CAPTURE AND SYNTHESIS OF HUMAN MOTION IN VIDEO SEQUENCES

Jia-Ching Cheng and José M. F. Moura
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA 15213
{moura, cjc@ece.cmu.edu}

Abstract - We present a knowledge-based framework to capture and represent human walkers in video. The system models the human body as an articulated object of twelve rigid body-parts whose motions are almost periodic and subject to dynamic constraints. The resulting representation is compact and composed of the motion, shape, and texture for each of the body-parts. We apply the representation to regenerate the original sequence and to synthesize articulated 3D human actions.

INTRODUCTION

Capturing a human and its motion from live video is important in many applications [4]. Reference [3] develops a framework - generative video (GV) - where a video sequence is compactly represented in terms of its contents by informational entities - the background and the moving objects - plus additional contextual data like the motions of each entity. To obtain GV-like representations is a challenge when the moving objects are not rigid. This is the case with humans. This paper develops a content-based representation framework for human walkers. The system extracts from the video the motion, shape, and texture for the walker, and provides tools to reproduce and manipulate the video contents.

It is complicated to capture a human and its motion from a video sequence because the human body is not rigid, it is capable of performing complex actions, and it can be highly self-occlusive. To accomplish this task requires solving the following problems: human modeling, action recognition, body-parts tracking, and texture recovery.

We describe here a *knowledge-based* approach for capturing and representing humans and their motion from video sequences. Figure 1 shows the block diagram of our system. Functionally, it consists of three components:

- *Knowledge database* contains the models for the human body, the body-parts, and the human motions, as well as possibly for other video contents.
- *Capture of the human walker* includes algorithms to extract the motion, shape, and texture for each rigid part of the human walker and for any other object exhibiting a distinct motion, including the background.

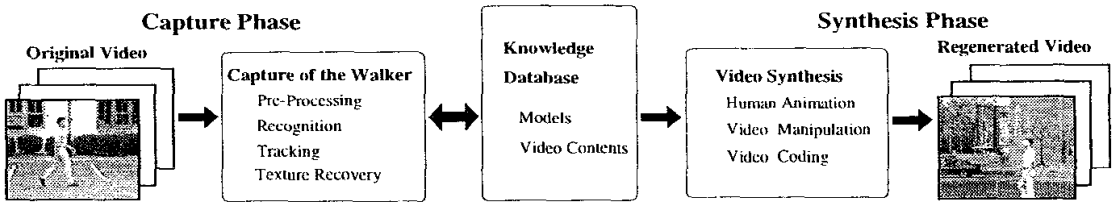


Figure 1: Human motion capture and representation.

- *Video synthesis* manipulates and resynthesizes video sequences by using the models and the video contents of the knowledge database.

KNOWLEDGE DATABASE

Our primary concern is with capturing humans and their motions from live video sequences. We develop a model-based approach to capture human walking. This approach utilizes models for the human body and the motions. The models are contained in the *knowledge database* component as shown in figure 1, see [1] for details.

The human body is described as an articulated object with twelve three dimensional (3D) rigid body-parts. Each body-part is a generalized truncated cone with semi-oval spheres attached to each end. The walking is described by twelve time-series that specify for each frame k the angle $\theta_i(k)$, $i = 1, 2, \cdots, 12$, between each pair of adjoining body-parts. The vector $\Theta(k) = [\theta_1(k), \cdots, \theta_{12}(k)]^T$ is referred to as the posture of the walker for frame k. We adopt prior walking patterns $\Theta_M(p) = [\theta_{M1}(p), \cdots, \theta_{M12}(p)]^T$ to characterize the walking, where p is referred to as the pose.

CAPTURE OF THE HUMAN WALKER

We now describe the *capture phase* of our system. This extracts from the video the posture and the texture of the walker. In other words, it estimates the posture parameters $\theta_i(k), i = 1, 2, \cdots, 12$, for each of the twelve body-parts, and recovers their texture. Our strategy is the following. We consider the walking as quasi-periodic. We have in our knowledge database a priori models for the walking as provided by ancillary data, see [1].

We capture the walker by establishing a correspondence between the posture of the walker in each video frame and the posture of a model walker generated using the prior motions. This correspondence is then used to fine tune the posture parameters and extract the texture for each body-part. The capture component is broken into four blocks: pre-processing, posture recognition, body-parts tracking, and texture recovery. We now detail each of these.

Pre-Processing

The pre-processing block estimates the 3D camera motion and the 3D position of the walker. It consists of four steps: (i) The estimation of the background motion. We model this motion with a 2D eight-parameter projective model. (ii) The detection of the human walker. This is done by low-level vision techniques that include background registration and motion based detection algorithms. (iii) The estimation of the motion for the walker's head-and-torso. We consider a 2D four-parameter affine model. (iv) The recovery of 3D motion from 2D. We infer 3D motion by using the estimated 2D motions of the background and the head-and-torso of the walker.

The output of the pre-processing block is: the 3D camera motion, and the position and the orientation of the walker across the video sequence.

Posture Recognition

The recognition block estimates $\Theta_{\text{fit}}(k) \stackrel{df}{=} \Theta_M(p_{\text{fit}}(k))$, where $p_{\text{fit}}(k) \stackrel{df}{=} f_p(k-1)+\phi_p$ is referred to as the fittest posture, $T_p \stackrel{df}{=} f_p^{-1}$ the period, and ϕ_p the phase of the walking posture. This is done by matching edge information of the walker with edge information of the model walker by a generate-and-test approach. We briefly describe below the posture recognition algorithm which we developed in [1].

We refer to the walker detected from the live video as the *data* walker, $W_D(k)$, where k is the corresponding frame number, and to the walker synthesized from the model as the *model* walker, $W_M(p)$, where $p \in [0,1)$ is the pose. For the data walker in frame k, $W_D(k)$, we search the pose space of the model walker, $W_M(p)$, $p \in [0,1)$, to estimate the pose, $p_{\text{sim}}(k)$, of the posture which is closest to the data walker $W_D(k)$, closest in the sense of maximizing a similarity measure $s(W_D(k), W_M(p))$.

The similarity measure, $s(W_D(k), W_M(p))$, quantifies the geometrical distance between the data walker $W_D(k)$ and the model walker $W_M(p)$. It requires knowledge of an edge map $E_D(k)$ and of a phase map $\Phi_D(k)$ for the data walker $W_D(k)$, as well as of an edge map $E_M(p)$, a phase map $\Phi_M(p)$, and a distance map $\Gamma_M(p)$ for the model walker $W_M(p)$. The distance map indicates the distance of a pixel to its closest edge pixel. The phase map 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. After finding the closest pose, $p_{\text{sim}}(k)$, for each of the data walkers $W_D(k), k = 1, 2, \cdots, K$, we determine the period T_p and the phase ϕ_p by a line fitting algorithm.

Body-Parts Tracking

Body-parts tracking locates accurately the position of all body-parts of the human walker across all the frames in the video sequences. This is necessary

to compensate for mismatches that remain at the output of the recognition block. These mismatches are attributed to two factors:

i. *Duration of the walking cycle*: In the recognition phase, the walking is assumed to be a periodic motion with a constant period. Walking, in reality, is better described as a cyclic motion where the duration of each cycle is not constant, i.e., the walking cycles Δ_i in figure 2 are uneven. We define the first frame of the walking cycle WC_m to be the anchor frame A_m, and the center frame to be the complement anchor frame A_m^c.

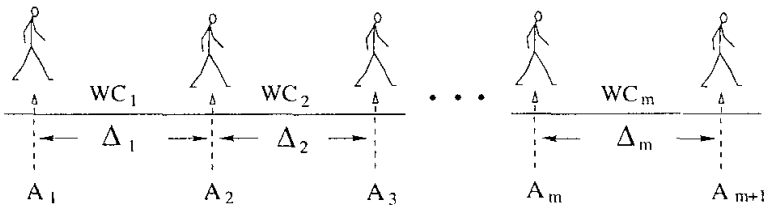


Figure 2: The anchor frames.

ii. *Posture of the real walker*: The recognition stage adopted a generic walking model to characterize the walking pattern. In the real world, of course, there are significant differences between the walking model and the real walker.

The goal of the tracking stage is to fine tune these two types of mismatches. We apply a tracking algorithm to significantly reduce the artifacts that persist after recognition has been accomplished. The output of the tracking is an improved estimate of the posture, we refer to this improved estimate as the data posture sequence $\Theta_D(k)$. Figure 3 shows a block diagram of the tracking algorithm which we refer to as Human Walking Tracking Algorithm (HWTA). An early implementation has been reported in [2].

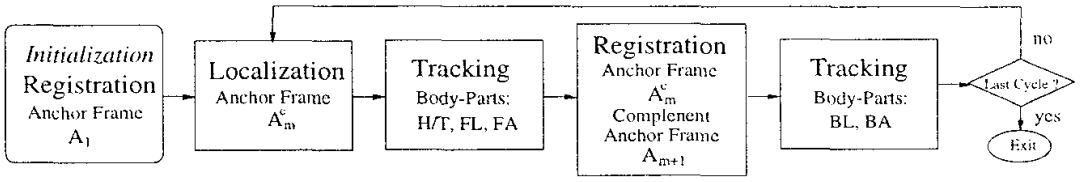


Figure 3: Block diagram of the Human Walking Tracking Algorithm.

This HWTA has three main modules:

- *Registration of frames* registers accurately the postures of the anchor frames A_m and the complement anchor frames A_m^c by determining the position of all body-parts.
- *Localization of anchor frames* locates for each walking cycle WC_m the anchor frames A_m. The anchor frames have poses close to 0.

- *Tracking of body-parts* estimates for each walking cycle WC_m the true posture of the walker in each frame. The recognition results provide initial reference templates for the tracking of the body-parts.

Texture Recovery

The last block of the capture component is the texture recovery. It extracts from the video the texture for the human walker. In this section, we discuss the construction of the texture.

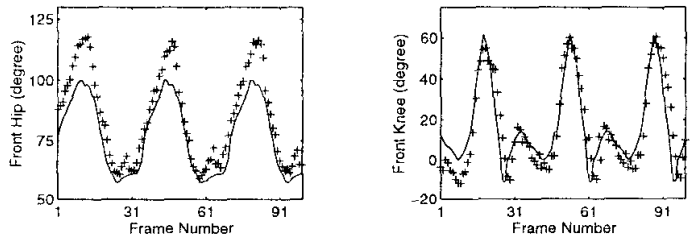
We assume that the 2D templates of the walker's body-parts, i.e., the 3D shape of the walker projected onto the image plane, change insignificantly during the walking cycle. This simplifies our representation of the shape of the body-parts to 2D templates. Using the modeling and the recovered motion for the body-part, we locate the corresponding position of the template in the video and extract the texture form the video.

Due to self occlusion, some body-parts, e.g., the torso and back leg, maybe partially occluded in some frames in the sequence. We recover the texture by integrating across the sequence. Once we obtain the 2D texture, we generate the 3D texture. This is done by inverse-mapping the 2D projection on the image plane to the 3D surface of the object.

EXPERIMENTS

We present experimental results for the *Pedro* sequence. The *Pedro* sequence is a live video of an outdoor scene of 300 frames. In the sequence the walker walks front-and-parallel to a moving camera.

Figures 4 (a) and (b) show the rotation angle of the hip joint and the ankle joint of the front leg, respectively. The dashed lines are the results from the recognition stage and the solid lines are after the tracking stage.

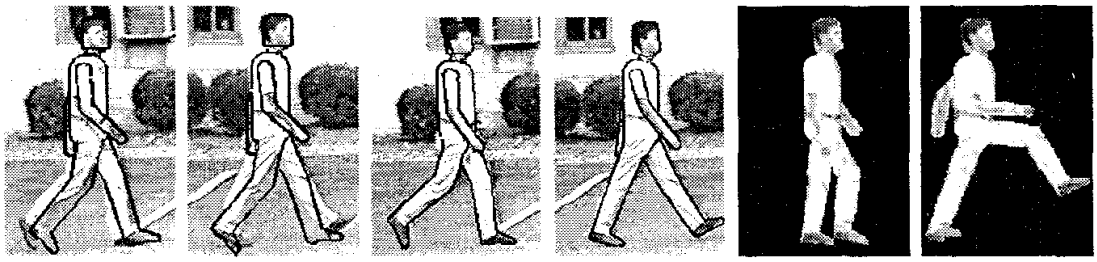


(a) (b)

Figure 4: Estimation results of posture parameters.

The images in figure 5 are generated by superimposing the contours of the synthesized model walkers to their corresponding data walkers. The model walkers in figures 5 (a) and (b) are generated by using the posture parameters acquired from the recognition stage. Some level of mismatch is apparent. The model walkers in figures 5 (c) and (d) are generated by using the posture

parameters obtained from the tracking stage. These results demonstrate very accurate tracking of the walker.



(a) (b) (c) (d) (e) (f)

Figure 5: Capture results: (a)-(d); synthesis results: (e)-(f).

With the motion, shape, and texture contents recovered for the human from the live video we can synthesize articulated human sequences. Figure 5 (e) shows a synthesized human using the recovered motion and texture. Figure 5 (f) shows a synthesized human using the recovered texture but with different posture.

CONCLUSIONS

Capturing humans and their motions from live video sequences has numerous applications such as content-based representation and human-machine interfaces. In this paper, we presented a system that captures a walking human from a monocular live video sequence. The result is a representation of the human walker in terms of motion, shape, and texture. This representation can be used to generate synthetic articulated human sequences. Future work will study other types of human actions.

References

[1] J. C. Cheng and J. M. F. Moura, "Automatic recognition of human walking in monocular image sequences," to appear in *Journal of VLSI Signal Processing*, 1998.

[2] J. C. Cheng and J. M. F. Moura, "Tracking human walking in dynamic scenes," in *Proceedings of IEEE Int. Conf. on Image Processing, ICIP'97*, **1**, pp. 137-140, 1997.

[3] R. S. Jasinschi and J. M. F. Moura, "Content-based video sequence representation," in *Proceedings of IEEE Int. Conf. on Image Processing, ICIP'95*, **2**, pp. 229-232, 1995.

[4] T. Molet, R. Boulic, D. Thalmann, "A real-time anatomical converter for human motion capture," in *Proceedings ot 7th EUROGRAPHICS Int. Workshop on Computer Animation and Simulation'96*, pp. 79-94, 1996.