Model-Based Recognition of Human Walking in Dynamic Scenes

J. C. Cheng J. M. F. Moura

Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract - In numerous content-based video applications, it is important to extract from a video sequence a representation for humans in motion. For example, in generative video (GV) [4], one needs to construct accurate *world images* for moving objects. Because humans are not rigid objects, this task is difficult. We propose here a model-based recognition of human walking in dynamic scenes. We model the human body as an articulated object connected by joints and rigid parts, and the human walking as a periodic motion. We determine the posture by using a recognition algorithm that estimates the period and phase of walking. We obtain promising results when testing our algorithm with real video.

1 Introduction

Tracking and recognition of humans and their actions is a challenging task in computer vision. Due to its complex nature, the human body is non-rigid, it is capable of performing a wide variety of actions, and can be highly self-occlusive. To overcome these problems in tracking humans and their actions, most systems in this domain resort to model-based approaches. They either adopt an apriori model of the human body [3, 6, 2] or make assumptions on the types of motion of the human [3, 6].

Hogg [3] considered human walking recognition in real image sequences. He modeled both the human body and the human motion. The human body is described as a set of elliptical cylinders; the motion model is acquired interactively from a prototype image sequence. A similar approach is taken by Rohr [6]. Rohr also adopted a cylindrical model for the human body. However, Rohr modeled the motion through a time series, averaging the kinematic data provided by the medical motion studies conducted by Murray [5].

Gavrila and Davis [2] studies tracking human movements based on a multi-view approach. Their model of a human body is constructed with super-quadrics and a large number of degrees of freedom. Their system can track and recognize unconstrained actions, yet it needs known initial pose as a start-up and several static cameras to provide sufficient views.

Our system accomplishes recognition of a walking person in a complex scene. Functionally, it is most closely related to the work of Hogg [3] and

0-7803-3780-8/97/$10.00 ©1997 IEEE

Rohr [6]. However, the systems of Hogg [3] and Rohr [6] require well-calibrated environments, and the human subject walks front-and-parallel to the *static* camera. Our system allows for camera motion during video capturing. The task is made more complicated by the camera mobility.

Our system consists of three components: pre-processing, modeling, and recognition. The pre-processing component detects human subjects and locates their positions. The modeling component describes the body and the walking. The recognition component recognizes the posture of walkers with assistance from the modeling component.

Section 2 to 4 consider each of these components. Finally, Section 5 describes preliminary experiments and concludes this proposal.

2 Pre-Processing

The pre-processing component isolates the walker from the background and estimates the position of the walker. First, we estimate the motion of the background for every two consecutive frames. We assume that the background motion between two consecutive frames is parameterized accurately by 2-D motion models such as the affine model or perspective transformations. Currently, we use the affine model. The computation framework is based on an iterative multiscale approach.

After determining the image background motion, we register consecutive images using this motion. As a result, we null the image background motion; the remaining motion is due to the walker. Following this, we detect for each consecutive pair of registered images the region corresponding to the walker.

Finally, we track the walker to obtain the position and height of the walker. Experimental evidence reveals that the motion between the head and torso of a walking person is negligibly small; thus, we treat these two parts as a single rigid body. We estimate the 2-D affine motion of the head-and-torso between two consecutive frames. This gives us the evolution of the 2-D position of the walker between frames.

3 Human Modeling

Human models facilitate the recognition described in Section 4. There are two major components to setting up a model for the human walker: (1) the model of the human body, which provides the geometrical knowledge about the walker; (2) the model of the walking, which provides the topological knowledge about the walker. We use these two types of knowledge to synthesize the walker.

Modeling the Human Body: The purpose of our modeling scheme is to generate the contour information of a walker. It suffices for our purposes to adopt an articulated cone-shaped model. This model is similar to that adopted by Hogg [3] and Rohr [6] in their work. The human body is considered to be composed of 12 rigid parts (head, torso, plus two primitives of arms and

three primitives of legs). Each part is represented by a truncated cone with an elliptical cross section and a semi-oval sphere attached to each end of the cone.

Modeling Human Walking: We adopt a kinematic approach in modeling the human movements. Murray [5] conducted experiments on measuring gaits of males and females in a wide range of ages and heights. Their results reveal that the movement patterns of different body parts are similar for different people. Rohr [6] used the average measurements of the movement patterns [5] in his work. Encouraged by his results, we adopt the same set of measurements in modeling the human walking. We assign every two jointed parts a joint angle; there are 11 joints and joint angles θ_i, $(i = 1, 2, \cdots, 11)$. For each of the joint angles, we take a set of equally-spaced samples from a walking cycle of its corresponding average measurement [5] to build the model posture $\Theta_M(p) \stackrel{df}{=} [\ \theta_{M1}(p)\ \theta_{M2}(p)\ \cdots\ \theta_{M10}(p)\ \theta_{M11}(p)\ ]^T$ where $p \in [0, 1)$, referred to as the pose, is the index of the angle series. These series are periodic with period of 1.

4 Recognition of Human Walking

We define the walker detected from the real video as the data walker, $W_D(k)$, where k is the corresponding frame number, and the walker synthesized from the model as the model walker, $W_M(p)$, where $p \in [0, 1)$ is the pose. Since we track a walker in a dynamic scene, we expect the edges to be cluttered. To reduce the noise introduced by these cluttered edges, we consider only edges falling within the region corresponding to the walker extracted by the motion detection process described in Section 2.

We estimate the posture by matching edge information of the data walker with edge information of the model walker by a generate-and-test approach. We introduce below a similarity measure that quantifies how close a data walker W_D is from a model walker W_M. This similarity measure involves a phase filtering operation. This is based on constructing a distance map and a phase map.

Distance and Phase Maps: For the model walker with pose p, $W_M(p)$, we create the edge map $E_M(p)$ by using the Canny edge detector. We construct the distance map $\Gamma_M(p; x, y)$

$$\Gamma_M(p; x, y) = \begin{cases} \alpha + \beta(\delta_\Gamma - \min\limits_{e \in E_M} \|e, (x, y)\|) & \text{if } \min\limits_{e \in E_M} \|e, (x, y)\| \leq \delta_\Gamma \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where (x, y) is a pixel position, α and β are positive constants, e is the position of an edge pixel in $E_M(p)$, and δ_Γ is a given threshold. Then, we construct the phase map $\Phi_M(p; x, y)$

$$\Phi_M(p; x, y) = \begin{cases} \tan^{-1} \frac{\nabla_y (W_M * G)}{\nabla_x (W_M * G)} & \text{if } \min\limits_{e \in E_M} \|e, (x, y)\| \leq \delta_\Gamma \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where ∇_x and ∇_y are the components of the gradient operator and G is a Gaussian lowpass filter.

The distance map indicates the distance of a pixel to its closest edge pixel. The phase map is derived from the gradient of a blurred model walker; it possesses the orientation information of the edge map. We use these two maps as geometry filters to measure the geometrical similarity between the model walker and a data walker. Functionally, our distance map is similar to the chamfer image [1] used for measuring the similarity between two sets of edge pixels. The chamfer matching method in [1] computes the similarity between two sets of edge pixels by only measuring the distance between them. It doesn't consider the orientation information between these edges, which we believe is as important as the distance information. Our phase map provides this information by measuring the orientation between these two sets of edge pixels.

Similarly, we construct for the data walker $W_D(k)$ an edge map $E_D(k)$ and a phase map $\Phi_D(k)$. In this step, we choose $\delta_\Gamma = 0$, i.e., $\Gamma_D(k;x,y) \equiv E_D(k)$.

Similarity Measure: For the data walker in frame k, $W_D(k)$, we determine its closest pose in the model by

$$p_{sim}(k) = \arg \max_{p \in [0,1)} s(W_D(k), W_M(p)) \tag{3}$$

where $s(W_D(k), W_M(p))$ is the similarity measure

$$s(W_D(k), W_M(p)) = \frac{\sum_{(x,y)} S_M(k,p;x,y) \cdot \Gamma_M(p;x,y)}{\sum_{(x,y)} S_M(k,p;x,y)} \tag{4}$$

where

$$S_M(k,p;x,y) = \begin{cases} 1 & \text{if } (x,y) \in E_D \text{ and} \\ & |\Phi_D(k;x,y) - \Phi_M(p;x,y)| \le \delta_\Phi \\ 0 & \text{otherwise} \end{cases}$$

where δ_Φ is a given threshold. We call the procedure defined by $S_M(k,p;x,y)$ in the equation above phase filtering.

Fittest Posture: We find the closest pose, $p_{sim}(k)$, for each of the data walkers in a number of consecutive frames $W_D(k), k = 1, 2, \cdots, K$, by using the aforementioned approach; then, determine the period, $T_p \stackrel{df}{=} {f_p}^{-1}$, (in frames/cycle) and the phase, ϕ_p, (or the pose of the walker in the first frame of the video) by a line fitting algorithm

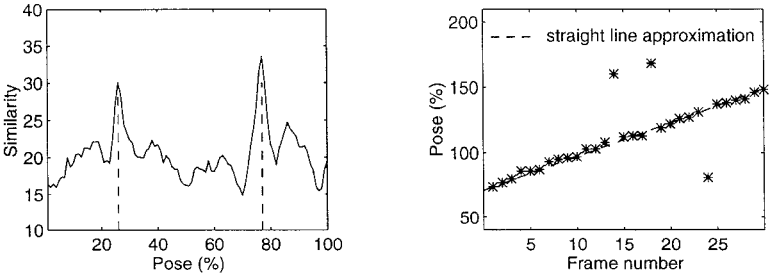
$$\begin{bmatrix} f_p & \phi_p \end{bmatrix} = \arg\min \sum_k \| p_{sim}(k), f_p(k-1) + \phi_p \| \tag{5}$$

We designate $p_{fit}(k) \stackrel{df}{=} f_p(k-1) + \phi_p$ to be the fittest pose of the data walker $W_D(k)$, and $\Theta_{fit}(k) \stackrel{df}{=} \Theta_M(p_{fit}(k))$ the fittest posture.

5 Experiments and Conclusion

We present results on recognizing the posture of a walker in the *Pedro* sequence. The *Pedro* sequence is a real video of an outdoor scene. We first apply the pre-processing component described in Section 2 to extract for each image the walking subject, which we refer to as the data walker. For each data walker, the recognition module searches the pose space: synthesizing a model walker, generating a distance map and a phase map for the model walker, and then determining the similarity measure to find the best match as explained in Section 4. To test the robustness of our approach, we apply our recognition algorithm to the first 30-frames segment of the *Pedro* sequence. We determine the pose for the data walker in each of the 30 frames by searching the entire pose space, i.e., from 0 to 1, with a pose increment of 0.01.

Figure 1 shows the results of matching the data walker of Frame 2. The horizontal axis is the percentage of pose in a walking cycle. The vertical axis is the similarity measure defined in (4). The result suggests that the model walker with pose of 0.77 is the best match to the data walker. We may notice in Figure 1 (a) that there is another peak centered around the 0.26 pose, which is about half a period apart from the major peak. This is due to the symmetric characteristic of walking. This large secondary peak may cause large errors (or outliers), see below.



(a) (b)

Figure 1: Initial posture recognition.

We perform the matching mentioned above on the data walkers for Frame 1 through Frame 30. The result is shown in Figure 1 (b). As can be seen, most of the data points scatter around a straight line except for three outliers, the three data points corresponding to Frames 14, 18, and 23. The outliers are due to the symmetric characteristic of walking as discussed above. These three data points will fall within the desired range if we compensate them by ± 0.50. We then determine the period and the phase of the posture for the data walker by applying equation (5). We obtain $f_p = 0.0267$ and $\phi_p = 0.7129$. This result shows that the fittest posture of the walker in frame k of the *Pedro* sequence is $\Theta_{fit}(k) = \Theta_M(p_{fit}(k))$ where $p_{fit}(k) = 0.0267(k-1) + 0.7129$. This indicates that the fittest pose for Frame 1 is $p_{fit}(1) = 0.7129$, and that

the period of the walking cycle is $T_p \cong 37.4$ frames/cycle.

We then superimpose the contours of the approximate model walkers to their corresponding data walkers. Some of the resulting images are shown in Figure 2. The results demonstrate that the posture recognition module is higly reliable.

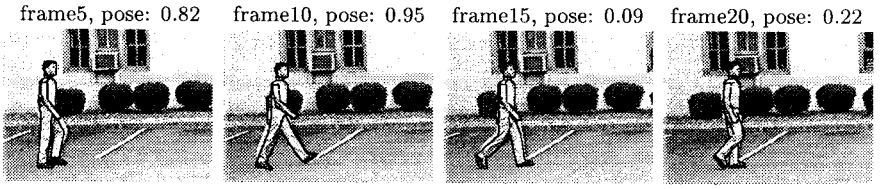


Figure 2: Superimpose contours of model walkers to walkers of real video.

In conclusion, content-based representation of humans in real video describes the humans according to their motion, shape, and texture. It involves solving the problems of action recognition, part segmentation, human modeling, and texture recovery. In this paper, we focus on action recognition. We propose a model-based recognition scheme for estimating the posture of a walking subject. We obtain promising results by testing our algorithm with real video. Our approach provides useful dynamic constraints for further segmentation of the body parts of the walking subject.

References

[1] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: two techniques for image matching," *Proc. of the 5th Annual Int. Joint Conf. on Art. Intell.*, pp. 659-663, Aug. 1977.

[2] D. M. Gavrila and L. S. Davis, "3-D model-based tracking of human in action: a multi-view approach," *Proceedings IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recogn.*, pp. 73-80, 1996.

[3] D. Hogg, "Model-based vision: a program to see a walking person," *Image and Vision Computing*, **1**(1), pp. 5-20, 1983.

[4] R. S. Jasinschi and J. M. F. Moura, "Content-based video sequence representation," *Proceedings of IEEE ICIP*, **2**, pp. 229-232, 1995.

[5] M. P. Murray, "Gait as a total pattern of movement," *American Journal of Physical Medicine*, **46**(1), pp. 290-332, 1967.

[6] K. Rohr, "Toward model-based recognition of human movements in image sequences," *CVGIP: Image Understanding*, **59**(1), pp. 94-115, 1994.