 35% of the O_1 length then the O_2 object will be a pole.

The 2D image projection of a pole is a rectangle. Thus, an alternative robust method could be the computation of the O_2 bounding rectangle, which is defined as the smallest rectangle enclosing the object. The O_2 eccentricity can be measured by the ratio $\frac{a}{b}$, where a, b denote the length and width of the rectangle, respectively. In addition, the ratio $\frac{ab}{O_2\ area}$ should be higher than a threshold.

3.2 High Jump Recognition

In the next step, the high jump is recognized. The high jump can be easily detected from triple jump and long jump because of the major human axis rotation during the jumping. Therefore, since the variance of the signal $A_1(t)$ will be higher in high jump than the triple jump and long jump cases, we used as threshold the value 0.3. This feature is independent from the camera motion and the athlete performance. A more robust computation of the variance can be done ignoring the 5% lower and the 5% higher values of $A_1(t)$.

3.3 Triple Jump and Long Jump Recognition

Finally, the triple jump and long jump are recognized. This is a difficult discrimination because of high similarity between them. In triple jump, the athlete gait period is lower during the jumps than the running stage (see Fig. 3(a)). On the other hand, in long jump, the athlete gait period during the jump is about the same as the running stage, or can not be measured when the legs are joined (see Fig. 3(b)). Therefore, we can discriminate them if we use a feature that measures how the gait period is changing during the sequence.

We apply the temporal signal segmentation method taking the running stage $[0, t_1]$. Let t_0 be the time so that the difference ($V(t_0)$) between the gait period measured at $[0, t_0]$ and the gait period measured at $(t_0, t_1]$ will be maximum. $V(t_0)$ is positive in triple jump and negative or close to zero in long jump. This feature is independent from the camera motion and the athlete performance.

4. EXPERIMENTAL RESULTS

We have tested the proposed algorithm on a data set containing 39 video sequences: 12 pole vault, 9 high jumps, 8 triple jumps and 10 long jumps. The correct classification rates were 100%, 88.9%, 87.5% and 80% for the pole vault, high jump, triple jump and long jump, respectively. Moreover, the pole vault and high jump classifiers don't produce any false alarms. Results of Human Points Tracking method are shown in Fig. 6. In many cases, the low quality silhouettes increase the errors of major human points computation (about $10 - 15\%$ of human height) mainly on leg points. Therefore, the 87.5% and 80% are high rates as the triple jump and long jump discrimination is only based on leg points motion. However, the discrimination between triple jump and long jump could possibly be improved using the x mass center point trajectory. We avoid to use it, because it is dependent on camera motion.
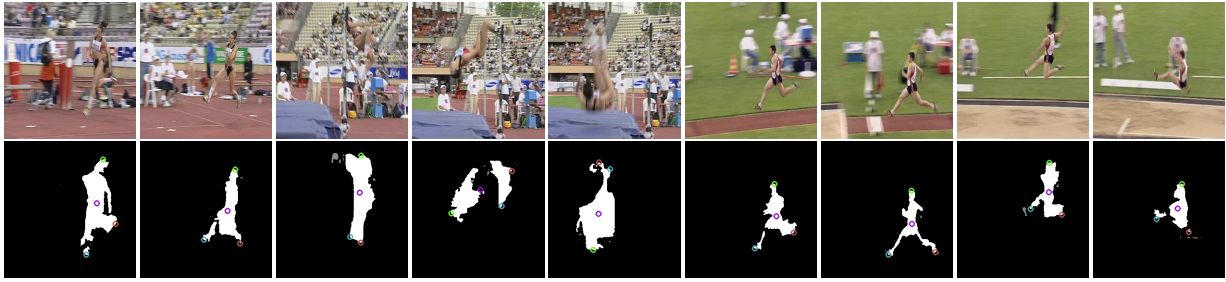
Fig. 6: Results of Major Human Points Tracking method on high jump and long jump sequence.

5. CONCLUSION

In this paper, an unsupervised, automatic human motion analysis and action recognition scheme is proposed tested on pole vault, high jump, long jump and triple jump videos. The silhouette analysis algorithm is color independent and it detects the major human points without tracking them. Consequently, if in some frames the silhouette estimation algorithm fails, the system will not loose its stability.

The pole vault recognition method, which is based on pole detection from silhouette shape, is executed first yielding 100% recognition ratio. The action recognition into long jump, high jump and triple jump is performed using independent from the camera motion and the athlete performance features, like gait period and human major axis signals. The human major axis feature is more robust to silhouette noise than the gait period feature. Thus, the high jump classification performance is the higher. An extension of the proposed methodology may include the addition of more sports and actions using more statistical features. Statistics analysis of athletics motion and video content based retrieval and indexing systems could be based on our method.

REFERENCES

[1] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[2] K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time: Part ii: Applications to human modeling and markerless motion tracking. *Int. Journal of Computer Vision*, 63(3):225–245, 2005.

[3] I. Grinias and G. Tziritas. Robust pan, tilt and zoom estimation. *Int. Conf. on Digital Signal Processing*, 2002.

[4] T. Mori, Y. Segawa, M. Shimosaka, and T. Sato. Hierarchical recognition of daily human actions based on continuous hidden markov models. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 779–784.

[5] A. S. Ogale, A. Karapurkar, and Y. Aloimonos. View-invariant modeling and recognition of human actions using grammars. In *ICCV Workshop on Dynamical Vision*, 2005.

[6] C. Panagiotakis and G. Tziritas. Recognition and tracking of the members of a moving human body. In *Proc. of AMDO 2004*, pages 86–98, 2004.

[7] E. Ramasso, D. Pellerin, C. Panagiotakis, M. Rombaut, G. Tziritas, and W. Lim. Spatiotemporal information fusion for human action recognition in videos. In *European Signal Processing Conference*, 2005.

[8] E. Sifakis, I. Grinias, and G. Tziritas. Video segmentation using fast marching and region growing algorithms. *EURASIP Journal on Applied Signal Processing*, pages 379–388, April 2002.

[9] E. Sifakis and G. Tziritas. Moving object localisation using a multi-label fast marching algorithm. *Signal Processing: Image Communication*, 16(10):963–976, 2001.

[10] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, 2003.

[11] L. Wang, H. Ning, and T. Tan. Fusion of static and dynamic body biometrics for gait recognition. *IEEE Trans. Circuits Syst. Video Techn.*, 14(2):149–158, 2004.