Project Report on Human Detection and Body Tracking

Ya Chang Changbo Hu

March 18, 2002

1. Introduction

The estimation of human motion from a monocular sequence of 2D images is difficult for many reasons: the non-linear dynamics of the limbs, ambiguities in the mapping from the 2D image to the 3D model, self occlusions, the similarity of the appearance of different limbs, and image noise etc. In this project, we learn and model background scenes statistically to detect foreground objects, distinguish people from other objects by checking skin-color blobs in the foregrounds. After we estimate the position of the body in the foreground region, we apply a Bayesian approach to track body motion, which is to compute the posterior probability distribution over the parameters of the 2D cardboard body model and dynamic model. Since the project has to be finished in 4 or 5 weeks, we make many simplifications in the realization. There are still many ideas that we plan to add into the current framework in the future.

There have been a significant number of projects on detecting and tracking people. Generally, they can be divided into three categories. The first category is to search over model parameters using the information from former images, and to measure the similarity between the predicted and the actual current image. This strategy has three standard steps in the literature:

I) Match template [1, 3~5]. Both [1] and [4] rely on background subtraction. In [4], the author located the body parts by silhouettes analysis. Due to the variety of body appearance in different pose, the accuracy of the algorithm is not satisfactory.

II) Find people by finding face according to the fact that people's normalized skin color is surprising constant across different skin pigmentation and radiation damage [6]. This approach is most successful when frontal faces are visible.

III) Search over correspondence between image configurations and object features [7~9].

The second category is to assemble image features into increasingly larger group, using the current group as a rough hypothesis about the object identity to select the next grouping activity. In the experiments using this strategy, the person is restricted to be naked or wear swimming suits. So the application will be highly limited.

Both of the two first strategies use quite sparse information from the images: such as edges [4], blobs [1] or other detected features. The third strategy is to use dense image information according to brightness constancy assumption [9]. Problem of detecting the model motion can be formulated as computing one of parameterized optical flow [24]. Though this method can provide dense image information, it is more sensitive to changes of appearance on human, such as wrinkles on the clothes, shading etc.

The choice of human model used for tracking depend on what kind of information has to be extracted, and also on what constraints can be introduced on the environments and on the activities of the tracked human. Generally, there are 3D limbs model for action recognition and 2D planar patches model for body pose recovery.

The way used in the project to detect and track human motion focuses on the motion of the figure. The human body is modeled as a 2D articulated object, parameterized by a set of joint angles and an appearance function for each of the rigid parts. With the aid of foreground detection, we compute the posterior distribution of these parameters by particle sampling, and propagate it through time.

2. System outline

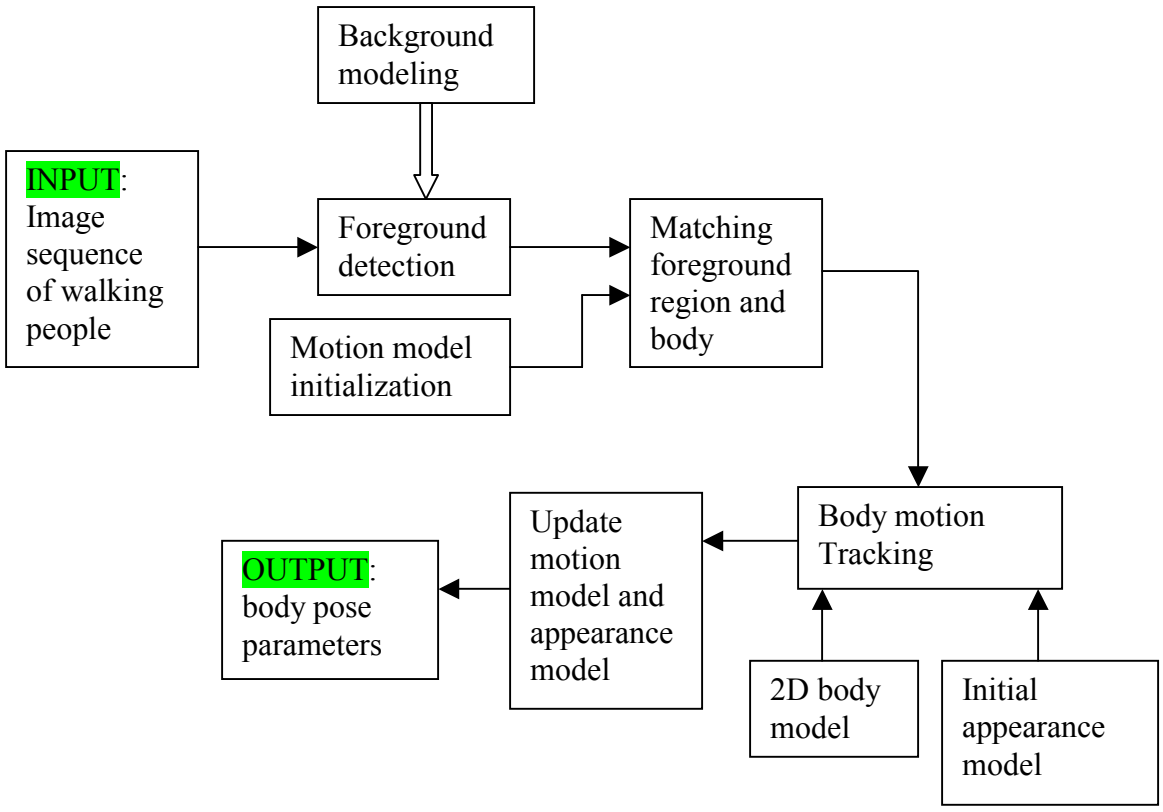

Fig.1. Detection and tracking system

The system diagram is shown in Figure 1. We use image sequence of walking people from [21]. Every frame is 320*240 resolution. The system is built by Visual C++. The working platform is 1.8 GHz Pentium IV PC.

3. Background learning and foreground detection

I. Haritaoglu et al [4] built a statistical model for a background scene to detect foreground regions even when the background scene is not completely stationary. Our project follows their algorithm, which is described as follows:

i) The pixel intensity of a completely stationary background can be reasonably modeled with a Gaussian distribution.

ii) The background scene is modeled by representing each pixel by three values: its minimum *m(x)* and maximum intensity values *n(x)* and the maximum intensity difference *d(x)* between consecutive frames observed during this training period.

We do not include the model of updating background model parameters in our project, because we only test short-term image sequence (about 100 frames). Illumination can be assumed not to change much in such short period of time.

A Gaussian filter is applied to the image sequence first to reduce the image noise. Foreground objects are segmented from the background in each frame of the video sequence by a four-stage process: thresholding, noise cleaning, morphological filtering, and skin-color detection. Each pixel is first classified as either a background or a foreground pixel using the background model. Pixel x from image I is a foreground pixel if :

$$B(x) = \begin{cases} 1 foreground \begin{cases} (I(x) - m(x)) < 2 * d(x) \\ \vee\ (n(x) - I(x)) < 2 * d(x) \end{cases} \\ 0 background \end{cases} \qquad (1)$$

Thresholding alone, however, is not sufficient to obtain clear foreground regions; it results in a significant level of image noise. After thresholding, one erosion process is applied to foreground pixels to eliminate one-pixel thick noise. Then, a fast binary connected-component operator is applied to find the foreground region. The small regions, whose areas are smaller than 8 pixels in our project, are eliminated. The regions are restored to their original sizes by a dilation process. In the bounding boxes of every remained foreground regions, we reapply background subtraction. Finally a binary connected component analysis is applied to the foreground pixels. Each foreground is assigned a unique label.

We judge whether the object is a person by detect whether there are skin color blobs in the bounding boxes [6]. Skin color is detected using a classifier with an empirically estimated Gaussian probability model of "skin" and "not-skin". By converting (R, G, B) triples into triples of the form (log(g), log(R)-log(G), log(B)-(log (G)+log(R))/2), skin cue is largely invariant to intensity or saturation, as this is robust to shading due to illumination. If there are skin color blobs with area great than a threshold, we classify the foreground region as a person.

4. Body motion tracking

After we detect the region that is likely to be a person, we want to estimate the motion of the head, limbs, and torso. In [7], a generative model of human appearance and motion is defined in Bayesian framework. The probabilistic formulation of the generative model provides the basis for evaluating the likelihood of the image measurements given the model parameters. A particle filtering approach is used to represent and propagate the posterior distribution over time, thus tracking multiple hypotheses in parallel.

Our project uses the Bayes' framework similar to [7], but has a simpler 2D body model instead of 3D cylindrical model in [7]. The appearance model is updated according to the extent that is occluded. We also use information from foreground detection in likelihood computing to increase accuracy. To reduce the computational complexity, the parameter space is partitioned according to the independency between some parameters. The parameters with higher hierarchy will be propagated first. The improved algorithm is

much more efficient than propagating all parameters simultaneously. Therefore we expect it to work better if the motion model can be trained by enough data using MPCA in the future.

4.1 2D cardboard body model

(a) (b)

Figure. 2.(a) Human body model (b)Match the model with the image

To track articulated human motion, we approximate the limbs and torso as rectangular planar [2]. In our simplified model, every joint has only one DOF, expressed as θ_i. The body is assumed to be upright. In our experiments, we assume that there are only two orientations of the torso: left or right body side to the camera. If the orientation is equal to 1, we map the body in the order of right arm, right leg, torso, head, left leg, and left arm; if the orientation is equal to –1, the mapping order is left arm, left leg, torso, head, right leg, and right arm instead. The limbs that are mapped first will be covered by the later mapped one if they are in the same position. So the entire pose of the body is given by 11 parameters, that is, 9 joint angles at neck, shoulders, elbows, hips, and knees, and the position of the torso. Let Φ be the vector containing these 11 parameters.

We assume the ratio between the sizes of the different body parts is constant in the image sequence. Actually this holds only when the orientation of the body keeps constant, the path of the motion is parallel to the image plane of the camera, and the static camera is modeled as a pinhole camera. The size of the model is therefore scaled by ratio between the height of the model and the height of the detected foreground region.

4.2 Appearance model

We assume that each limb and torso is textured mapped with an appearance model, $R(.)$. Moreover, it is desirable to estimate the appearance parameters through time to reflect the changing appearance of the object in the video. Here we compute the probability that a limb is occluded. If it is below the threshold, the appearance function $R_t(.)$ at time t is taken to be the mapping of the image at time t-1 onto the shape model by

the shape parameters at time *t-1*, otherwise, it is taken as the initial mapping function that is learned offline.

Given the parameters in the shape model and an appearance function for each of the rigid part, we can render images of how the body is likely to appear.

4.3 Motion Model

Our project uses a first order motion model, including velocity V only, to estimate the parameters of the shape model in subsequent frames.

4.4 Tracking Algorithm

4.4.1 Bayesian Formulation

The goal of tracking a human figure can now be formulated as the computation of the posterior probability distribution over the parameters R_t , Φ_t and V_t of the model at time t, given a image sequence $\bar{I}_t$. Using Bayes' rule and the Markov assumptions, the posterior distribution can be expressed as [7]:

$$p(\phi_t, V_t, R_t \mid \bar{I}_t) = cp(I_t \mid \phi_t, R_t) *$$

$$\int [p(\phi_t \mid \phi_{t-1}, V_{t-1}) p(V_t \mid V_{t-1}) p(R_t \mid I_{t-1}, \phi_{t-1}) p(\phi_{t-1}, V_{t-1}, R_{t-1} \mid \bar{I}_{t-1})] d\phi_{t-1} dV_{t-1} dR \quad (2)$$

4.4.2 Particle sampling

Due to the nonlinearity of the likelihood function over the model parameters, we cannot derive an analytic expression for it in the entire state space. So we represent the posterior as a weighted set of state samples, which are propagated using a particle filter with sequential importance sampling [23]. The detailed algorithm in [7] can be briefly described as below.

Each state, s_t, is represented by a vector of parameters in shape model and motion model. Appearance model can be determined by the shape parameters and the images, so we need not consider it here. The posterior at time *t-1* is represented by N state samples ($N \approx 10^4$ in our experiments). We first draw N samples according to the posterior probability distribution at time *t-1*. For each state sample, we propagate the angular velocities and shape parameters forward in time by sampling from their prior. At this point we have new values of Φ_t and R_t which can be used to compute the likelihood $p(I_t \mid \phi_t, R_t)$. The N likelihoods are normalized to sum to one and the resulting set of samples approximates the posterior distribution $p(\phi_t, V_t, R_t \mid \bar{I}_t)$ at time t.

4.4.3 Likelihood computation

The likelihood $p(I_t \mid \phi_t, R_t)$ is the probability of observing image I_t given the human model that has configuration Φ_t and appearance R_t at time t. In the 2D cardboard body model, the shape of each rigid part does not change in the video. So the projection is very fast.

We can represent the likelihood as:

$$p(I_t \mid \phi_t, R_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{f(I_t, \bar{I}_t)}{2\sigma^2}\right) \qquad (3)$$

where $\bar{I}_t = R^{-1}(R(I_{t-1}, \phi_{t-1}), \phi_t)$

$$f(I_t, \bar{I}_t) = \sum_{(x,y)\in S} \begin{cases} (I_t(x,y) - \bar{I}_t(x,y))^2, if\,(x,y) \in S_f \cap S_m \\ \max, if\,(x,y) \in S_m, but\,(x,y) \notin S_f\, or\,(x,y) \notin S_m, but\,(x,y) \in S_f \\ 0, otherwise \end{cases}$$

Where S is the bounding box of foreground region. S_f is the foreground region. S_m is the region with mapped body model. Max is an empirical large value to penalize the mismatch between S_f and S_m. σ is empirically determined.

4.4.4 Dynamic model

For a constrained activity such as walking or running, we assume that the angular velocity of the joints and the velocity of the body are constant over time. So the dynamics are represented by

$$p(\phi_t \mid \phi_{t-1}, V_{t-1}) = G(\phi_t - (\phi_{t-1} + V_{t-1}), \sigma_\phi)$$

$$p(V_t \mid V_{t-1}) = G(V_t - V_{t-1}, \sigma_V)$$

Where $G(x, \sigma)$ denotes a Gaussian distribution with zero mean and standard deviation σ, evaluated at x. The standard deviations σ_ϕ and σ_V are also empirically determined.

4.4.5 Parameter space partition

It is very time-consuming to search in the high degrees of freedom parameter space. According to the independency between the velocity of the torso and the velocity of the joint angles, and the fact that torso belongs to a higher level hierarchy in the body structure, the distribution over the position of the torso is propagated first, and the distributions of other joint angles are propagated conditioned on the already found distribution of the torso position. This is obviously much more efficient than propagating all parameters simultaneously.

Function $f(I_t, \bar{I}_t)$ has to be changed in likelihood computation of propagating torso position, because we cannot penalize the pixels that are in foreground but not in mapped torso region now. σ in (3) will be changed at the same time.

$$f(I_t, \bar{I}_t) = \sum_{(x,y)\in S} \begin{cases} (I_t(x,y) - \bar{I}_t(x,y))^2, if\,(x,y) \in S_f \cap S_m \\ \max, if\,(x,y) \in S_m, but\,(x,y) \notin S_f \\ 0 otherwise \end{cases}$$

Where S_m is the region with mapped torso.

5. Experiment results

On a Pentium IV 1.8GHz PC, the Visual C++ implementation takes approximately 3seconds/frame for experiments with 10,000 state samples. At frame 0, the posterior distribution is initialized manually with a Gaussian prior. To visualize the posterior distribution we display the contour of the every rigid part in the 2D model corresponding to the expected value of the model parameters.

We track a person walking on a straight path parallel to the camera plane over frames. The global rotation of the torso held constant. The model successfully tracks the person.

The legs are estimated accurately in most frames, but the arms drift in some frames due to the occlusion and the ambiguities of their poses.

Figure 2. Tracking results of a walking human

6. Conclusion

Our project combines the method of background modeling used in W4 and the Bayesian formulation in [7]. It can track articulated human figures in 2D using monocular image motion information.

Because search in such high degrees of freedom of articulated body motion is exponential computational complexity, our project is far from practical uses. There are several approaches to reduce the computational complexity. One is to relax constraints arising from articulation, and track limbs as if their motion were independent. The other is to introduce constraints, such as labeling using markers or color coding, prior assumptions about motion trajectories [7] or view restrictions. J Deutscher [10] develops an algorithm, called annealed particle filtering, that is to take a series of simplified versions of the evaluations function, use the converged result of every simpler version as a start point for a search on a less simple version, ending at an extremum of the original evaluation function. This method works well in decreasing the search in high dimensional configuration spaces. Another method in [7, 18] is to train the motion model using PCA/MPCA. Because many human activities are highly constrained and the body is often moved in symmetric and repetitive patterns, the parameters in the model can be reduced to 50% and the time for computation per frame is decreased greatly. Both of the two methods need great training data, so we cannot test it in our current project. We plan to test our system more thoroughly on other image sequences and try different algorithms to increase robustness and efficiency in the future.

Reference:

[1] C. Wren et al: “Pfinder: real-time tracking of the human body” PAMI July 1997

[2] S. Ju et al: “Cardboard people: A parameterized model of articulated image motion” IC on Face and Gesture Analysis, 1996

[3] A. Bobick et al: “The recognition of human movement using temporal templates” PAMI March 2001

[4] I. Haritaoglu et al: “W4: Real-time surveillance of people and their activities” PAMI, August 2000

[5] I. Haritaoglu et al: “W4S: A real-time system for detecting and tracking people in $2\frac{1}{2}$D”

[6] T. Darrell et al: “Integrated person tracking using stereo, color, and pattern detection” IJCV 2000

[7] H. Sidenbladh et al: “Stochastic tracking of 3D Human Figures using 2D image Motion”, ECCV 2000

[8] H. Sidenbladh et al: “Learning image statistics for Bayesian tracking” ICCV 2001

[9] H. Sidenbladh: “Probabilistic tracking and reconstruction of 3D human motion in monocular video sequences” Ph. D thesis

[10] J Deutscher et al: “Articulated body motion capture by annealed particle filtering” CVPR 2000

[11] A.Jepson et al: “Robust online appearance models for visual tracking” CVPR 2001

[12] S. Ioffe et al: “Probabilistic methods for finding people” IJCV 2001

[13] S. Ioffe et al: “Huamn tracking with mixtures of trees” ICCV 2001

[14] N. Jojic et al: “Tracking self-occluding articulated objects in dense disparity maps” ICCV 1999

[15] C. Bregner et al: “Tracking people with twists and exponential maps” CVPR 1998

[16] C. Bregner: “Learning and recognizing human dynamics in video sequences” CVPR 1997

[17] S. Intille et al: “Real-time closed-world tracking” CVPR 1997

[18] Y. Wu et al: “Capturing natural hand articulation” ICCV 2001

[19] R. Polana et al: “Low level recognition of human motion”

[20] A. Azarbayejani et al: “Real-time self-calibrating stereo person tracking using 3D shape estimation from blob features”

[21] http://www.nada.kth.se/~hedvig/data.html

[23] M. Isard et al: “Contour tracking by stochastic propagation of conditional density” ECCV 1996

[24] C. Bregler et al: “Tracking people with twists and exponential maps” CVPR 1998