

DENSITY PROPAGATION FOR TRACKING INITIALIZATION WITH MULTIPLE CUES

Cheng Chang, Rashid Ansari and Ashfaq Khokhar

ECE Dept., University of Illinois at Chicago

ABSTRACT

The paper presents an automatic initialization procedure for visual tracking of human motion. Instead of relying merely on low-level image features to give a single estimate of the initial human posture, the system seeks to find a set of samples that carries multiple hypotheses of the pose. By accumulating different image cues in the first 3-15 consecutive frames and combining dynamic information regarding human motion, the system builds a human body model for the person to be tracked from a video sequence and produces a sample set as an estimate of the posterior distribution of the initial posture. The sample set provides a good starting point for tracking with sequential Monte Carlo methods.

1. INTRODUCTION

There is a growing need for robust tracking of articulated human body motion in the fields of human-computer interaction, visual surveillance, gait analysis and person identification. There has been considerable previous work on developing such systems [1]. However, almost all tracking methods assume that a good target model is already available prior to tracking and the target state is known at least for the first video frame. The initialization problem is considered as a separate recognition problem and is not addressed in these methods.

Initialization of tracking is challenging since the estimation of target state has to be performed with only partial information about the target. While shape (edge) information is generally available for tracking, a region (color) model may not be feasible as colors and textures may vary from target to target. A robust initialization should selectively make use of different video cues and accumulate evidence from different features. Initialization is essentially a problem of how to effectively exploit the rich content in video data to prepare for tracking. This has not been sufficiently addressed in past work.

One of the methods of human pose estimation for tracking [2] estimates the initial pose of a walking human by matching the boundary of a human model with image observation for the first 10-15 frames. The author observes that

This research was supported in part by the NSF under the grant BCS-9980054 and CCR 0196365.

although it was possible to estimate the correct initial posture with this procedure, the problem is very hard and deserves further investigation. In [3], many manually marked 2D views of the human body are stored. Test images are matched against deformed exemplars until a good match is reached. Both methods employ only edges for pose estimation. In [4], a human tracker is initialized by using a parallel edge detector to collect body segment candidates. The method implicitly uses motion information since static candidates are discarded.

All the above methods give a single estimate of the subject's initial pose. We believe that the initialization process, like tracking, should carry multiple hypotheses as well, especially when no accurate target appearance model is available. In this paper, we propose an automatic initialization approach that builds human models for the subject to be tracked from a video sequence and obtains all necessary parameters for tracking. Use of a walking dynamic model is motivated by the fact that in many applications such as surveillance, people's ingress into the scene may occur from one of several possible entrances. Tracking can be initialized by estimating the person's walking pose in a specific view. The method combines motion and shape information to form a motion-enhanced shape image as a robust image feature. In addition, skin color is used as a supplementary cue to give a rough initial estimate of the walking pose prior to more accurate initialization process. Instead of giving a single point estimate, the method provides an estimate of the initial posterior distribution of the target state for subsequent tracking. The result of the initialization procedure is a sample set that accumulates information from previous observations and carries multiple hypotheses.

We first define the initialization problem in section 2. Model construction and parameter estimation from multiple video cues is described in section 3. Section 4 presents the sample-based initialization process. The remaining sections describe experiments and present conclusions.

2. PROBLEM DEFINITION

Denote the target state and observation at time t as \mathbf{x}_t and \mathbf{y}_t respectively. Let $\mathbf{Y}_t = \{\mathbf{y}_0, \dots, \mathbf{y}_t\}$ be the history of observations up to time t . In human tracking, \mathbf{x}_t is the

configuration of an articulate human body model, such as the one shown in Fig. 1. The initialization process should also estimate parameters such as the dimension of the body model \mathcal{M} , walking cycle T and walking velocity \mathbf{v} . Denote $\Theta = \{\mathcal{M}, T, \mathbf{v}\}$.

Given a video sequence containing a walking human in a given view and a possibly cluttered background, the initialization problem seeks to use as few frames (M) as possible to find a good estimate of $p(\mathbf{x}_M, \Theta | \mathbf{Y}_M)$ without prior knowledge about the subject’s region appearance.

We next show how various cues can be used to infer the conditional distribution of \mathbf{x}_t and Θ .

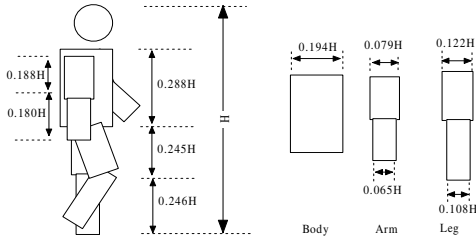


Fig. 1. The human body model used in this work.

3. MULTIPLE VIDEO CUES

3.1. Motion

Assuming a static camera and that the human subject is the only large moving region in the scene, motion information is extracted through background subtraction. The algorithm operates in RGB space and normalized rg space in order to eliminate shadows, where $r = R/(R + G + B)$ and $g = G/(R + G + B)$. A number of background frames are used to train a background model. For each pixel, the mean in all 5 channels and variance in chromatic channels σ_r, σ_g are computed and recorded. Given a new pixel, its (R, G, B, r, g) value is compared with the corresponding background pixel. If $|r - \mu_r| + |g - \mu_g| > C(\sigma_r + \sigma_g)$ or the difference in RGB is larger than a large threshold T_{RGB} , the pixel is declared as a foreground pixel.

After moving regions in the scene are detected, holes are removed by performing a ‘closure’ operation on the regions. In our experiments outliers were less detrimental than holes. The motion silhouette, denoted by m_t , is then defined by the largest moving region, or the two largest regions if their projections on the x -axis overlap for more than 40% of the width of the larger region. Fig. 2 shows an example of background subtraction of a subject walking and casting a shadow.

Given the silhouette, the size of the model can be estimated from the silhouette height H . The dimensions of various body segments and their proportions have been ex-

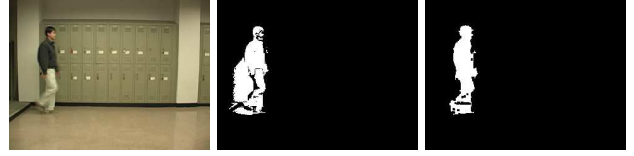


Fig. 2. Examples of background subtraction. Left: The Original color image. Middle: The result using only RGB background model. Right: The result using combined RGB and rg models with region operation.

tensively studied in anatomical literature [5]. Fig. 1 shows the body proportion used in this work, which is derived from [5]. Walking cycle T can also be estimated now by extracting the stride lengths (widths of the bounding boxes) from the silhouettes for M frames ($M \geq T/2$) and detecting the peaks in the strides. The pelvis location at time t is estimated by horizontally scanning the silhouette at the height of $0.491H$ and taking the mean of the midpoint of the silhouette and the midpoint of the bounding box. The walking velocity can then be estimated by taking the difference of pelvis locations in consecutive frames. Velocity is computed for every two consecutive frames for the first M frames and the median of these is taken as the estimate for \mathbf{v} .

We observe that the above procedure is generally robust for estimating Θ . Nonetheless we assume Θ could be slightly misleading and may undergo small changes and take that into account in our density estimation.

3.2. Motion-enhanced shape

A gradient map g_t is computed by convolving the original image with a edge detection kernel, $g_t = \mathcal{G}(i_t)$ where \mathcal{G} denotes convolution operation and i_t is the video sequence. This gradient map is then masked by the motion silhouette to set background gradient to zero.

Although g_t offers a good cue for accurate posture estimation, this information can sometimes be very weak, especially when the color of a person’s clothes is similar to that of the background. On the other hand, motion boundary is very pronounced since the boundary is detected on the binary motion mask m_t . However, we avoid using motion boundary directly as it is usually noisy with holes and outliers and lacks gradient inside the boundary. Instead we form a motion-enhanced gradient map by increasing the gradient value for every pixel if it is greater than certain threshold and is on the motion boundary, i.e., $g_t(x, y) = g_t(x, y) + \mathcal{G}(m_t)(x, y)$ if $g_t(x, y) > T_g$. Fig. 3 shows an example of motion-enhanced gradient map.

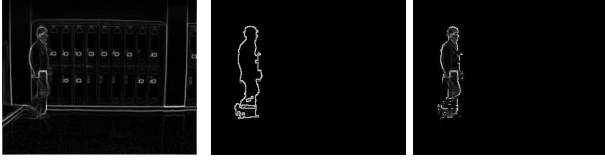


Fig. 3. From left to right: original gradient, motion boundary, motion-enhanced gradient.

3.3. Skin color

We observed that the estimated human pose based on gradient information is sometimes offset by half of the walking cycle from the true pose, which is also observed in [2]. Extra information needs to be used for disambiguation. Skin color is used in this work. The idea is to track hand locations and detect the occlusion of one of them. First skin color regions are detected within motion region m_t . The head is then detected by assuming the head is a large skin blob near the top of the motion region. Candidate hand regions are given by large blobs within a belt region of $0.35H$ to $0.65H$. Hands are located by associating each blob with the closest blob in previous frame and eliminating the blobs with too short trajectories. See Fig. 4 for an example of hand detection. Once the outer hand and the occluded hand are identified, a binary decision can be made about which half of the walking cycle the person is in for the first frame.

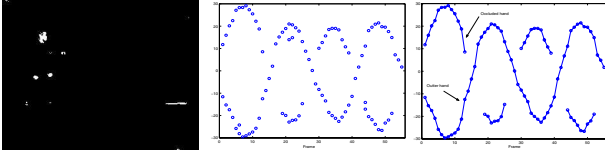


Fig. 4. Hand detection. Left: Detected skin blobs. Middle: The x -coordinates of hand candidates over time with the head location as reference point. Right: Two hands traced out.

4. DENSITY ESTIMATION

We employ a sequential Monte Carlo method to generate multiple hypotheses and propagate them over time.

4.1. Density propagation

The purpose of this initialization procedure is to estimate $p(\mathbf{x}_t|\mathbf{Y}_t)$. If samples from $p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1})$ is given, the new posterior can be obtained by propagating the samples from previous frame like in particle filter [1]. For the very first frame, with insufficient information regarding the subject, we assume that the initial body configuration is uniformly

distributed over the half walking cycle decided by skin color in section 3.3.

Given the sample set from $p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1})$, propagation is implemented through a *projection-drift-diffusion* process [6]. If we denote the walking motion as $X(t)$ whose value is the instantaneous configuration of the subject at time t , $X(t)$ would form a closed trajectory in its configuration space and each pose can be determined by $\phi_t \in [0, 1]$, known as *pose*. Given a sample configuration $\mathbf{s}_{t-1}^{(n)}$, its *projection* on the motion trajectory $\phi_{t-1}^{(n)} \in [0, 1]$ is found by finding the closest point on the trajectory to the sample. The projection then undergoes a *drift* described by $\tilde{\phi}_t^{(n)} = \phi_{t-1}^{(n)} + \Delta\phi_t^{(n)}$ where the pose changing rate $\Delta\phi_t^{(n)}$ follows a Gaussian distribution centered at $1/T$. The predicted sample then takes a random walk in the configuration space (*diffusion*).

4.2. Evaluation

Initialization is next evaluated for being acceptable for unimodal assumption. It is observed that after propagating the samples for a few frames, a high peak in the sample distribution emerges. A measure of effectiveness of a unimodal assumption is used as a criterion to stop the initialization process. Since dimensionality of $\{\mathbf{x}_t, \Theta\}$ is high ($d > 10$), we measure the distribution of the pose $p(\phi_t|\mathbf{Y}_t)$ instead. By projecting samples onto the motion trajectory, pose samples and associated weights $\{\phi_t^{(\cdot)}, w_t^{(\cdot)}\}$ can be computed.

Given $\{\phi_t^{(\cdot)}, w_t^{(\cdot)}\}$, an estimate of the pose density $\hat{p}_k(\phi_t|\mathbf{Y}_t)$ is obtained using kernel density estimation with a Gaussian kernel. On the other hand, under the unimodal assumption, a Gaussian density $\hat{p}_u(\phi_t|\mathbf{Y}_t)$ can be estimated from the sample set through Maximum Likelihood estimation. Kullback-Leibler divergence is then used to measure the difference of the two densities:

$$KL(\hat{p}_u, \hat{p}_k) = \int \hat{p}_u \log(\hat{p}_u/\hat{p}_k) d\phi_t.$$

The initialization process will stop if the change in KL divergence over two consecutive frames is negligible or M frames have been used.

5. EXPERIMENTS

We tested the system on several video clips containing both indoor and outdoor scenes and a variety of subjects. In these test videos the system produces a good sample set that are suitable to initiate more advanced Sequential Monte Carlo tracking techniques. Here we describe one of the test videos. The video contains 55 frames of a subject walking in a corridor in near frontal-parallel view. The person walks at normal speed (about 30 frames/cycle), with some abnormal movement of the arms. The background contains mainly straight edges, which would cause problems

in methods relying only on gradient. A few small regions with colors that are similar to skin color are also present in the background. About 70 background images are used to train a background model.

Using the human model in Fig. 1, the state \mathbf{x}_t is a 10-dimensional vector, with two joint angles for each limb plus pelvis location. The motion-enhanced gradient map is then used to compute sample weights. The weights are determined by the pixel values of the gradient map on the perimeter of the model. In addition to gradient, the silhouette image is used as a secondary cue to incorporate some region information. Samples that cover more silhouette regions are preferred. Gaussian models are used for the observation model, for which we start with a relatively large variance and gradually decrease the variance in order to prevent the samples from converging to a mode too early.

Some results of the experiment are shown in Fig. 5. Starting with 50 uniformly distributed samples over the first half of the walking cycle, the system gradually located the correct pose after propagating the samples for 5 frames, as seen in Fig. 5(a). For clarity 20 samples are shown as stick models over the original images. Fig. 5(b) depicts the change of estimated pose distribution during initialization. It can be clearly seen how the distribution evolves from a multi-modal distribution to a unimodal one. Fig. 5(c) shows the change of KL Divergence during initialization.

The system is evaluated on different starting poses. It is found that for normal walking speeds, a fair sample set can be produced within 5 frames for starting poses that are clear of self-occlusion. Poses with heavy self-occlusion create more modes in the sample distribution and sometimes need more time to converge. However, a half walking cycle is usually enough for any starting poses. Notice that we can always choose a frame with a good starting pose based on silhouette stride information. To avoid delays the system can propagate samples backward as initialization does not have to be a forward sequential process.

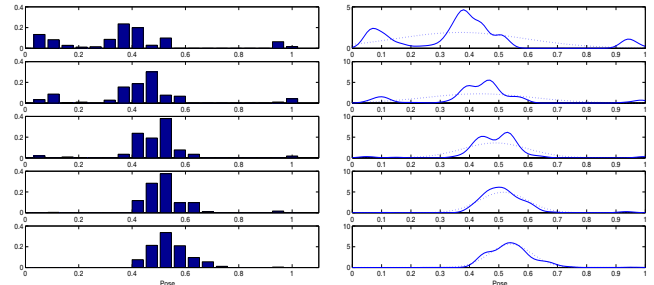
6. CONCLUSIONS

A system that performs automatic initialization for tracking human motion is described. The system combines bottom-up feature extraction that explores multiple video cues with a top-down sample propagation process that employs motion dynamics. The result of the initialization process is a sample set that accumulates information from multiple frames and carries multiple hypotheses.

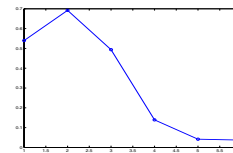
Other than background subtraction, techniques such as optical flow and motion segmentation provide more detailed motion information. We will further investigate these techniques and extend our work to initialization of multiple persons and multiple activities tracking.



(a) The sample set in the first, third, and fifth frames.



(b) Evolution of pose distribution. Left: Distribution represented by histogram. Right: Kernel density estimation (solid) and unimodal estimation (dotted).



(c) Evolution of KL Divergence during initialization.

Fig. 5. Some experiment results.

7. REFERENCES

- [1] H. Sidenbladh, M. Black, et al., “Stochastic tracking of 3d human figures using 2d image motion,” in *Proc. of ECCV*, 2000, pp. 702–718.
- [2] K. Rohr, “Human movement analysis based on explicit motion models,” in *Motion Based Recognition*, M. Shah and R. Jain, Eds., pp. 171–198. Kluwer Academic Publishers, 1997.
- [3] G. Mori and J. Malik, “Estimating human body configurations using shape context matching,” in *Proc. ECCV*, 2002, pp. 666–680.
- [4] D. Ramanan and D. A. Forsyth, “Finding and tracking people from the bottom up,” in *Proc CVPR*, 2003, pp. 467–474.
- [5] D. A. Winter, *The Biomechanics and Motor Control of Human Movement*, John Wiley and Sons, 2nd edition, 1990.
- [6] C. Chang, R. Ansari, and A. Khokhar, “Robust tracking of cyclic nonrigid motion,” in *Proc. ICIP*, 2003.