A Multi-Cue-Based Human Body Tracking System[1]

Yihua Xu, Xiao Deng and Yunde Jia

School of Computer Science, Beijing Institute of Technology
Beijing 100081, PR China
{yihuaxu, dergx, jiayunde}@bit.edu.cn

Abstract. This paper presents a real-time vision-based system for the tracking of human upper body with both color images and depth maps. We combine the color-histogram-based particle filtering and mean shift algorithm to track face and hands and estimate other body parts by human kinetics. A multi-cue approach that integrates depth information and the color-based method is introduced to handle the rapid and complex motions of human hands. Real-time depth is recovered based on simple hardware configuration, which makes our system easy to be popularized in many real-world applications like digital entertainment. The system runs at 20 fps for images with 320x240 pixels on a 2.8GHz PC.

Keywords: Multi-Cue, Stereo Vision, Human Body Tracking.

1. Introduction

Visual tracking of human motion is receiving increasing attention from computer vision researchers due to its great potentials in many applications, such as human-computer interaction (HCI), visual surveillance, video conferencing and content-based image retrieval. Within last decades, lots of methods have been proposed in this area. Some of them use only a single visual cue like color and contour, and are unlikely to achieve robust tracking due to changing illuminations, cluttered background, rapid and complex human motions. To increase robustness, many approaches that combine multiple complementary cues have also been proposed. Isard and Blake [1] explored the ICONDENSATION algorithm through the combination of standard CONDENSATION [2] and importance sampling technique. In their work, a hand tracker integrating color blob-tracking with a contour model was demonstrated, which was experimented to be robust to rapid motion, heavy clutter and hand-colored distractors. Wu et al. [3] also combined color and contour cues for object tracking and achieved adaptation for the color model by co-inference technique. The integration of color and motion cues is also being widely used [4], [5]. Usually motion detection is used to find a coarse range of the object, and then color measurement is applied to search for the exact location of the object within the range.

[1] This work was partially supported by The High-Tech Program and Natural Science Foundation (60675021) of China.

It is more effective to apply depth than the above 2D image information to the subjects like background subtraction, occlusion and recovery of 3D postures. Nanda and Fujimura [6] presented a contour based tracking method in which edge detection is applied to depth map rather than gray image to increase the robustness towards highly cluttered environments. Darrell et al. [7] integrated depth, color and face detection into a single real-time person tracking system. Processing based on depth information was used to output the silhouettes of different users, which played an important role for robust multi-person tracking. With dense disparity map, Jojic et al. [8] explored a statistical image formation model. This model accounts for self-occlusions among the articulated 3D Gaussian models by picking the minimum depth. Towards the applications in HCI and digital entertainments, Li [9] proposed a 3D human body tracking and modeling system. Particle filtering relying on color information was used to estimate the 2D positions of body parts, and depth measurement was applied for the acquisition of 3D parameters. However, due to lack of effective motion prediction, the system performs poorly when the tracking object moves quickly and arbitrarily.

This paper proposes a real-time human upper body tracking system in which both color and depth cues are used to track face, hands and other parts, and recover the 3D posture of the upper body. Based on simple hardware configuration, a binocular stereo vision platform is built to product color images and dense depth maps in real-time. Depending on this platform, color-histogram-based particle filtering and mean shift algorithm are combined for the tracking of face and hands. Furthermore, in order to effectively handle the rapid and complex motion of human hands, it is necessary to increase the hit ratio of sampling of the particle filtering method. In our work, according to the analysis on depth map, hand candidate region can be extracted to build the importance function, which is then incorporated into the particle filtering framework to improve sampling efficiency. This importance sampling approach integrates the depth and color cues and therefore greatly increases the tracking robustness. Finally, given the tracking results of face and hands, we adopt some simple kinetic constraints to estimate other body parts and complete the 3D posture recovery.

The remainder of this paper is organized as follows: Section 2 describes the binocular stereo vision platform and the background subtraction process. Section 3 shows the combined algorithm of color-histogram-based particle filtering and mean shift tracking. Section 4 discusses the depth-based importance sampling technique. Section 5 refers to the 3D posture recovery. Experimental results are demonstrated in Section 6 and the conclusion is given in Section 7.

2. Depth Recovery and Background Subtraction

In this paper, a binocular stereo vision platform is built with two web cameras (Logitech QuickCam Pro 4000) and a personal computer, as pictured in Fig. 1(a). The flexible technique proposed in Zhang's work [10] is adopted to calibrate the stereo system. We firstly place a planar pattern like chessboard into the common view of the cameras and obtain their intrinsic and extrinsic parameters respectively, and then

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

estimate the rigid transformation between these two cameras by setting one of them as the reference. After calibration, the block-based stereo matching algorithm is employed to implement depth recovery. To achieve video frame rates, we take a computational optimization strategy based on incremental calculations [11].

Given the above platform, background subtraction is realized using the real time depth sequence. In previous works, color-based background models are widely used for static background subtraction [9]. However, such models run into trouble when the background or illumination varies significantly. In person tracking applications, however, usually there is distinct difference between the depths of the foreground and background. Therefore, threshold segmentation on the depth map can be used to extract foreground regions efficiently. The threshold depth can be simply predefined, or computed from the depth at person head which can be located by face detection algorithm. A result of the subtraction is shown in Fig. 1(d). It is supposed that the user is facing the camera and the upper body maintains upright.

Fig. 1. The binocular stereo vision platform and background subtraction (a. reference configuration; b. color image; c. depth map; d. mask of foreground region extracted by depth segmentation)

3. Color-Based method for Face and Hand Tracking

The system integrates color-histogram-based particle filtering [12] and mean shift alogrithm [13] to track face and hands. Particle filtering usually performs well even under the conditions of background distractions and partial or intermittent occlusions. To improve tracking accuracy, the particle filtering tracking result is used as the initial solution of the subsequent mean shift computation, which provides an optimization process to obtain more accurate result based on the deterministic search strategy.

In the framework of particle filter, hypothesized object region can be parameterized as rectangle $W=(x,y,\alpha\cdot w,\alpha\cdot h)$, where (x,y) is coordinates of rectangle center, (w,h) is the constant reference size as prior knowledge, and α acts as the scale factor. The state evolution is described by the model $x_t - x_{t-1} = x_{t-1} - x_{t-2} + \boldsymbol{w}_t$, where $\boldsymbol{w}_t$ is the white noise. At time t, the normalized color histogram $H_n = \{h_{i,n}\}_{i=1}^{B-1}$ of the hypothesized region is calculated according to the H channel in HSV color space, where B is the histogram bin number. Then, the confidence π_t^n of sample $\boldsymbol{x}_t^n = (x_t^n, y_t^n, \alpha_t^n)$ can be evaluated based on the distance between H_n and reference color histogram H_{ref}:

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

$$\pi = \exp\{-D^2(H_n, H_{ref})\} \quad (1)$$

where $D(H_n, H_{ref})$ is derived from the Battacharyya similarity coefficient:

$$D(H_n, H_{ref}) = (1 - \sum_{i=0}^{B-1} \sqrt{h_{i,n} \cdot h_{i,ref}})^{1/2} \quad (2)$$

The state vector $\boldsymbol{x}_t$ is estimated with mean of all samples. Afterwards, starting with rectangular window determined by $\boldsymbol{x}_t$, the mean shift algorithm is applied on the skin-color probability map obtained by the back projection of the reference color histogram to the source image. Mean shift algorithm proceeds iteratively to move the window and locate it at the peak of probability distribution when algorithm converges, more details is introduced in [13]. Finally, we redistribute the samples of the particle filter according to the optimized tracking result.

Invalid color reference histogram due to changing illumination and pose will lead to low confidence of samples in particle filter. So, when the maximum confidence of all samples is less than a threshold in few consecutive frames, a new color reference histogram is calculated in the face region obtained by face detection method and is used to replace the old one. Besides, face detection is used for the initialization of face tracking, as described in [9].

4. Depth-Based Importance Sampling

In the standard formulation of particle filtering method, samples are only determined by the state density at the previous time step as well as the motion model, which we term standard sampling. Hand candidate region at the current time step obtained by depth segmentation can be adopted to generate the importance function. Then, the depth-based importance sampling is integrated into the framework of color-histogram-based particle filtering, which makes the tracking performance more robust even if the hand moves quickly and arbitrarily.

In order to find the hand candidate region, we assume that human motions are subjected to two constraints as follows based on the observation to general HCI: 1. The face and torso have the same depth value, and the hands are usually in front of the torso. 2. Depth value descends from shoulder to hand along the arm, so the region with local minimum depth value is considered as human hand. Based on the first constraint, the region including hand can be extracted according to the depth of face. Some parts of the arm might also be segmented out together with the hand, so we adopt an iterative method to search for the local minima of depth value according to the second constraint, and then determine hand candidate region. The procedure of hand candidate region extraction is given in Table 1.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Table 1. Algorithm of hand candidate region extraction

Step1: Calculate the mathematical expectation d_f and standard deviation sd_f of points within the face region in depth map (Fig. 2(a), 2(b)).

Step2: Generate image I by depth segmentation:

$\begin{cases} I(x,y) = d(x,y), & d(x,y) > ds_{thre} \\ I(x,y) = 0, & else \end{cases}$, where $d(x,y)$ is the pixel value at position (x,y) in depth map, $ds_{thre} = d_f + \lambda \cdot sd_f$ is the threshold and $\lambda(\lambda > 3)$ is a constant coefficient.

Step3: Apply connected component analysis to image I to eliminate noise and choose the component with proper size as the coarse hand candidate region (Fig. 2(c)). Draw the bounding rectangle R of the region (yellow window in Fig. 2(c)) and calculate its parameters: top-left corner (x_l, y_l), bottom-right corner (x_u, y_u) and geometric center (x_c, y_c).

Step4: Refine rectangle parameters using iteration algorithms and finally determine the region with local minimum depth:

a) Calculate the centroid (x_c', y_c') of R:

$x_c' = \sum_x \sum_y x \cdot d(x,y) / \sum_x \sum_y d(x,y)$

$y_c' = \sum_x \sum_y y \cdot d(x,y) / \sum_x \sum_y d(x,y)$, where $x \in [x_l, x_u], y \in [y_l, y_u]$.

b) Update the lower- and upper-boundary of R to make geometric center coincide with centroid:

$\begin{cases} x_l = x_u - 2(x_u - x_c') = 2x_c' - x_u, & x_c' > x_c \\ x_u = x_l + 2(x_c' - x_l) = 2x_c' - x_l, & x_c' < x_c \end{cases}$

$x_c = (x_l + x_u)/2$.

update y_l, y_u and y_c based on the same principle.

c) Iterate a) and b) until $|x_c - x_c'| < \varepsilon$ and $|y_c - y_c'| < \varepsilon$ or $|x_l - x_u| < \varepsilon'$ and $|y_l - y_u| < \varepsilon'$, where both ε and ε' are thresholds set in prior.

d) Set the rectangle $R = (x_c, y_c, w, h)$ (blue window in Fig. 2(c)) returned by above iterative method as the hand candidate region where (w,h) is the size of R.

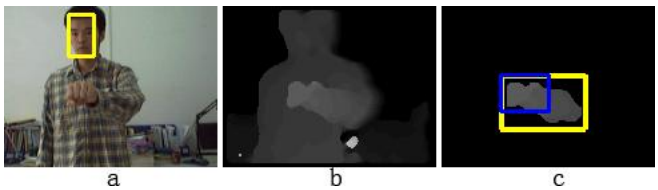

Fig. 2. Hand candidate extraction (a. the face tracking result; b. the smoothed depth map; c. the extracted results, yellow window is the bounding hand-covered region computed by depth segmentation and blue window is the hand candidate region extracted by the iterative method)

Based on the hand candidate region $R = (x_c, y_c, w, h)$, the importance function is defined as a 2D Gaussian distribution:

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

$$g(\boldsymbol{x}_{trans}) = N(\boldsymbol{u}, K) \quad (3)$$

where $u = (x_c, y_c)$ and $K = \Lambda\{(w/2)^2, (h/2)^2\}$ are the mean and covariance respectively. Importance sampling proceeds based on $g(\boldsymbol{x}_{trans})$ which generates the translation part $\boldsymbol{x}_{trans} = (x_t^n, y_t^n)$ of the sample vector $\boldsymbol{x}_t^n$, while α_t^n is computed by $\alpha_{t-1} + w_t$. Then, confidence of the sample is evaluated by comparing the color content of hypothesized region with reference color histogram, as described in previous section. During the tracking process, weights for importance and standard sampling are q and $1-q$ respectively.

At the initialization stage, all samples are generated through depth-based importance sampling, and the color reference histogram is obtained according to the color content of face. If average confidence of samples exceeds the threshold set in prior, we consider that initialization is successful.

Two candidate regions will be extracted given the assumption that the two hands never overlap when both moving in front of the torso. So we can use the two candidate regions to generate importance functions for the tracking of left and right hand respectively.

5. 3D Posture Recovery

In this paper, we use simple geometric figures to represent the body parts and recover the 3D posture by estimating the parameters of these figures, such as position, size, etc.

As shown in Fig. 3, in the 3D space, head (face), hands and torso are represented as rectangles while upper arms and forearms are modeled with line segments. For rough estimations, we consider that the size of torso, upper arm and forearm are in a fixed proportion to that of head and use these constraints to simplify the process of 3D posture recovery.

Firstly, given the depth map as well as the 2D tracking results, it is easy to obtain the 3D representations of head and hands based on the calibrated camera geometry. Then, we can determine the torso relying on the head, and denote the upper left and right vertexes of torso as the positions of the two shoulders respectively (Fig. 3). Given the 3D locations of hand and shoulder as well as the lengths of upper arm and forearm, the rough range of elbow can be found according to human kinetics. Finally, when a point within the range best matches its corresponding point in depth map, it is selected as the elbow. If no proper point to select due to hand-elbow occlusion, we set the position of elbow as $(X_h, Y_h, Z_h + l_f)$, where (X_h, Y_h, Z_h) is the position of hand and l_f is the length of forearm. The 3D posture recovering is shown in Fig. 3.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

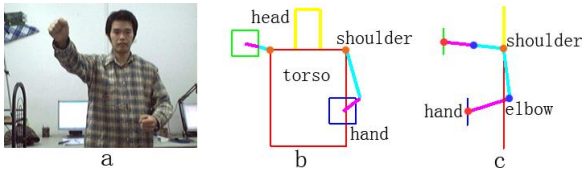

Fig. 3. 3D posture recovery (a. posture of user in real-world; b. frontal view of the recovered 3D posture; c. profile view of the recovered 3D posture)

6. Experimental Results and Application

Based on the methods described in the previous sections, we complete the implementation of the human body tracking system. It runs at 20 Hz for 320x240 resolution images on a 2.8G PC. A qualitative tracking experiment is shown in Fig. 4. Since depth-based importance sampling provides reasonable predictions, the tracking results are accurate even when the hands motion speeds up or turns around suddenly. We also developed a series of games in which the 3D posture recovered by human upper body tracking system is used to drive the virtual objects. Fig. 5 shows a music playing game in which the player is controlling Snoopy with his motions.

We integrate color-based particle filtering (Color-PF), depth-based importance sampling (Depth-IS) and mean shift (MS) to achieve better tracking performance compared with other two combination strategies (listed in Table 2), which is verified by experimental result demonstrated in Fig. 6 and Table 2. About 300 consecutive frames are used for test. Tracking accuracy of these three methods are evaluated according to the Euclidean distance between tracking results of hands and manually marked ground truth.

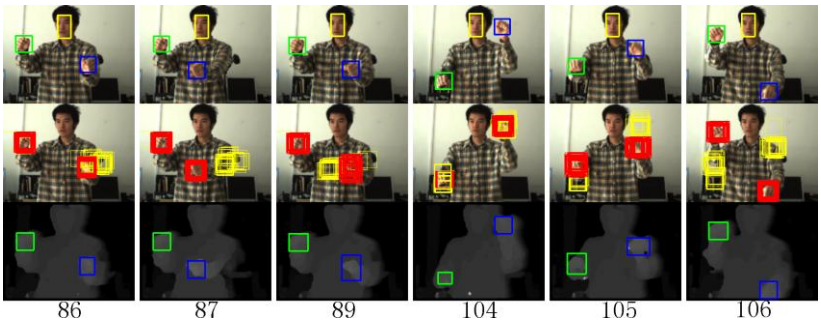

Fig. 4. Face and hands tracking by the proposed system (First row: tracking results with yellow, green and blue boxes denote face, left and right hand, respectively. Second row: red and yellow boxes denote the importance and standard sampling, respectively. Third row: candidate hand regions extracted from the depth maps. In this experiment, number of samples is 200 and the weight for importance sampling is 0.5)

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

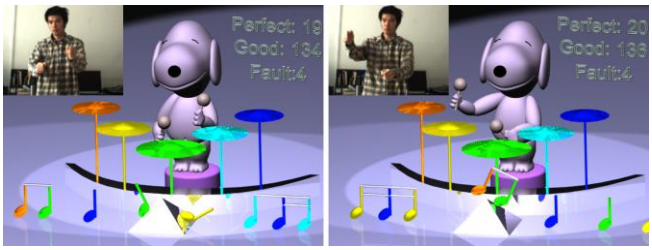

Fig. 5. Playing music with the proposed human tracking system.

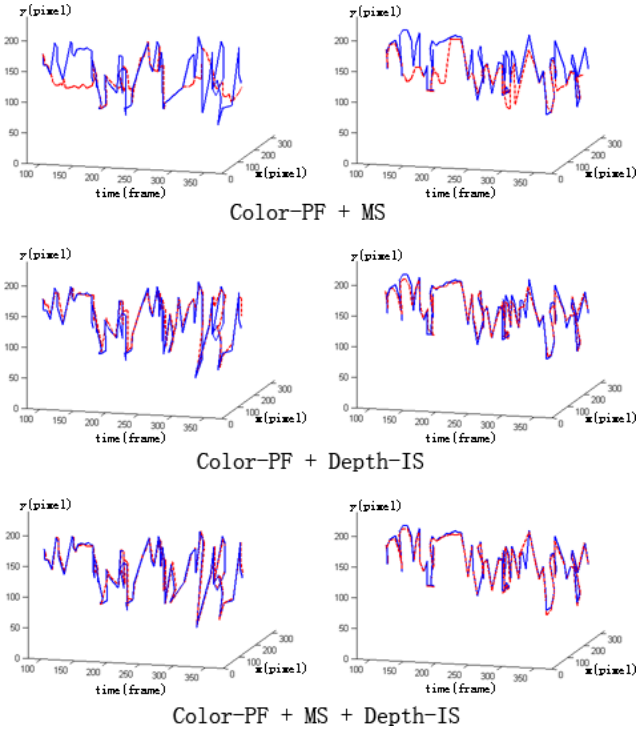

Fig. 6. The comparisons between ground truth and hands tracking results by three different combinations of methods (the two columns denote the tracking for left and right hand respectively, the blue and red curves denote the ground truth and the tracking result respectively)

Table 2. Accurate rates of the hand tracking by three different combinations of algorithms.

	Accurate Rate	
	Left hand	Right hand
Color-PF + MS	45.97%	51.21%
Color-PF + Depth-IS	86.00%	84.00%
Color-PF + MS + Depth-IS	98.95%	90.12%

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

7. Conclusions

In this paper, we developed a multi-cue based real-time human tracking system. Relying on a simple-configured binocular stereo vision platform, color and depth information are efficiently combined to achieve robust and accurate tracking. Experimental results show that the depth-based importance sampling greatly increases the tracking robustness to rapid and complex human motions. 3D postures are also recovered that makes possible the driving of virtual scenarios by user motions. The system can be applied to many real-world applications like HCI based digital entertainments.

References

1. Isard, M., Blake, A.: ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In Proc. Europ. Conf. Computer Vision, Vol. 1. (1998) 767-781
2. Isard, M., Blake, A.: CONDENSATION - Conditional Density Propagation for Visual Tracking. Int. J. Computer Vision, Vol. 29(1). (1998) 5-28
3. Wu, Y., Huang, T.S.: A Co-inference Approach to Robust Visual Tracking. In Proc. IEEE Int. Conf. Computer Vision, Vol. 2. (2001) 26-33
4. Pérez, P., Vermaak, J., Blake, A.: Data Fusion for Visual Tracking with Particles. Proc. IEEE, Vol. 92(3). (2004) 495-513
5. Vermaak, J., Pérez, P., Gangnet, M., Blake, A.: Towards Improved Observation Models for Visual Tracking: Selective Adaptation. In Proc. Europ. Conf. Computer Vision, Vol. 1. (2002) 645-660
6. Nanda, H., Fujimura, K.: Visual Tracking Using Depth Data. In Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop. (2004) 37-37
7. Darrell, T., Gordon, G., Harville, M., Woodfill, J.: Integrated Person Tracking using Stereo, Color, and Pattern Detection. In Proc. IEEE Conf. Computer Vision and Pattern Recognition. (1998) 601-609
8. Jojic, N., Turk, M., Huang, T.S.: Tracking Self-Occluding Articulated Objects in Dense Disparity Maps. In Proc. IEEE Int. Conf. Computer Vision, Vol. 1. (1999) 123-130
9. Jingfeng, Li, Yihua, Xu, Yang, Chen, Yunde, Jia: A REAL-TIME 3D HUMAN BODY TRACKING AND MODELING SYSTEM. In Proc. IEEE Conf. Image Processing. (2006)
10. Zhang, Z.: A Flexible New Technique for Camera Calibration. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22(11). (2000) 1330-1334
11. Stefano, L.Di, Marchionni, M., Mattoccia, S., Neri G.: A Fast Area-Based Stereo Matching Algorithm. In Proc. IAPR/CIPRS Int. Conf. Vision Interface. (2002) 146-153
12. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In Proc. Europ. Conf. Computer Vision. (2002) 661-675
13. Bradski, G.R.: Computer Vision Face Tracking as A Component of A Perceptual User Interface. Applications of Computer Vision Workshop. (1998) 214-219

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de