Tracking Human Walking in Dynamic Scenes

Jia-Ching Cheng   José M. F. Moura
Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213-3890

Abstract

Extracting from a video sequence a representation for humans in motion has numerous applications. This task is difficult due to the complex nature of the human body which is non-rigid and capable of performing a wide variety of actions. We propose here a model-based approach to tracking human walking in dynamic scenes. We model the human body as an articulated object connected by joints and rigid parts, and describe the human walking as a periodic motion. The posture of the walker is determined by a recognition scheme that estimates the period and phase of walking. This result is then used to establish dynamic constraints for the human posture. These constraints along with kinematic constraints that govern the linkage of the articulated human body are then adopted to facilitate the tracking of the body parts of the human. The paper illustrates the results of testing our algorithm with real video.

1 Introduction

Extracting from a video sequence a representation for humans in motion has numerous applications [4]. It involves solving problems of action recognition, part decomposition, part tracking, shape recovery, and texture recovery. These tasks are difficult due to the complex nature of the human body. 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, 7, 5, 2] or make assumptions on the types of motion of the human [3, 7].

Hogg [3] and Rohr [7] recognized human walking in real images. They adopted cylindrical models for the human body, and kinematic data measured from real walking subjects to model the walking.

Gavrila and Davis [2], and Kakadiaris and Metaxas [5] tracked 3-D human movements. Gavrila and Davis [2] modeled a human body using superquadrics. Kakadiaris and Metaxas [5] modeled the body parts of a human as deformable contours. Their systems can track unconstrained actions, yet they need to work in well-calibrated environments and with several static cameras to provide sufficient views.

Our system accomplishes recognition and tracking of a walking person in a complex scene. Functionally, it is most closely related to the work of Hogg [3] and Rohr [7]. There are two major differences between our approach and those in [3, 7]: (1) Hogg [3] and Rohr [7] required that the human subject walks front-and-parallel to the *static* camera, while we allow for camera motion during video capturing. The camera motion makes the task much more difficult; (2) the human body and the walking posture described by Hogg [3] and Rohr [7] are solely based on the prior model. They don't compensate for mismatches due to discrepancies between the model and the data. Our system adapts the models to the real video data by estimating the motions.

The method described in this paper consists of four components: pre-processing, modeling, recognition, and tracking. The pre-processing stage detects human subjects and locates their positions. The modeling step describes the body and the walking. The recognition block recognizes the posture of walkers with assistance from the modeling component. The tracking task tracks human walking using the recognized body parts as references. Sections 2 to 5 consider each of these blocks. Section 6 presents experimental results. Finally, Section 7 concludes the paper.

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 an affine model or a perspective

0-8186-8183-7/97 $10.00 © 1997 IEEE

transformation. In this work, we use the affine model. The computation framework is based on an iterative multiscale approach.

Once the image background motion has been determined, 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 and tracking described in Sections 4-5. 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 [7] 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 [6] 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 [7] used the average measurements of the movement patterns [6] 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 [6] to build the model posture $\Theta_M(p) \overset{df}{=} [\ \theta_{M1}(p)\ \theta_{M2}(p) \cdots \theta_{M10}(p)\ \theta_{M11}(p)\ ]^T$ 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

The goal of our recognition component is to estimate the period and phase of walking. It determines the posture by matching edge information of the data walker with edge information of the model walker by a generate-and-test approach.

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.

We introduce 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. The distance map indicates the distance of a pixel to its closest edge pixel. The phase map describes 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.

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, ${f_p}^{-1}$, (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.

$$[\ f_p \quad \phi_p\ ] = \arg\min \sum_k \| p_{sim}(k), f_p(k-1) + \phi_p \| \tag{1}$$

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

The task of the fittest posture estimation is to determine the period and the phase that best characterizes the posture of the data walker using the model posture, i.e., $\Theta_M(p_{fit}(k))$. Since the model assumed is generic and not as yet tuned to the specific values in the video under study, we expect some level of mismatch between the model and the data walker.

5 Tracking Human Walking

In order to generate an accurate representation for the walker, we need to refine our recognition by identifying each body part. The task of the tracking component is to construct a data posture $\Theta_D(k) \overset{df}{=} [\ \theta_{D1}(k)\ \theta_{D2}(k)\ \cdots\ \theta_{D10}(k)\ \theta_{D11}(k)\ ]^T$ where k is the frame index, which characterizes precisely the posture of the data walker in the video.

Our tracking component consists of two processes: a key frame registration which determines the true posture for some selected key frames, and a gradient-based tracking algorithm that estimates the true posture of the walker in each frame by using the determined posture of a key frame as an initial reference.

Key Frame Registration: We designate the walkers with least occlusions as key frames; thus, walkers with poses close to 0 and 0.5, whose arms and legs are widely open, are the candidate frames. We employ a divide-and-conquer strategy to fine tune the posture to its true value. The motion of the torso has already been obtained in the pre-processing stage; only the body parts of the limbs need to be identified. The fittest posture provides dynamic constraints which describe the occlusive relationship of the body parts and suggest the possible positions of the body parts. We use the matching method as described in the beginning of this section to determine the posture parameters $\theta_{D(i)}(key)$, $(i = 1, 2, \cdots, 11)$, for a key frame key.

We start the posture estimation from the limb with the highest visibility. For each limb, we first estimate the posture parameter $\theta_{D(\cdot)}(key)$ of the body part which connects the limb to the torso. Once the posture parameter is recovered, we go on to estimate the posture parameter between the newly identified body part and its jointed body part.

Tracking Algorithm: We adopt a gradient-based method for estimating the motion of the body parts between consecutive frames. Dynamic constraints of the posture and kinematic constraints of the articulation are incorporated to improve stability and reliability.

We decompose the human body into 5 parts: head and torso, and four limbs. Each part therefore consists of two or three rigid segments. We develop an algorithm to track a multiple-segmented articulated object. Below is our algorithm for tracking an object with two segments.

Let $I(\mathbf{x}, t)$ be the articulated object at time t, as shown in Figure 1. It consists of two rigid segments $I_1(\mathbf{x}, t)$ and $I_2(\mathbf{x}, t)$. $J_1(t)$ and $J_2(t)$ are the joints, and the distance between these two joints is d. We assume that this articulated object only has two degrees of freedom, rotating around the two joints. Let $I(\mathbf{x}, 0)$ be the reference image, and $I(f(\mathbf{x}, \mathbf{q}), \tau)$ be the corresponding image of $I(\mathbf{x}, 0)$ at time τ, where $f(\mathbf{x}, \mathbf{q})$ is the motion of pixel $\mathbf{x}$, which is parameterized by $\mathbf{q} = [\theta_1, \theta_2]^T$.

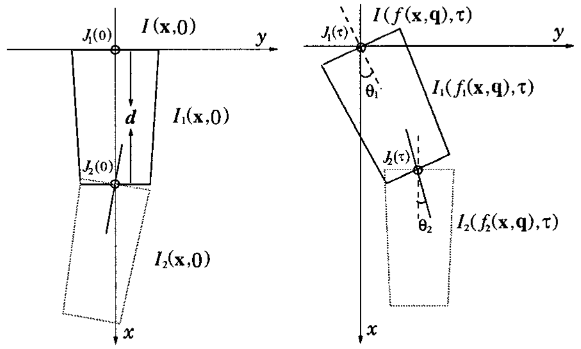


Figure 1: The structure of two-segment articulated object.

We obtain

$$f_1(\mathbf{x}, \mathbf{q}) = \mathbf{R}[\theta_1]\mathbf{x} \tag{2}$$

$$f_2(\mathbf{x}, \mathbf{q}) = \mathbf{R}[\theta_1] \begin{bmatrix} d \\ 0 \end{bmatrix} + \mathbf{R}[\theta_1 + \theta_2](\mathbf{x} - \begin{bmatrix} d \\ 0 \end{bmatrix}) \tag{3}$$

where $f_1(\mathbf{x}, \mathbf{q})$ is the motion of pixel $\mathbf{x}$ in the upper part; likewise, $f_2(\mathbf{x}, \mathbf{q})$ is the motion of pixel $\mathbf{x}$ in the lower part, and $\mathbf{R}[\theta]$ is a 2×2 rotation matrix.

We want to estimate the motion vector $\mathbf{q}$ by minimizing the cost function

$$C(\mathbf{q}) = \sum_{\mathbf{x} \in \Re} (I(f(\mathbf{x}, \mathbf{q}), \tau) - I(\mathbf{x}, 0))^2 \tag{4}$$

We expand $I(f(\mathbf{x}, \mathbf{q}), \tau)$ in a Taylor series as follows

$$\begin{aligned} I(f(\mathbf{x}, \mathbf{q}), \tau) = I(\mathbf{x}, 0) + \theta_1 I_\theta + \theta_2 I_{2\theta_2} - \\ d\,\theta_2 I_{2y} + \tau I_t + h.o.t. \end{aligned} \tag{5}$$

where $I_\theta \overset{df}{=} I_{1\theta_1} + I_{2\theta_2}$, $I_{i\theta_i} \overset{df}{=} -yI_{ix} + xI_{iy}$, $I_{ix} \overset{df}{=} \frac{\partial I_i(\mathbf{x},0)}{\partial x}$, $I_{iy} \overset{df}{=} \frac{\partial I_i(\mathbf{x},0)}{\partial y}$, $I_t \overset{df}{=} \frac{\partial I(\mathbf{x},0)}{\partial t}$, and $h.o.t.$ denotes the higher order terms. Under the assumption of small motions, we discard the higher order terms. Then we have

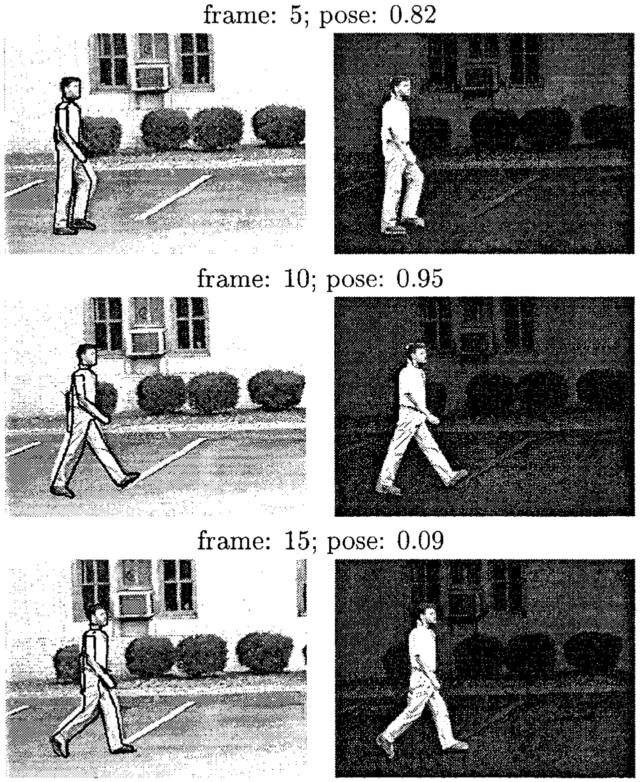
$$C(\mathbf{q}) \cong \sum_{\mathbf{x} \in \Re} (\theta_1 I_\theta + \theta_2 (I_{2\theta_2} - dI_{2y}) + \tau I_t)^2 \tag{6}$$

Differentiating the cost function in equation (6) with respect to $\mathbf{q}$ yields the solution

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = -\tau \begin{bmatrix} \sum I_\theta^2 & \sum I_\theta (I_{2\theta_2} - dI_{2y}) \\ \sum I_\theta (I_{2\theta_2} - dI_{2y}) & \sum (I_{2\theta_2} - dI_{2y})^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum I_\theta I_t \\ \sum (I_{2\theta_2} - dI_{2y}) I_t \end{bmatrix} \tag{7}$$

6 Experiments

We present results on recognizing and tracking the posture of a walker in the *Pedro* sequence. The *Pedro* sequence is a real video of an outdoor scene. 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. We perform the matching mentioned above on the data walkers for Frame 1 through Frame 30. We then determine the period and the phase of the posture for the data walker. We obtain $f_p = 0.0267$ and $\phi_p = 0.7129$. This result shows that the fittest pose of the walker in frame k of the *Pedro* sequence is $p_{fit}(k) = 0.0267\ (k-1) + 0.7129$. A detailed example can be found in [1]. We then superimpose the contour of the approximate model walkers to their corresponding data walkers. Some of the resulting images are shown in the left side of Figure 2. After determining the fittest posture, we estimate for key frames in the sequence the precise posture. This leads to an accurate segmentation of the human body. We then use these segmented body parts as initial references for tracking the interframe motion. The right side of Figure 2 shows the tracking results corresponding to the frames in the left side of Figure 2. These results represent very accurate tracking of the walker.



frame: 5; pose: 0.82

frame: 10; pose: 0.95

frame: 15; pose: 0.09

Figure 2: Recognition and tracking results.

7 Conclusions

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 decomposition, part tracking, shape recovery, and texture recovery. In this paper, we propose a model-based scheme for tracking human walking. We model the human body as an articulated object connected by joints and rigid parts, and describe the human walking as a periodic motion. We recognize the human posture by finding the frequency and phase of walking. We then use these recognition results along with kinematic constraints to track the body parts of the human.We obtain accurate results when applying our algorithm to real video.

References

[1] J. C. Cheng and J. M. F. Moura, "Model-based recognition of human walking in dynamic scenes," *Proceedings of IEEE First Workshop on Multimedia Signal Processing*, pp. 268-273, 1997.

[2] D. M. Gavrila and L. S. Davis, "3-D model-based tracking of human in action: a multi-view approach," *Proceedings of IEEE CVPR*, pp. 73-80, June 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] I. A. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection," *Proceedings of IEEE CVPR*, pp. 81-87, June 1996.

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

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