Fast detection and modeling of human-body parts from monocular video

Fast Detection and Modeling of Human-Body Parts from Monocular Video

Weilun Lao[1], Jungong Han[1], and Peter H.N. de With[1,2]

[1] Eindhoven University of Technology
P.O. Box 513, 5600MB Eindhoven, The Netherlands
[2] CycloMedia Technology B.V
P.O. Box 68, 4180BB Waardenburg, The Netherlands
{w.lao,jg.han,P.H.N.de.With}@tue.nl

Abstract. This paper presents a novel and fast scheme to detect different body parts in human motion. Using monocular video sequences, trajectory estimation and body modeling of moving humans are combined in a co-operating processing architecture. More specifically, for every individual person, features of body ratio, silhouette and appearance are integrated into a hybrid model to detect body parts. The conventional assumption of upright body posture is not required. We also present a new algorithm for accurately finding the center point of the human body. The body configuration is finally described by a skeleton model. The feasibility and accuracy of the proposed scheme are analyzed by evaluating its performance for various sequences with different subjects and motion types (walking, pointing, kicking, leaping and falling). Our detection system achieves nearly real-time performance (around 10 frames/second).

Keywords: motion analysis, trajectory estimation, body modeling, object detection.

1 Introduction

Successful estimation of the pose and modeling of human body facilitates the semantic analysis of human activities in video sequences [1,2]. The detection of human-body parts lays a solid ground to capture the human motion in more detail, which is essential for object/scene analysis and behavior modeling of deformable objects. Such semantic analysis can be explored for specific applications, such as surveillance, human computer interaction, virtual reality, sports analysis and 3-D gaming.

Accurate detection and efficient tracking of various body parts are ongoing research topics. However, the computation complexity needs significant reduction to meet a real-time performance, especially for surveillance applications. Existing fast techniques can be classified into two categories: appearance-based and silhouette-based methods. *Appearance-based* approaches [3,4] utilize the intensity or color configuration within the whole body to infer specific body parts. They can simplify the estimation and collection of training data. However, they

F.J. Perales and R.B. Fisher (Eds.): AMDO 2008, LNCS 5098, pp. 380–389, 2008.

are significantly affected by the variances of body postures and clothing. For the *silhouette-based* approach [5,6,7,8], different body parts are located employing the external points detected along the contour, or internal points estimated from the shape analysis. The geometric configuration of each body part is modeled prior to performing the pose estimation of the whole human body. However, the highly accurate detection of body parts remains a difficult problem, due to the effectiveness of segmentation. Human limbs are often inaccurately detected because of the self-occlusion or occlusion by other objects/persons. Summarizing, both silhouette and appearance-based techniques do not offer a sufficiently high overall accuracy of body-part detection. Also, the assumption of upright posture is generally required.

To address the challenging problem of accurately detecting and modeling human-body parts in a fast way, we contribute in two aspects. First, various differentiating body features (e.g. body ratio, shape, color) are integrated into one framework to detect different body parts without the assumption of the human's posture being upright. Second, we have proposed a novel scheme for capturing human motion, that combines the trajectory-based estimation and body-based modeling. This is effective to improve the detection accuracy. Our approach differs from current state-of-the-art work in the sense that it lacks training, while efficiently preserving the overall quality of the final results. More generally, the presented work aims at the object/scene analysis and behavior modeling of deformable objects. As our system is efficient and achieves nearly real-time performance (around 10 frames/second), we facilitate its application in a surveillance system.

The structure of this paper is as follows. Section 2 briefly presents the scheme. Section 3 introduces every detection component involved. The body-part detection that is based on seamless integration of different observation clues, is explained in detail. Promising experimental results and analysis are presented in Section 4. Finally, Section 5 discusses conclusions and our future work.

2 System Architecture

When combining the trajectory-based estimation and body-based detection, we intend to capture the human motion and locate the body parts using a skeleton model. The block diagram of our proposed scheme is shown in Figure 1. First, each image covering an individual body is segmented to extract the human silhouette after shadow removal. Second, both the trajectory-based and body-based modules are co-operating based on a particular sequence of internal functions. The position of the moving object in every frame is extracted. Occurring situations (behaviors) can be validated along the estimated trajectory for every individual person. Based on the trajectory-based estimation, the system initializes the local body-part detection. In this body-modeling module, various features are applied, such as appearance, body ratio and posture direction. As the fundamental anchor point in our skeleton modeling scheme, the center point of the whole body is also extracted. After different body parts are detected,

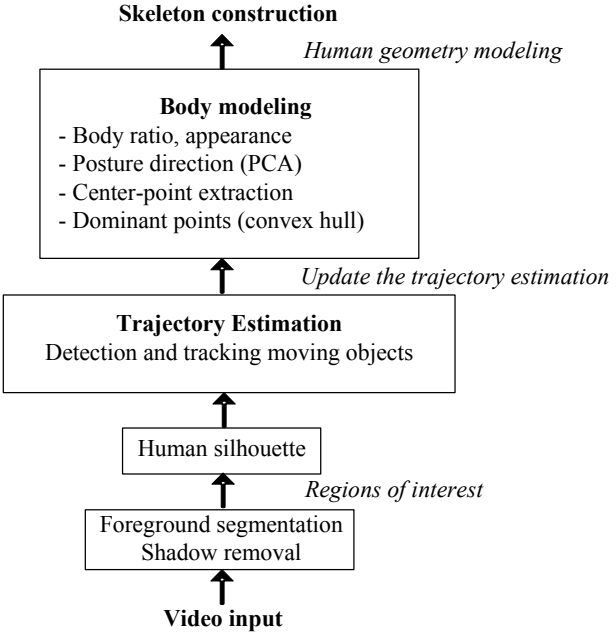

Fig. 1. Block diagram of our body-part modeling system

the human geometry is modeled. Finally, the skeleton model of every person is constructed.

3 Component Algorithms

3.1 Background Substraction

Background modeling is generally the first step of detection and/or analysis of moving objects in a video sequence. We perform an adaptive background subtraction to support person-behavior analysis. The intention is to maintain a statistical background model at every pixel.

In the case of common pixel-level background subtraction, the scene model has a probability density function for each pixel separately. A pixel from a new image is considered to be a background pixel if its new value is well described by its density function. For a static scene the simplest model could be just an image of the scene without the intruding objects. After the background modeling, the next step would be to e.g. estimate appropriate values for the variances of the pixel intensity levels from the image, since the variances can vary from pixel to pixel. Pixel values often have complex distributions and more elaborate models are needed. The Gaussian mixture model (GMM) is generally employed for the background subtraction. We apply the algorithm from reference [9] to produce the foreground objects using a Gaussian-mixture probability density. The parameters for each Gaussian distribution are updated in a recursive way.

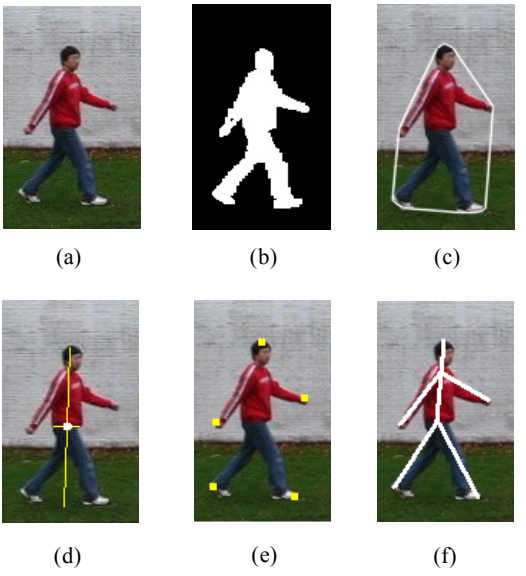

Fig. 2. Procedure of body-based processing: a) original frame, b) foreground segmentation (after shadow removal), c) body modeling based on convex hull, d) center-point estimation, e) body-part location and f) skeleton construction in single-person motion

Furthermore, the method can efficiently select the appropriate number of Gaussian distributions during pixel processing so as to fully adapt to the observed scene.

In the actual segmentation of foreground and background, shadow removal is another important issue. Based on the assumption that shadows decrease the brightness of pixels but do not affect their color, shadows are detected and removed [9]. To consider lighting changes during the process of video acquisition, the pixels labelled as background are used to update in a recursive manner. Finally, the labelled foreground pixels are grouped together to represent potentially moving objects.

3.2 Trajectory Estimation

The trajectory-based module estimates the human position over time, i.e. the movement, which is regarded as a fundamental function of surveillance systems. In our trajectory-based module, we apply blob tracking in two approaches. In a simple setting (e.g. static background, no occlusion), the first approach is based on an object's segmented binary mask. In the second approach, we employ the broadly accepted mean-shift algorithm for tracking persons, based on their individual appearance model represented as a color histogram. When the mean-shift tracker is applied, we detect every new person entering the scene and calculate the corresponding histogram model in the image domain. In subsequent frames for tracking that person, we shift the person object to the location whose

histogram is the closest to the previous frame. After the trajectory is located, we can conduct the body-based analysis at the location of the person in every frame.

3.3 Body-Based Modeling

The body-based processing block models the human motion by a skeleton model. The detailed procedure is illustrated in Figure 2. In the example of single-person motion, the input frame (Fig. 2a) is segmented to produce a foreground blob after shadow removal is applied(Fig. 2b). Then the convex hull is implemented for the whole blob (Fig. 2c). The dominant points along the convex hull are strong clues, in the case of single-person body-part detection. They infer the possible locations of body parts, like head, hands and feet. Here we employ a *content-aware* scheme (Section 4.1) to estimate the center point (Fig. 2d), which is fundamentally used to position the human skeleton model. Meanwhile, dominant points along the convex hull are selected and refined (Section 4.2) to locate the the head, hands and feet (Fig. 2e). Finally, different body parts are connected to a predefined skeleton model involving a center point, where the skeleton is adapted to the actual situation of the person in the scene (Fig. 2f).

4 Construction of Skeleton Model

We represent the body by using a skeleton model, which is used to infer the relative orientation of body parts and body posture. The center point is first estimated from the silhouette. Afterwards, it is connected to different body parts to construct the skeleton model.

4.1 Center-Point Extraction

The center point plays an important role in the skeleton model as a reference point. Its estimation accuracy significantly affects the detection of body parts. Here we apply a *content-aware* scheme to detect the center point c_i at the frame with index i. Contents of posture direction, human-body ratio and appearance are taken into account.

The posture direction of a human body can be estimated by the major axis m_i of the body's foreground region at the frame i. The major axis is determined by applying the *Principal Component Analysis* (PCA) to the foreground pixels. Its direction is given by an Eigenvector v associated with the largest Eigenvalue of its covariance matrix. Along the above direction and based on the somatological knowledge, we initially classify the whole body into three segments: head, upper body (including torso and hands) and lower body (two legs). Also, an initial body boundary b_i, dividing upper body and lower body, is produced. Next, within the neighboring area A from body boundary b_i, we perform the Laplacian filter $L_i(x, y)$ to each pixel (x, y) prior to a thresholding function $f(.)$ by value δ.

If $L_i(x,y) > \delta$, $f(.) = 1$. Otherwise, $f(.) = 0$. Then we search the optimal boundary line $\hat{b'_i}$ between the upper body and lower body in Equation (1) by

$$\hat{b'_i} = \arg\max_{b'_i} \sum_{(x,y)\epsilon b'_i} f(L_i(x,y),\delta), \tag{1}$$

where $L_i(x,y)$ indicates the Laplace operation with the 3×3 kernel at point (x,y). Finally, the center point C_i is located by the crossing point of the major axis m_i and the boundary line $\hat{b'_i}$ in Equation (2), hence

$$C_i = m_i \odot \hat{b'_i}, \tag{2}$$

where "$\odot$" denotes returning the intersection position between two lines. During our experiments, we have found that this center-point extraction is effective and accurate, and it is superior to the centroid-of-gravity (CoG) approach of the whole blob, as used in [5]. An example is visualized in Figure 3. Our proposed scheme is simple but effective, even when disturbed by residual noise after shadow removal. If the clothes between the upper body and the lower body are similar in the appearance, only the silhouette feature is employed. The center point is estimated based on the domain knowledge of the human-body ratio.

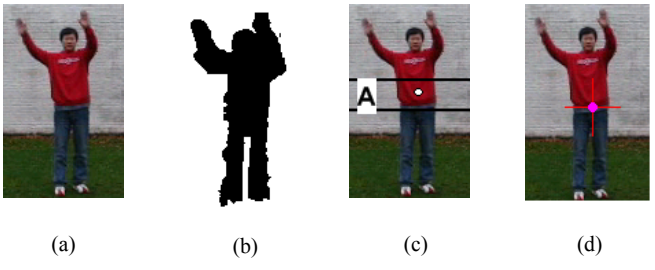

Fig. 3. Estimation of center point: (a) original frame, (b) silhouette after foreground segmentation, (c) result of CoG approach, (d) result of *content-aware* center point

4.2 Skeleton-Model Extraction

Different body parts are connected to the center point according to a predefined human geometry model, which is similar to the one reported in [8]. Every individual part is estimated according to the Euclidean distance between the center point C_i and every dominant point along the convex hull at the frame i. Based on the body-ratio knowledge, we initially select a set of dominant points P_i with the maximum distance in the three body segments, i.e. head, upper body and lower body. These dominant points are used to infer the locations of potential body parts. As we obtain the body segments (head, upper body, lower body) along the posture direction from Section 4.1, we can refine the points P_i in each individual segment to locate the body parts. Then we use a simple nearest-neighbor

filtering scheme to correlate different body parts over time. Afterwards, a Double Exponential Smoothing (DES) filter is added to refine the results. This filter provides good performance for moving object tracking [10].

The DES smoothing operator is defined by

$$\begin{cases} s_i = \alpha \cdot o_i \ + \ (1-\alpha) \cdot (s_{i-1} + d_{i-1}) \ , \\ d_i = \gamma \cdot (s_i - s_{i-1}) \ + \ (1-\gamma) \cdot d_{i-1} \ , \end{cases} \tag{3}$$

where o_i is the observed body-part position value at the frame i. The parameter s_i refers to the position after smoothing the observed position, d_i represents the trend of the change of body-part position, and α and γ are two weighting parameters controlling motion smoothness. Equation (3) applies to every detected body-part position for the individual person. The first smoothing equation adjusts s_i directly for the trend of the previous period with d_{i-1}, by adding it to the last smoothed value s_{i-1}. This helps to eliminate possible position discontinuities. The second smoothing equation updates the trend, which is expressed as the weighted difference between the last two position values.

After the smoothing filter is performed on the observed body parts, another post-processing step is implemented to improve the detection accuracy. If the distance between the detected hands and the center point is below a predefined threshold, we set the location of the hands as a default value, i.e. the position of center point. This additional processing can remove some inaccurate observation and improve the accuracy, especially in the self-occlusion case.

5 Experimental Results and Analysis

In our experiments, we have tested the algorithm for different monocular video sequences covering more than 2,500 frames. The video sequences were recorded at 15-Hz frame rate at a resolution of 320×240 samples (QVGA). The sequences cover different persons, background, clothes and behaviors in both indoor and outdoor situations.

We have evaluated our scheme with different motion types such as walking, pointing, kicking, leaping and falling. We implemented two state-of-the-art contour-based methods [5,8] for performance comparison. Figure 4 summarizes the accuracy comparison when using the different methods. In our experiments, the ground truth of body-part locations were manually obtained. The maximum tolerable errors in the evaluation is set to 15 pixels. Some visual examples of our experimental results are illustrated in Figure 5. After the body-part detection, the skeletons are superimposed on the images. Our system is implemented in C++ on a 3.0-GHz PC. The detection system operates at nearly real-time speed (around 10 frames/second).

From our experiments, we have found that the dominant points (with high curvature) along the contour play an important role in the three presented contour-based methods. If the dominant points are highly observable, e.g. in the motion types of pointing and kicking, all three methods yield similar performance. However, as we integrate the temporal constraints by employing the DES filter, our

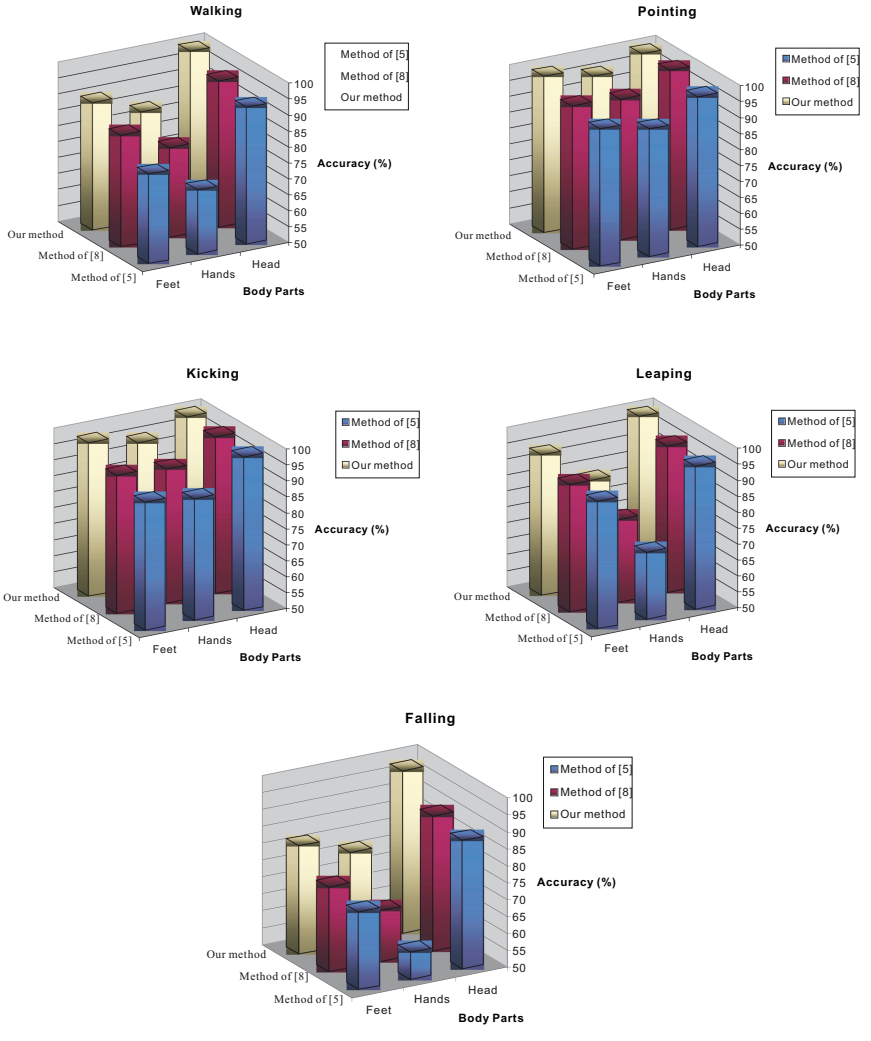

Fig. 4. Comparison of the detection accuracy of three different methods for walking, pointing, kicking, leaping and falling

detection accuracy is higher by around 5%, especially in the case of the self-occlusion when the hands/legs appear within the silhouette. Another interesting point is that our method does not assume that the human posture is upright. Moreover, the posture direction can be estimated in our algorithm. In the falling case, our method clearly outperforms the other two [5,8] by around 20% in the detection of hands.

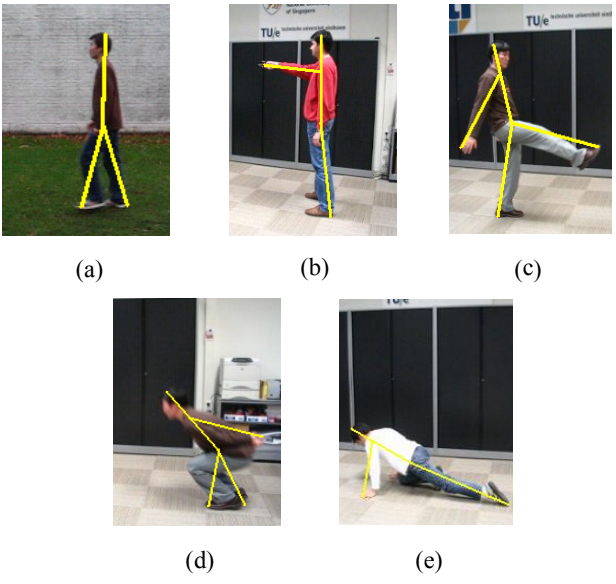

Fig. 5. The modeling result of single-person motion: (a) walking, (b) pointing, c) kicking, (d) leaping and (e) falling

6 Conclusions and Future Work

We have proposed a novel dual-module scheme for human-body modeling, that combines trajectory-based estimation and body-based analysis in a co-operating way, to capture the human motion and locate the different body parts. The trajectory-based module provides a platform for performing body-based analysis. The body-based module updates the tracking process, infers the posture of the human body and describes the body geometry efficiently by a skeleton model. We have presented a new algorithm for accurately locating the body center point, using the body silhouette and an upper/lower-body separation line. This algorithm outperforms the conventional center-of-gravity approach from existing literature, addressing the same center-point usage. Body-part detection was performed after estimation of the center point, analysis of body ratio, silhouette and appearance. An advantage is that the conventional assumption of upright body posture is not required. The above scheme has proven to be a fast (nearly real-time speed at 10-Hz frame rate) and effective technique for the automatic detection of different body parts within monocular video sequences in indoor/outdoor areas.

However, the current system has a few limitations. The self-occlusion problem is not completely solved, requiring additional exploration, as the dominant points along the convex hull fail to differentiate and locate the underlying body parts within the silhouette. We have found that the color appearance of the person is important in the case of self-occlusion. The region-based nature of color

will be utilized to improve the body-part segmentation. Also, we are going to capture motion sequences from different viewpoints and train the optimal parameters for different motion types, aiming at becoming more view-independent in performance.

References

1. Moeslund, T.B., Hilton, A., Kruger, V.: A Survey of Advances in Vision-Based Human Motion Capture and Analysis. Computer Vision and Image Understanding 104, 90–126 (2006)
2. Lao, W., Han, J., de With, P.H.N.: A Matching-Based Approach for Human Motion Analysis. In: Cham, T.-J., Cai, J., Dorai, C., Rajan, D., Chua, T.-S., Chia, L.-T. (eds.) MMM 2007. LNCS, vol. 4352, pp. 405–414. Springer, Heidelberg (2006)
3. Viola, P., Jones, M., Snow, D.: Detecting Pedestrians Using Patterns of Motion and Appearance. In: Proc. Int. Conf. Computer Vision, pp. 734–741 (2003)
4. Aggarwal, K.: Simultaneous Tracking of Multiple Body Parts of Interacting Persons. Computer Vision and Image Understanding 102, 1–21 (2006)
5. Fujiyoshi, H., Lipton, A., Kanade, T.: Real-time Human Motion Analysis by Image Skeletonization. IEICE Trans. Information and System 87, 113–120 (2004)
6. Haritaoglu, I., Harwood, D., Davis, L.: W4: Real-Time Surveillance of People and Their Activities. IEEE Trans. Pattern Analysis and Machine Intelligence 22, 809–830 (2000)
7. Yu, C., Hwang, J., Ho, G., Hsieh, C.: Automatic Human body Tracking and Modeling from Monocular Video Sequences. In: IEEE Proc. Int. Conf. Acoustics, Speech and Signal Processing, Hawaii, vol. 1, pp. 917–920 (2007)
8. Peursum, P., Bui, H., Venkatesh, S., West, G.: Robust Recognition and Segmentation of Human Actions Using HMMs with Missing Observations. EURASIP Journal on Applied Signal Processing 13, 2110–2126 (2005)
9. Zivkovic, Z., van der Heijden, F.: Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction. Pattern Recognition Letters 27, 773–780 (2006)
10. Han, J., Farin, D., de With, P.H.N., Lao, W.: Real-Time Video Content Analysis Tool for Consumer Media Storage System. IEEE Trans. Consumer Electronics 52, 870–878 (2006)