A matching-based approach for human motion analysis

A Matching-Based Approach for Human Motion Analysis

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]LogicaCMG Netherlands
P.O. Box 7089, 5600JB Eindhoven
The Netherlands
{w.lao, jg.han, P.H.N.de.With}@tue.nl

Abstract. This paper presents a novel approach to implement estimation and recognition of human motion from uncalibrated monocular video sequences. As it is difficult to find a good motion description for humans, we propose a matching scheme based on a *local descriptor* and a *global descriptor*, to detect individual body parts and analyze the shape of the whole body as well. In a frame-by-frame process, both descriptors are combined to implement the matching of the motion pattern and the body orientation. Moreover, we have added a novel spatial-temporal cost factor in the matching scheme which aims at increasing the temporal consistency and reliability of the description. We tested the algorithms on the CMU MoBo database with promising results. The method achieves the motion-type recognition and body-orientation classification at the accuracy of 95% and 98%, respectively. The system can be utilized for an effective human-motion analysis from a monocular video.

1 Introduction

Human beings can estimate and recognize human motions with high accuracy and robustness. To simulate this capability, efforts have been taken using various algorithms [1,2] with encouraging results. Successful estimation of the pose and motion type of people would allow the semantic analysis of human activities in video sequences. This process is very important and useful in various applications such as surveillance, human computer interaction, virtual reality and content-based video database query and retrieval.

Some research has been done to address the problem of effective human-motion analysis from uncalibrated monocular video sequence. There are two different approaches: appearance-based and body-part-based methods. Appearance-based approaches make use of the configuration of the whole body instead of specific body parts. In literature [3], specific static and stride parameters are used to perform motion recognition. Hidden Markov Models (HMM) can be used to perform the task of gait-based identification when the appearance of different shapes is

T.-J. Cham et al. (Eds.): MMM 2007, LNCS 4352, Part II, pp. 405–414, 2007.

learned as an initial distribution [4]. The appearance-based approaches can simplify the estimation and collection of training data since detailed labeling of the body components is not required. However, appearance-based techniques are significantly affected by the body postures and the camera viewpoint. For example, they cannot effectively distinguish the sequence captured from the front and the back of the person as their appearances are very similar to each other.

For the body-part-based approaches [5,6,7], different body parts (face, torso, the limbs, etc.) are located for detecting the person, using different features. The geometric configuration of each body part is modeled prior to performing the pose estimation of the whole human body. In other words, the estimation of the body-part positions can be used to interpret a person's pose and activity. These component-based approaches extract some elements of the body which guide the whole-body tracker. Then, a human activity can be represented as a collection of body parts moving in a specific pattern. However, the highly accurate detection of body parts remains a challenging problem due to the variances of pose and clothing.

The core of our method is a technique for matching an object in a data set, which is measured by investigating the similarity of body components and the silhouette. In this paper, we use a novel representation for the body parts and body shape, referred to as *local descriptor* and *global descriptor*, in order to facilitate the motion analysis in a robust way. Having a query frame, the combination of body parts and the body appearance is an effective aid for searching of the corresponding labeled frame in the data set. This combination guarantees a successful match, even if the detected contour of foreground objects is not precisely known. After this, the human-motion analysis related to a specific query sequence can be obtained. Furthermore, although not discussed explicitly in this paper, our scheme can successfully infer the pose, sometimes even when partial self-occlusion occurs. This improvement is important for potential applications like surveillance, when both the motion type and the specific orientation are required.

To solve the challenging problem of accurately analyzing human motion from uncalibrated video sequences, our contributions lie in two aspects. First, we propose a novel matching scheme to implement motion recognition, based on a weighted linear combination of local and global descriptors for a detected person. The properties of the query frame can be obtained after its labeled matching frame in the data set is retrieved. The other advantage involved is that we have employed a simple but effective spatial-temporal matching scheme. It can discriminate cyclic motions when spatial features are not sufficient. As a whole, our approach captures the human motion and analyzes its activity classification, which are essential for object/scene analysis and behavior modeling of deformable objects. Such scene analysis at the semantic level can be explored for specific applications, such as surveillance, tennis sports analysis and 3D gaming.

The structure of this paper is as follows. We briefly introduce every step involved in the algorithms in Section 2. Section 3 introduces our proposed matching functions. Local and global descriptors and spatial-temporal matching

approaches are explained in detail. Promising experimental results are presented in Section 4. Finally, Section 5 dicusses conclusions and future work.

2 Algorithm Overview

The top-down block diagram of our proposed algorithm is depicted in Figure 1, which contains four different steps. First, at the pre-processing step, each image covering an individual body is segmented to extract the blob representing foreground objects. The detected blobs are refined to produce a human silhouette afterwards. Second, at the modeling step, we implement the body-part detection, referred to as a set of local descriptors. For the shape-based analysis, we define a global descriptor. In our scheme, the similarity of different shapes plays an important role. Both the local descriptors and global descriptor are combined to implement a matching scheme. Every input frame is tested to find its corresponding human-motion image in the prerecorded data set. Then, a spatial-temporal cost function is proposed to distinguish different motion patterns of cyclic motion. Next, the outcome of modeling module is fed into the semantics module for further analysis. It can implement motion classification associated with the data set. The above techniques can be applied to a specific application. For example, the geometry/motion information of the reconstructed 3-D model is essential for tennis-sports performance analysis. The players and coaches can benefit from the video-based performance analysis to improve their training. The technical details involved at all the steps are described in the following section.

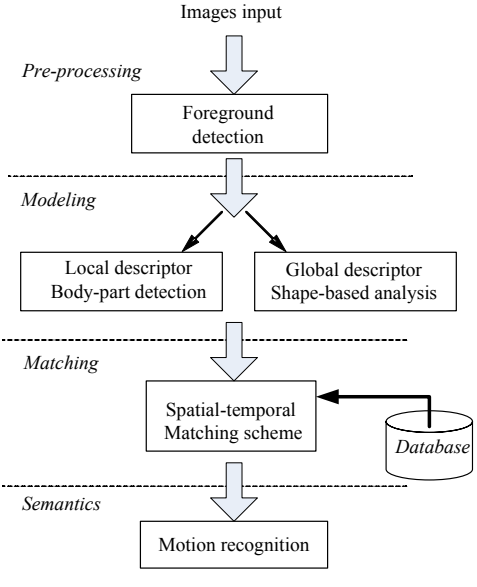

Fig. 1. The block diagram of our proposed algorithm

3 Matching Function for Body-Motion Estimation

3.1 Foreground Object Detection

As the first step, the foreground objects are detected employing background subtraction. This general approach can be used to segment moving objects in a scene assuming that the camera is stationary and the lighting condition is fixed. First, we store the image of the background $I_{bg}(x, y)$, as a reference image, without the foreground object. Then, given an original image $I_{ori}(x, y)$ from a particular sequence, feature detection of moving objects is performed within that image, but is restricted to areas where $||I_{ori}(x, y) - I_{bg}(x, y)|| > \delta$. Parameter δ is an adaptively chosen difference threshold. During experiments, we have found usually distorted and split blobs, which are still corresponding to the same person. To improve the blob segmentation, we perform a few iterations of morphological operations. Moreover, shadows cast on the background can be erroneously labeled as foreground. The shadow-removing approach of [8] is used in our scheme. The false segmentation caused by shadows can be minimized by computing differences in a color space that is less sensitive to intensity changes.

3.2 Local Descriptor and Global Descriptor

After the human silhouette is available, we propose a new hierarchical approach to describe the human motion. We design several local descriptors for relevant human body parts and a global descriptor for the body shape. This representation enables a better matching criterion for the motion of the complete human body.

Local Descriptor. The positions of n body components $B = (b^1, b^2, ...b^n)$ (like face, torso, limbs, etc.) are referred to as the set of *local descriptors*. These descriptors significantly represent the geometric structure of a particular person. Afterwards, the orientation of the human body concerned can be estimated based on its known structural position. For example, sequences captured from the front and the rear show a similar shape, but the availability and location of the person's face can successfully classify these two cases. Face and hands can be reliably detected, as their unique skin color contributes to the accurate detection. The detection of other body parts is comparably difficult. Since the color-based detection approaches have proved to be effective, these two body components are currently chosen. The distribution of skin color is trained off-line on a large database of images. We can also use an adaptive approach taking different lighting conditions into account and update the model of the skin color continuously. Additionally, face detection can be used to learn the skin color and obtain more precise and robust detection of hands with an on-line trained skin model. Next, tracking the face and hands is performed by following their respective detected blobs over consecutive frames. The tracking is initialized by the size and position of the blobs using the skin blobs in the image and tracked by finding the nearest blob in the next frame. During occlusion with two hands (when crossing each other), left and right hand share the same blob until they split again. During

occlusion with the body, the last detected position of a hand blob is maintained until a new blob re-appears close to this position and is assigned to the lost hand. For each frame in the sequence, the detected positions of the face and hands are represented as the normalized coordinates in the bounding box fitting to the detected human silhouette. If the face/hands are not detected, their coordinates are set to zero. Finally, for every frame l, we denote the position of the detected body component j as b_l^j. In our current work, $j \in \{1, 2, 3\}$, as only the face and two hands are the detection targets. Evidently, more body components can be incorporated in the scheme.

Global Descriptor. In addition to the detection of body components, we develop a shape-based *global descriptor*, based on the feature of shape contexts (SCs) [9]. This feature proves to be an effective tool for analyzing a particular shape. SCs are based on representing a shape by a set of sample points from the internal and external contours of an object, which are found by an edge detector (e.g. Canny detector). Suppose the shape of a detected body is represented as a set of n points $S_p = \{p_1, p_2, ..., p_n\}$, sampled from the internal and external contours of the shape. We use bins that are uniform in log-polar space, making the descriptor more sensitive to positions of sample points with smaller radius than to sample points farther away. For each point p_i on the shape, we compute a histogram h_i of the relative coordinates of all the remaining $n - 1$ points in S_p (these remaining points are in the set S_r, hence S_r excludes p_i). Taking p_i as the center point, we divide the space around p_i in a log-polar scale and define histogram bins on this grid accordingly. Then we count the number of points from S_r that are enclosed in each bin[9]. The number of points in each bin is the outcome of the function histogram $h_i(k)$, where k denotes the bin number. The histogram $h_i(k)$ is called the *shape context* of the point p_i. Similarly, we can calculate the shape context of every point q_i on the query shape with a set of points $S_q = \{q_1, q_2, ..., q_n\}$. Based on the shape contexts for all the points p_i and q_i, we can measure the similarity between the shape S_p and the shape S_q. Let $d(S_p, S_q) = \sum_i d(p_i, q_i)$ denote the distance between these two shapes, then this distance can be calculated by

$$d(S_p, S_q) = \sum_{i=1}^{n} d(p_i, q_i) = \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{[h_i(k) - h'_i(k)]^2}{h_i(k) + h'_i(k)} \ , \tag{1}$$

where n represents the number of sampled points on the shape, $h_i(k)$ and $h'_i(k)$ denote the K-bin normalized histogram of p_i and q_i, respectively. The above concept allows us to formulate our matching of a test (query) frame to a reference frame. In essence, the proposed global descriptor intends to investigate the similarity between the shape of a given person and its matching shape in the data set. It embeds both the external contour and internal contour of the shape and attempts to derive more information about the shape description.

3.3 Matching Using Local and Global Descriptors

After the local descriptor and global descriptor for a detected person in the frame are available, we combine them to produce a distance metric as shown in Equation (2) for matching. The estimation based on this metric aims at finding the optimal match between the query frame and its corresponding frame in the data set. The distance for matching one image to another is defined as a sum of two terms. The first term measures the relative spatial similarity in the detected body structure, and the second one measures the shape similarity in the two frames. Given a query frame l, now we can find its matching frame m in the data set for which the following function is minimal, hence

$$\arg\min_{m} \; w_{local} * (\sum_{j=1}^{ncom} w_j \|b_l^j - b_m^j\|^2) + w_{global} * d(S_l, S_m) \;, \tag{2}$$

where w_{local} and w_{global} are the weighting parameters for local descriptor and global descriptor, respectively. Parameter $d(S_l, S_m)$ represents the distance between the shape S_l and S_m, $ncom$ denotes the number of body components for detection, w_j is the scalar weighting factor for different detected body parts and $\|b_l^j - b_m^j\|^2$ denotes the distance of those body parts between two frames. From Equations (1)-(2), we are able to find the matching frame m. Then the key properties (motion type, orientation, etc.) of the query frame l are estimated from the labeled reference frame m. To improve the analysis accuracy, a temporal consistency can be enforced. For example, if more than half of the frames in a test sequence belongs to a particular motion type T, the whole sequence is labeled accordingly.

3.4 Spatial-temporal Matching for Cyclic Motion

In some classification cases, e.g. to distinguish a walking person carrying a subject from a normal walking person, only the spatial information is sufficient for classification. The combination of both local and global descriptors of a shape is applicable in such case. However, when the temporal information is essential, an elegantly designed spatial-temporal matching scheme is necessary. For example, to distinguish fast walking from slow walking, the temporal consistency should be considered. For this reason, we propose an approach that exploits motion dynamics and yet enforces temporal consistency.

Figure 2 shows the connectivity of different nodes, where each node represents a frame in a reference video sequence including a cyclic human motion. Suppose a cycle of motion (like walking) in the data set is composed of N frames in a reference set $F = (F_1, F_2, ..., F_N)$. We define in advance a so-called *state-transition cost* $C_{a,b}$, where a and b represent the (time-)index of the frames stored in the data set F (for simplicity, we have left out the frames F in the subscript to obtain a simplified notation and to avoid double subscripts). The value of $C_{a,b}$ depends on a distance metric between the a^{th} frame and the b^{th} frame. We propose to use a simple cost function, like the difference between the

index numbers, hence $C_{a,b} = b - a$ for $b \geq a$ and $C_{a,b} = 0$ for $b < a$. In the data set F, it should be noted that the N^{th} frame is followed by the first frame of the second cycle of the cyclic motion. Given a particular test sequence F' with N' frames, the matching is performed as explained in Section 3.2 on a frame-by-frame basis. After the matching, we obtain a set of shape-matching indexes $I = (I_1, I_2, ..., I_{N'})$ for every test frame F'_a with $1 \leq a \leq N'$, which is exactly the corresponding index of the frame in the data set F. For example, if the first test frame F'_1 is matched with the reference frame F_5, we obtain $I_1 = 5$.

Let us now distinguish between various motion patterns by computing the most probable class (e.g. motion type) for the test sequence. This class can be found by calculating the total path cost in the test sequence

$$C_{N'} = \sum_{l=2}^{N'} C_{I_{l-1},I_l} \ , \tag{3}$$

where C_{I_{l-1},I_l} indicates the path cost from the test frame F'_{l-1} to F'_l. In this way, the total cost $C_{N'}$ is used to classify different motion types.

Let us illustrate this classification with an example. Suppose we have a reference sequence of a motion cycle containing 30 frames for a slowly walking person. Given a test sequence with 10 frames which can be matched to the reference frame in the index order $\{5, 8, 11, ..., 28, 1, 4\}$, the path cost C_{10} can be calculated by $C_{10} = C_{5,8} + C_{8,11} + ... + C_{28,1} + C_{1,4}$. Since all frame index differences are 3, the result of this example gives $C_{N'} = 30$. Finally, with this result, we estimate the motion type (fast or slow walking) of the test sequence F' by comparing it to the cost for the reference sequences.

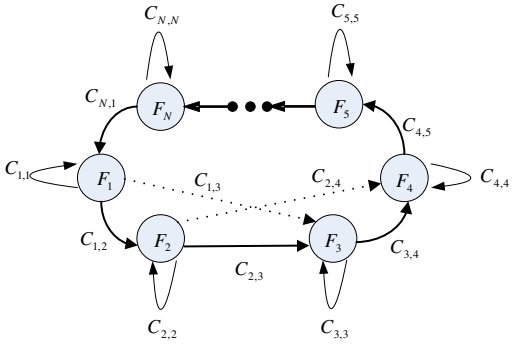

Fig. 2. Computation path for the classification of a cyclic motion

4 Experimental Results

We have tested the presented algorithms on the CMU MoBo database. This database contains a number of video sequences, which contain various subjects performing different types of motion on a treadmill. For each subject, the

Fig. 3. Examples of the retrieved images after the matching

database provides six sets of video sequences collected by six stationary cameras from different viewpoints. In our experiment, we selected the sequences of one subject (number 04006), each sequence being 30 frames (frames numbered 400-429) in length. For the data set, we selected from the aforementioned preselected sequences, the sequences with three types of motion (slow walk, fast walk, walking with a ball). For each motion type, five different sequences were used, covering five different viewpoints (vr03_7, vr05_7, vr07_7, vr13_7 and vr17_7). Thus, 15 sequences were contained in the data set. For testing, three other subjects (numbers 04022/04037/04068) were chosen to provide 60 test sequences. Every test sequence is 1 second long, recorded at 30 frames/s.

Using the techniques described in Section 3, we processed the sequences to evaluate our proposed method. First, the face and hands are detected and the shape context of the human silhouette is calculated for every frame in the data set to provide reference data. Afterwards, we input the test sequences to the system and implement the matching function driven by both local and global descriptors. After every frame is matched to a corresponding one in the data set, the type of the motion and body orientation can be labeled accordingly. Some examples of the matching frames are shown in Figure 3. The 1^{st}, 3^{rd} and 5^{th} rows of images are from test sequences. The 2^{nd}, 4^{th} and 6^{th} rows of images are from the data set. Moreover, the motion-type and orientation classification results are summarized in Table 1. It shows that we achieve the activity classification and body-orientation classification at the accuracy of 96% and 98%, respectively. These results indicate that our proposed approach is reasonably accurate.

Table 1. Recognition results on the CMU database

Motion type	Number of sequences	Motion recognition errors	Body-orientation recognition errors
Fast walk	20	2	0
Slow walk	20	1	0
Ball	20	0	1
Total	60	3	1

5 Conclusions and Future Work

This paper has presented a matching-based approach for the human motion analysis from a monocular video. The whole process detects the human motion, classifies the motion types and detects the body orientation. We have introduced a novel matching scheme to implement motion recognition based on a combination of local and global descriptors for detecting a moving human body. The local descriptors refer to body parts where we basically use the hands and the face. We applied a simple skin-color detection and prior knowledge about most likely positions to identify the hands and the face. The global descriptor is based on sample points along the contour of the human body. We have adopted the

concept of shape contexts to derive histograms where sample points occur in the image. The histograms are used for matching a query sequence with the human motion in a data set. We have defined a matching function that is based on the accumulated Euclidean distances of the body parts and the global difference between the shape histograms of the global descriptor.

We have also defined a spatial-temporal matching function to distinguish query motion cycles in human motion and match those to already stored sequences in a data set. We have shown that a simple cost function based on the differences between time indexes of video frames within a sequence can be used to distinguish motion patterns. Our approach was evaluated and showed a good effectiveness, as it implements the activity classification and orientation classification at the accuracy of 96% and 98%, respectively, in the CMU MoBo database.

We are currently working on a faster matching method, as well as the collecting and labeling more training sets with a large variety of poses and activities. The presented work should finally lead to the object/scene analysis and behavior modeling of deformable objects.

References

1. Moeslund, T.B., Granum, E.: A Survey of Computer Vision-Based Human Motion Capture. Computer Vision and Image Understanding, Vol. 81 (2001) 231–268
2. Wang, L., Hu, W., Tan, T.: Recent Development in Human Motion Analysis, Pattern Recognition, Vol. 36 (2003) 585–601
3. Bobick, A. and Johnson, A.: Gait Recognition Using Static Activity-Specific Parameters, Proc. Conf. Computer Vision and Pattern Recognition, Vol. 1 (2001) 423–430
4. Wang, L., Ning, H., Hu, W. and Tan, T.: Gait Recognition Based on Procrustes Shape Analysis, Proc. *IEEE* Conf. Image Processing, Vol. 3 (2002) 24–28
5. Haritaoglu, I., Harwood, D. and Davis, L.: W4: Real-Time Surveillance of People and Their Activities, *IEEE* Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No. 8 (2000) 809–830
6. Ioffe, S. and Forsyth, D.A.: Probabilistic Methods for Finding People, Int'l J. Computer Vision, Vol. 43, No. 1 (2001) 45–68
7. Lee, M.W. and Malik, J.: A Model-Based Approach for Estimating Human 3D Poses in Static Images, *IEEE* Trans. Pattern Analysis and Machine Intelligence, Vol. 28, No. 6 (2006) 905–916
8. McKenna, S.J., Jabri, S., Duric, Z. and Wechsler, H.: Tracking Interacting People, Proc. *IEEE* Conf. Automatic Face and Gesture recognition (2000) 348–353
9. Belongie, S., Malik, J. and Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts, *IEEE* Trans. Pattern Analysis and Machine Intelligence, Vol. 24, No. 24 (2002) 509–522
10. Shi, J., Gross, R.: The CMU Motion of Body (MoBo) Database, Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University (2001)