Contour Cue Based Particle Filter for Monocular Human Motion Tracking

Li Hanlu, Zhou Zhong
State Key Laboratory of
VR Technology and Systems
Beihang University
lihl@vrlab.buaa.edu.cn

Zhang Shujun
Dept. Information Science
Technology
Qingdao Technology University

Wu Wei
State Key Laboratory of
VR Technology and Systems
Beihang University

Abstract

Particle filter is widely used in human motion tracking but its efficiency is low. A contour cue based particle filter algorithm is proposed in this paper for the human motion tracking in a markerless monocular video. The likelihood of sampled particles is measured by chamfer distance between two contours. One contour is extracted from the video image. The other is transformed from the sampled particle. The value of likelihood is the weight of corresponding particle. Then weighted particles are optimized by Levenberg-Marquardt method to make the final estimation closer to the posterior distribution of motion state. Apart from this, the skin part of human body is detected to constrain the sampled particles when the contour feature points are not sufficient with large occlusion. The experiment result shows that the contour cue based method is more efficient than the edge method.

Keywords: particle filter; human motion tracking; contour cue; estimation; likelihood measurement

1 Introduction

Video based human motion tracking has important applications in many fields including computer animation, computer games, virtual reality, intelligence monitoring, HCI, etc. In this paper, a contour cue based optimized particle filter approach is proposed for markerless human upper body tracking in monocular video. To improve the efficiency, the contour cue is used to compute the likelihood function. It is more efficient than the common used cues, like edge, silhouette, and color et al. After the likelihood measurement, a mathematic optimization process is adopted. A skin detection is applied to refine the particle weights, when the contour information is not sufficient enough to track, such as the situation of complete self-occlusion of one limb.

The paper is organized as follows: Section 2 describes related work of human motion tracking in markerless monocular video. Section 3 introduces the principle of tracking using particle filter and elaborates on the details of the proposed contour cue. Section 4 gives the outline of the contour cue based monocular human tracking algorithm. Section 5 provides the analysis of experimental results and Section 6 concludes the paper.

2 Related Work

In the research of human motion tracking in markerless monocular video, model-based probability approach is usually applied[Agarwal and Triggs 2004]. Particle filter is a multiple hypothesis probability approach, which is based on sampling and has no requirement of the object state distribution, so it is suitable for the non-linear motion[Poppe 2007]. Low efficiency is a main problem of this method, because the algorithm needs to compute the likelihood for all the particles. Some methods used optimization strategies to select more effective particles[Deutscher et al. 2005][Sminchisescu and Triggs 2003]. These approaches only improve the efficiency from reducing the particle number. Moreover, reducing the time consumption of likelihood function will also improve the tracking efficiency. To do this, more effective features needed. [Joachim et al. 2006] used edge, ridge and adaptive color cues to measure the likelihood between the model and observation images. In [Grest et al. 2006], it used corners which are tracked with KLT feature tracker to acquire the likelihood.

In this paper, we use an optimized particle filter approach to track markerless human body in monocular video. At the measurement stage, the proposed contour feature is used as the main cue to measure the likelihood. The contour feature is more concise and stable compared with the other cues. It saves the time of likelihood measurement. Although it will be failure when limbs are self-occlusion, we use the skin color detection to assist the sampling process in this situation. Contour cue with the assistance of skin detection improves the tracking efficiency, and also guarantees the tracking quality.

3 Contour Cue based Particle Filter

In this section the principle of the particle filter algorithm is introduced first, and then the proposed contour cue is analyzed, compared with the edge cue.

3.1 Principle of Particle Filter

Particle filter method for human motion tracking is essentially an approach of Bayes Posterior Probability Estimation. The objective is to estimate motion state by previous motions and video images. Assuming the motion process is a Markov process, the estimation equation of the motion state at time *t* is:

$$p(X_t|Z_t) = kp(X_t|X_{t-1})p(Z_t|X_t) \tag{1}$$

Where X presents a vector of the state space: $\{d_0, d_1, \ldots, d_m\}$. Each state X uniquely fixes one motion posture. Z presents the observation features of video images, such as edge, silhouette, color et al. These features can also be used combined with others. In our algorithm, we mainly use contour feature as the cue to match. k is the normalization constant. $p(X_t|X_{t-1})$ is a kind of dynamic model, used to predict the next motion state. It can be implemented through learning data, or directly by added a noise component. Using the dynamic model, a number of original particles are sampled. Each particle is a motion state, the particle set for time t is described as $\{x_t^1, x_t^2, \ldots, x_t^N\}$. $p(Z_t|X_t)$ is the likelihood function. For each particle sampled, the likelihood function measures the similarity between the motion state determined by particle and features extracted from the video image. To N particles, we have N similarity values.

VRCAI 2010, Seoul, South Korea, December 12 – 13, 2010.
© 2010 ACM 978-1-4503-0459-7/10/0012 $10.00

Taken these values as weight for each particle and normalize them as $\{\overline{\omega}_t^1, \overline{\omega}_t^2, \ldots, \overline{\omega}_t^N\}$, the finally estimation of motion state at time *t* is $X_t = \sum_{i=1}^{N} \overline{\omega}_t^i x_t^i$. The sampled particle number *N* is exponential growth with dimension of state space: $N \geq \frac{D_{min}}{\alpha^d}$, where D_{min}, α are constants, $0 < \alpha < 1$.

3.2 Contour Cue for Likelihood Measurement

The contour refers to the silhouette boundary. For the boundary always has strong intensity changes, so it is the part of image edge. From Figure 1 we can see, compared to the edge, contour feature points are more concise. Although edge feature sometimes contain more information than the contour, it is sensible to the noise and the change of environment conditions. While the contour is stable because the computation of it does not rely on the gradient. The contour feature points just save information about each whole limb and do not consider the internal changes of each limb. This is consistent with the assumption of rigid body. Figure 1 also shows the edge image and contour image in different light condition. Using contour as the cue, on the one hand it saves the time of compute $p(Z|X)$; on the other hand, it also avoids the failure measurement caused by the redundant information. Therefore, using contour cue to calculate the likelihood is most efficient. In this paper, we use the contour feature as the main cue to compute the likelihood function.

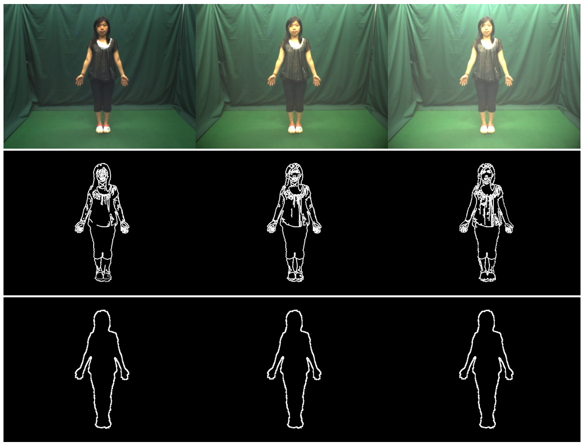

Figure 1: *the edge and contour image in different light condition, the first row is the original image, the second row is the corresponding edge image, and the last row is the corresponding contour image*

4 Tracking Algorithm

Based the above principle, the optimized particle filter tracking algorithm with contour cue of is presented in this section. The whole tracking process is first outlined, then the computation of contour feature and likelihood function is given, and the assist of skin color is described at last.

4.1 Contour Cue based Tracking

An articulated kinematic model with 8 DoF(degrees of freedom) is used to present the human motion state $S = \{x_1, x_2, \ldots, x_8\}$, with the known skeleton length. The shoulder is modeled as a ball joint with 3 DoF, and elbow as a hinge joint with 1 DoF. The appearance model used in this algorithm is the contour Y_c. Y_c^I presents the contour points of human body extracted from the video image. Y_c^S presents the contour points registered with the motion state. M_c is the number of contour feature points of the body.

Before the tracking begins, an initialization is manually taken to make the motion model *S* register with contour Y_c extracted from the starting frame. After initialization, the following process is repeated for each frames until the video capturing is stopped. At time step *t*, the prediction process predicts the motion state of current frame from the previous state. Assuming the motion difference between neighbor frames is small, the current state $\widetilde{S}_t$ is acquired from S_{t-1} by adding noise B_i to each DoF. The noise component B_i is a gaussian random variable with variance matrix d_i and mean 0. The variance of each degree of freedom varies with the history changes of that degree of freedom. In this step, *N* states from the motion space is sampled as the particles $\{\widetilde{S}_t^1, \widetilde{S}_t^2, \ldots, \widetilde{S}_t^N\}$. After prediction, the likelihoods between particles and captured image is calculated by matching the selected features. As described in Section 3, the contour cue is used to construct the likelihood function. The likelihood function is constructed as below:

$$p(Y_c|S) = \exp\{-\frac{\sum_{i=1}^{M_c}(r_i(Y_c^S, Y_c^I))^2}{M}\} \quad (2)$$

In equation(2), $r_i(Y_c^S, Y_c^I)$ presents the difference between the features from model state and video image at point *i*. It calculates the chamfer distance of the two features, for each contour feature points in Y_c^S, it finds the nearest distance to points in Y_c^I.

To make the particles more effective, the particles is optimized according to the measurement result. If the distance $R(Y_c^S, Y_c^I)$ of one particle is too large, this particle is moved to make the distance smaller using Levenberg-Marquardt method. After optimization, the particles are changed to $\{S_t^1, S_t^2, \ldots, S_t^N\}$, and the likelihood function values for these particles are recorded as $\{\omega_t^1, \omega_t^2, \ldots, \omega_t^N\}$. When the distance is reduced to the accepted range, these likelihood values are normalized as the weights of particles. The expectation of these weighted particles is used to approximate to the posterior distribution of motion state. When the distance stops at a large value due to the failure of contour feature, the skin part is detected to refine the particle set. The final motion state of current frame is estimated as $S_t = \Sigma_{i=1}^{N} \omega_t^i S_t^i$.

4.2 Computation of Contour Feature and Likelihood Function

The calculation of contour feature points is a process combined image segment and boundary points search. As the contour depends much on the quality of the image foreground, an efficient image segment method is need, which is insensitive to the shadow and change of lighting. An adaptive background subtraction is applied for this purpose. The process of human body contour is shown in Figure 3. The main steps of contour calculation include: background subtraction, fake patches judgment and contour extraction. For eliminating the influence of shadow, the color space of original video image is converted from *RGB* to *YH*, which *Y* is the luminancy components and *H* is the hue components. First, we compute the expectation E_Y, E_H and variance D_Y, D_H for *Y* and *H* from a certain number of background images. These variance is the adaptive threshold of segmentation, for the derivation of each pixel to its mean value is central distribution. When subtracted the background, if just the *Y* component has difference with the background model, while the *H* is stable, we consider this pixel is a shadow pixel. After subtraction, the fake foreground and background patches caused by similar fore-color or little change of background are fixed by judging the areas of connection regions. If the area of one connection is little enough, this region is considered to be fake, that means if it is a back-hole, it is judged to

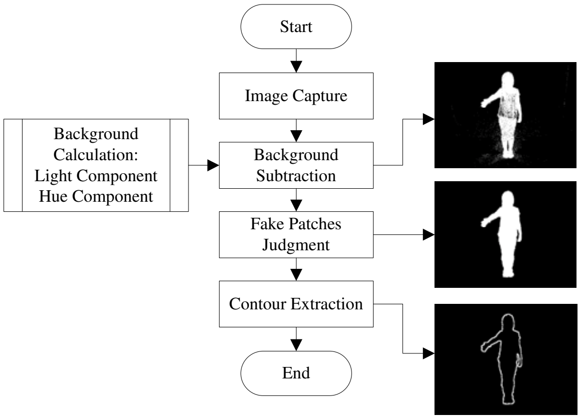

Figure 2: *The contour calculation process*

be a fore-region and if this region is a fore-region, it is deleted as noise. Then the contour points Y_c^I are computed after the suitable foreground of the human body are got.

After the contour Y_c^I is extracted from video frame and the contour Y_c^S is registered with particle *S*, the likelihood value is calculated using Equation 2. the chamfer distance of each point in the appearance model is calculated, which is the distance to the nearest extracted contour pixel in the video frame. For all the points in the appearance model, a vector is composed by these chamfer distances as $R = [r_1, r_2, \dots, r_{M_c}]_T$, we compute the samples weight using Equation 2, in which each $r_i(Y_c^S, Y_c^I)$ is one element of this vector.

4.3 Skin Color Constrain

Skin color is often the stable feature for locate human face and hands in tracking systems. It allows fast processing and is invariant to the change of motion state. When the limb contours are lost due to occlusions, the skin detection can be used to assist fixing the body parts. The skin detection algorithm in this paper should have the following two properties: fast and no need training. In this paper, the region based color segment method is used to detect the skin. For the hue of human skin is stable in a range, the skin parts of body is detected, including head, arms, especially the hands, as shown in Figure 4. When the limb occlusion is large, the skin detection is used to constrain the sampled particles. When the occlusion of limbs is large, we calculated the ratio of pixels as skin or non-skin for right and left arms respectively. Then the ratio is given to corresponding elements of particle as their new weight. The weights for elements of left limbs and right limbs are defined as ω^l, ω^r is calculated as below:

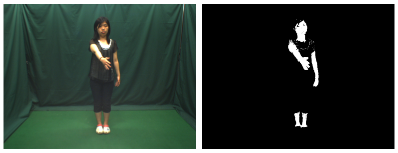

Figure 3: *The skin parts of body*

$$\omega^{l|r} = \frac{1}{M_s^{l|r}} \Sigma_{i=1}^{M_s^{l|r}} p(i) \qquad (3)$$

In equation(3), $M_s^{l|r}$ is the skin points number of left or right limbs, $p(i) = 1$if the i^{th} point belongs to the skin and $p(i) = 0$if not. At last, we use particle $\{\omega^l S_l^1, \dots, \omega^l S_l^{\frac{N}{2}}, \omega^r S_r^1, \dots, \omega^r S_r^{\frac{N}{2}}\}$ to replace the previous $\{S^1, S^2, \dots, S^N\}$ whose weights are only calculated through contour.

5 Experiment Results

In this section experiments for tracking human motion are given based on the proposed tracking algorithm. We also compared our method with edge-based method and give the result analysis.

The camera we used is FL2G-13S2C-C from PointGrey Company with 4mm lens of FV0420. It has various acquisition modes and we chose one with 30fps frame rate and 640*480 image resolution. Focusing on the upper limbs, the tracking result is presented by a 2.5D cardboard model which was projected by the 3D human articulated model. Assuming the torso is static, there're altogether 8 DoF for tracking, each arm 4 DoF. For tracking a human who is oriented towards the camera, this reduction of model is acceptable. The initialization of appearance model for edge cue and contour cue is shown in Figure 5. We use the measurement time and optimization time of edge cue method and contour cue method to denote the efficiency of these two cues. 200 particles are sampled for estimating. From Table 1, we can see the contour cue method saved much time at the likelihood measurement stage. This is because the contour points is much fewer than the edge points. The tracker score proposed by [Wang and Rehg 2006] is used to evaluate the tracking quality. For all the contour points in the appearance model, a residual vector is composed by these chamfer distances:$R = [r_1, r_2, \dots, r_{M_c}]^T$. The tracker score *TS* is computed as the absolute value of the residual vector.

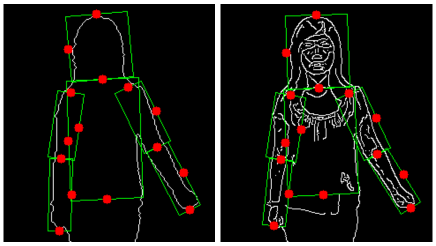

Figure 4: *(a) contour appearance (b) edge appearance*

Table 1: *Computation time of the two methods*

Procedure	Contour(ms/f)	Edge(ms/f)
Preprocess	56	34
Measurement	71	128
Optimization	228	576
Total	355	738

$$TS = R^T R \qquad (4)$$

A smaller score denotes a better tracking. From Figure 6 we can see, the contour cue method is better than edge cue. While at some frames the result is bad, these are mainly due to occlusions. Euclidean distance between tracked joints P_J^T and manually labeled joints P_J^R is used to denote the accuracy of the tracking. It is calculated as:

$$D = \sum_{i=1}^{M_J} \sqrt{(x_i^T - x_i^R)^2 + (y_i^T - y_i^R)^2} \qquad (5)$$

Figure 7 shows the distance in three situation: edge cue only, contour cue only, and contour cue with skin constraint. We can see the

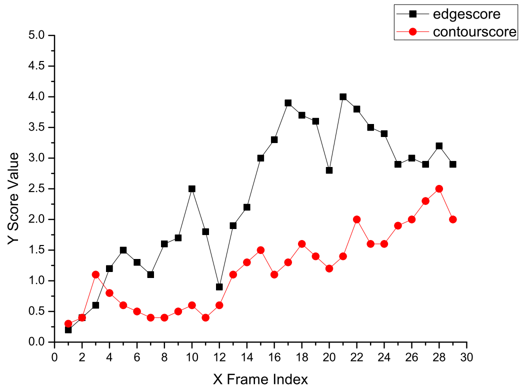

Figure 5: *The tracker score of two feature cues*

result of contour cue is much closer to the ground truth data than that of edge cue. While at some frames, for example at frame 3, the distance is high. This is mainly caused by the limb occlusions. With the assistance of skin detection, the distance keeps low and stable. The tracking results with cardboard bounding of the contour cue and edge cue method are shown in Figure 8. We can see that contour cue method is more stable and accurate in tracking than edge cue based tracking.

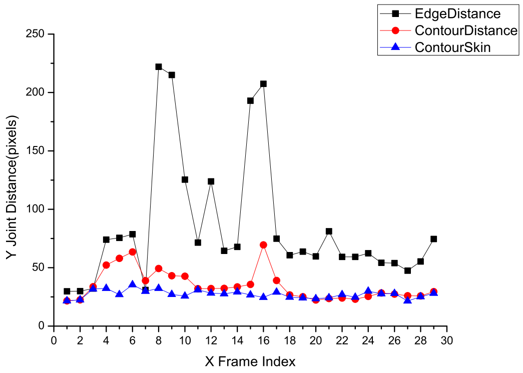

Figure 6: *The Euclidean distance between tracking joint position and manual joint position*

6 Conclusions

This paper proposed a probability framework which used the optimized particle filter algorithm to track human motion in a markerless monocular video. When measuring the likelihood of particles, we use contour cue to calculate the likelihood values. In the self-occlusion situation, we detect the skin parts to assist the sampling process, making a more reliable particle set. The experiment results showed that the proposed contour cue based particle filter algorithm is more efficient both in time and in space than the common feature-based methods. Meanwhile, with the assistance of skin, it can also produce more stable and higher quality tracking results.

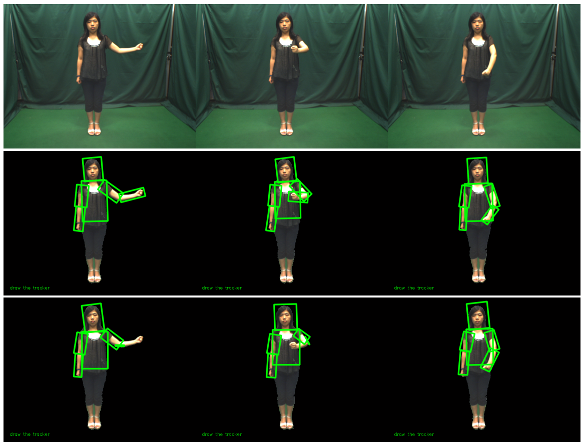

Figure 7: *The tracking results images with contour cue: the first row is the original image, the second is the contour cue based result, the last row is the edge cue based result.*

7 Acknowledgment

This paper is supported by the National Grand 973 Program of China (Grant No. 2009CB320805) and the National Natural Science Foundation of China (Grant No.60903064).

References

AGARWAL, A., AND TRIGGS, B. 2004. Tracking articulated motion using a mixture of autoregressive models. In *ECCV04*, Vol III: 54–65.

DEUTSCHER, J., AND REID, I. 2005. Articulated body motion capture by stochastic search. *Int. J. Comput. Vision 61*, 2, 185–205.

GREST, D., HERZOG, D., AND KOCH, R. 2006. Monocular body pose estimation by color histograms and point tracking. In *DAGM06*, 576–586.

RONALD, P. 2007. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding 108*, 1-2, 4–18.

SCHMIDT, J., FRITSCH, J., AND KWOLEK, B. 2006. Kernel particle filter for real-time 3d body tracking in monocular color images. In *FGR '06*, IEEE Computer Society, Washington, DC, USA, 567–572.

SMINCHISESCU, C., AND TRIGGS, B. 2003. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research 22*, 2003.

WANG, P., AND REHG, J. M. 2006. A modular approach to the analysis and evaluation of particle filters for figure tracking. In *CVPR '06*, IEEE Computer Society, Washington, DC, USA, 790–797.