Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Visual Tracking Algorithm Based on CAMSHIFT and Multi-cue Fusion for Human Motion Analysis

Ge Yang
Key Laboratory of Integrated Microsystems,
Key Laboratory of Machine Perception,
Shenzhen Graduate School, Peking University, China
yangge@szpku.edu.cn

Hong Liu
Key Laboratory of Machine Perception,
Key Laboratory of Integrated Microsystems,
Shenzhen Graduate School, Peking University, China
liuh@szpku.edu.cn

***Abstract*— It is still a challenging problem for tracking objects in complex visual situations, such as an object is occluded or the object's color features are very similar to its background. Therefore, a novel visual tracking algorithm is proposed for multiple cues fusion based on three common cues: color, target position prediction and motion continuity in this paper. Color feature is free of translation and rotation and robust to partial occlusions and pose variations. Features of target position prediction and motion continuity can handle the condition that the color difference between the foreground and the background is similar. Combining with CAMSHIFT (Continuously Adaptive Mean Shift) technique, experimental results show that the proposed visual tracking algorithm is more robust than traditional single cue and gets better tracking effect than CMST (Collaborative Mean Shift Tracking). Successful rates of the proposed algorithm are 70% to 100% in 4 different complex conditions.**

Keywords*—Visual Tracking, Multiple Cue Fusion, CAMSHIFT, Occlusion Handling.

I. INTRODUCTION

With the development of information technology and intelligent science, computer vision has been the front of IT and high technology domain. It is a next problem how to use a computer to comprehend more video information. Visual analysis is important in HRI (Human Robot Interaction), intelligent control, virtual reality, image coding based on model, content search of streaming media and so on. It is necessary and urgent to study it.

Avidan et al. [1] regarded tracking as a binary classification problem. It trained a weak classifier to distinguish between object features and background features by an online way. A strong classifier was used to compute a confident image of a next frame. A peak value (a new position of an object) was found by mean shift algorithm. Through training a weak classifier a strong classifiers was updated. It made a tracker get more robust at low computation cost. Reference [1] could track a target at many scenes including a camera motion or stillness, a gray image or an infrared image, objects of different sizes. The tracking algorithm could handle some occlusion: classification and marking technologies were used to detect occlusion; a particle filter was used to overcome occlusion. The algorithm didn't handle the occlusion of a long time and a large range. Its feature space didn't consider space information.

For a non-rigid object Comaniciu et al. [2] proposed an object expression and a locating way. A locating issue was transformed to an issue of finding a local maximum interesting area. An object model and an object candidate area were showed through a Bhattacharyya coefficient. It could handle the condition that a camera moved, a target was partly occluded and an object size was changed and so on. It was not a tracking algorithm for a special task without combining with prior knowledge for a special task.

Dawei et al. [3] proposed an adaptive feature selective way in a mean shift tracking framework. A most decisive feature was chosen according to Bayesian error rate. A weight image was constructed. Mean shift technology was used to track an object. It supposed that Bayesian error rate of the selected features obeys Gaussian distribution. Its application field was limited. In tracking an object based on a kernel unchangeable model and unfit scale were two main limit factors. Anbang et al. [4] proposed a new tracking way based on a kernel for the two factors. It combined scale estimation with updating an object model. It was impacted by an object's scale and change of its appearance. However, it was validated through tracking a hand and was lack of generality.

Nouar et al. [5] proposed a tracking algorithm based on a two dimensional histogram of an image. The two color channels were selected based on a denotation standard. The standard was a most suitable expression of tracking an object and a least expression of meet background. It could track and manage the target whose shape and light changed very large. It only considered color cue and didn't consider other cues. If color feature was similar between a target and background tracking results wouldn't be very well.

Many researchers had studied a visual tracking algorithm based on multi-cue [6-9]. A traditional visual tracking algorithm based on a single cue didn't work well when environment changed. Shortcoming of a tracking algorithm based on color feature was that it would fail if color feature was similar between interesting objects and non-interesting objects or other areas [10]. Wu et al. [6] proposed a visual tracking algorithm based on cooperative study inference combined with multi-cue. It used a sequential Monte Carlo way. Multi-cue included shape feature and color feature. Cooperative study

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

inference was that the inference in high dimension space could be inferred through an iterative method from low dimension space. However, it is complex.

To increase reliability of visual tracking Tao et al. [7] proposed a dynamic Bayesian network way for multi-cue. Multi-cue included color of skin, ellipse shape and face detection. It combined multi-cue with concealed motion states. Tao et al. [7] used approximate inference based on a particle to estimate practice motion states. Simple linear Gaussian distribution need not be hypothesized. However, it only adapted a tracking way of a face. Cheng et al. [8] proposed a match way based on multi-cue of data fusion for a target motion or light change. Multi-cue included color feature, contour feature and target position prediction. It could adjust a weight value of every cue. But it didn't consider a situation of empty hole when bounder of a target was generated through contour feature. It didn't give out how to get a threshold value.

In this paper a visual tracking algorithm based on CAMSHIFT and multi-cue fusion for human motion analysis is proposed. Liu et al. [9] proposed CMST (Collaborative Mean Shift Tracking) algorithm. It combined color feature, position feature and prediction feature. It could update a weight value of every cue according to background. It used Mean Shift technology and auxiliary objects. However, it supposed that a model of background obeyed a single Gauss model and need train video sequences without a motion target before. This limited its application. Probability distribution of mean shift technology was based on static distribution and was not dynamically adjusted. The difference between [9] and our work is that we need not suppose a background model and train video sequences without a motion target before. We consider color feature, feature of target position prediction and feature of motion continuity and use CAMSHIFT technology and can dynamically adjust its probability distribution.

Section Ⅱ analyzes other visual tracking algorithms, especially their shortcoming. Section Ⅲ describes CMF (Visual Tracking Algorithm Based on CAMSHIFT and Multi-cue Fusion for Human Motion) algorithm based on CAMSHIFT (Continuously Adaptive Mean Shift) and multi-cue fusion. CMF algorithm, CAMSHIFT algorithm and CMST algorithm are analyzed through experiments in section Ⅳ. Results of experiments show that CMF algorithm is better than other algorithms even if color feature is similar between foreground and background. The whole paper is summarized and directions of future research are pointed to in section Ⅴ.

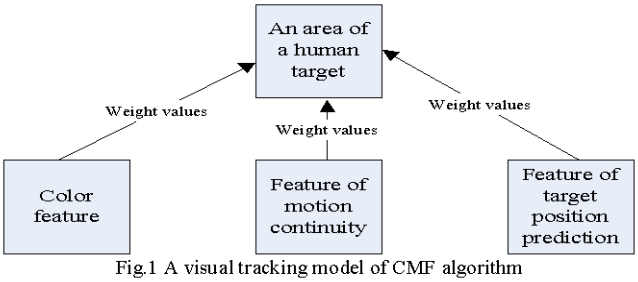


Fig.1 A visual tracking model of CMF algorithm

II. CMF ALGORITHM

CMF algorithm includes color feature, feature of target position prediction and feature of motion continuity, shown as Fig. 1.

A. *Color feature*

Color information is an important feature. It isn't impacted by rotation. It is robust for partly occlusion and gesture change. In this paper color feature acts as an important feature. CAMSHIFT technology realizes tracking. However, its tracking window is unchangeable. This will limit the management of appearance and occlusion. It needn't consider the situation that some background areas are taken as a part of a target and it can also realize tracking purpose.

A histogram uses m bins in this paper. There are n pixels. Their position and corresponding values in the histogram are $\{x_i\}_{i=1,\ldots,n}$, $\{q_u\}_{u=1,\ldots,m}$. Define a function, as in (1). The function denotes the discrete area value corresponding to every pixel. In a histogram, the uth color area corresponds to a value, as in (2) and (3).

$$b: R^2 \to \{1,\ldots,m\} \tag{1}$$

$$q_u = \sum_{i=1}^{n} \delta\left[b(x_i) - u\right] \tag{2}$$

$$p_u = \min\left(\frac{255}{\max(q_u)} q_u, 255\right) \tag{3}$$

In a color probability distributed image zero moment of a window area is (4).

$$M_{00} = \sum_x \sum_y p_u(x,y) \tag{4}$$

$$M_{10} = \sum_x \sum_y x p_u(x,y) \tag{5}$$

$$M_{01} = \sum_x \sum_y y p_u(x,y) \tag{6}$$

$$x = \frac{M_{10}}{M_{00}} \tag{7}$$

$$y = \frac{M_{01}}{M_{00}} \tag{8}$$

One moment of a window area is (5) and (6). Coordinates of a tracking point are (7) and (8).

B. *Feature of target position prediction*

This paper uses frame difference and computes difference values of every point between a prior and a rear frame, then judges which one is a moving point through setting a threshold [10]. If the threshold is set through experience there is much blindness. It will only adapt some special situations. Therefore this paper uses Otsu algorithm to dynamically determine the threshold F of frame difference [11] [12]. Basic idea of Otsu algorithm is that an appropriate t is found and makes scatter moment be least in a class. Namely, the threshold puts an image of frame difference partition two classes. It will make

the variance of the two divided classes be largest. In (9), 1 denotes pixels of foreground, 0 denotes pixels of background.

$$B=\begin{cases}1 & I(t)\text{-}I(t\text{-}2)>F\\ 0 & I(t)\text{-}I(t\text{-}2)\le F\end{cases} \quad (9)$$

Suppose that there are N gray values in an image of frame difference. There are m_i pixel points whose gray value is equal to i. A threshold F puts an image of frame difference partition two classes: M= [1, F] and Q= [F+1, N]. The probability that a gray value is i is P_i, as in (10). The probability that a pixel belongs to class M is P_M, as in (11). The probability that a pixel belongs to class Q is P_Q, as in (12). The average value of class M is μ_M, as in (13). The average value of class Q is μ_Q, as in (14). The average value of an image of frame difference is μ, as in (15). The variance between class M and class Q is σ^2, as in (16). F is set to make σ^2 be largest.

$$P_i=\frac{m_i}{\sum_{i=1}^{N} m_i} \quad (10) \qquad P_M=\sum_{i=1}^{F} P_i \quad (11)$$

$$P_Q=\sum_{i=F+1}^{N} P_i \quad (12) \qquad \mu_M=\sum_{i=1}^{F}\frac{iP_i}{P_M} \quad (13)$$

$$\mu_Q=\sum_{i=F+1}^{N}\frac{iP_i}{P_Q} \quad (14) \qquad \mu=\sum_{i=1}^{N} iP_i \quad (15)$$

$$\sigma^2=P_M(\mu_M-\mu)^2+P_Q(\mu_Q-\mu)^2 \quad (16)$$

C. Feature of motion continuity

In a short time among frames there is strong continuity for a target motion. The velocity of a target is considered to be unchangeable [8] [13]. Velocity of a tracked target is estimated through prior frames. A current position of a target is estimated through prior frames.

$$X(t,row)=X(t\text{-}1,row)\pm(X(t\text{-}1,row)-X(t\text{-}2,row)) \quad (17)$$

$$X(t,col)=X(t\text{-}1,col)\pm(X(t\text{-}1,col)-X(t\text{-}2,col)) \quad (18)$$

Suppose that X (t, *row*) is row coordinate of current position of a target at time t, as in (17). X(t, *col*) is column coordinate of current position of a target at time t, as in (18) and *rows* is the largest row number and *cols* is the largest column number in an image. Considering the continuity of human motion a current position is estimated through a linear predictor [14]. The relation among X (t, *row*) 、X (t-1, *row*) and X (t-2, *row*) is (19). The relation among X (t, *col*) 、X (t-1, *col*) and X (t-2, *col*) is (20).

$$X(t,row)\in[A,\min(B,rows)] \quad (19)$$

where A and B are defined as

$$A=max(X(t\text{-}1,row)-(X(t\text{-}1,row)-X(t\text{-}2,row)),1)$$

$$B=X(t\text{-}1,row)+(X(t\text{-}1,row)-X(t\text{-}2,row))$$

$$X(t,col)\in[C,D] \quad (20)$$

where C and D are defined as

$$C=max(X(t\text{-}1,col)-(X(t\text{-}1,col)-X(t\text{-}2,col)),1)$$

$$D=min(X(t\text{-}1,col)+(X(t\text{-}1,col)-X(t\text{-}2,col)),cols)$$

Suppose that row width of a tracking window is *width* and column width is *length*. Row coordinate of a target is (21) and column coordinate of a target is (22). Namely a target is in the rectangle area.

$$Y(t,row)\in[max(X(t,row)-width,1),min(X(t,row)+width,rows)] \quad (21)$$

$$Y(t,col)\in[max(X(t,col)-length,1),min(X(t,col)+length,cols)] \quad (22)$$

D. Tracking algorithm based on multi-cue fusion

This paper uses color feature, feature of target position prediction and feature of motion continuity. Even though a cue fails the tracking algorithm still works well. This increases its robustness.

Realizing steps of CMF algorithm are:

step 1: Set an interesting area I in a current image.

step 2: Set an initial position of a searching window and a selected position or an interesting area is the tracked target G.

step 3: Compute a probability distributed image M_1 of color feature and set weight value of the probability image M_1 be a_1.

step 4: Compute a probability distributed image M_2 of feature of target position prediction and set weight value of the probability image M_2 be a_2.

step 5: Compute a probability distributed image M_3 of feature of motion continuity and set weight value of the probability image M_3 be a_3.

step 6: Make a final probability distributed image M, as in (23).

step 7: Compute zero moment and one moment of a window area in the probability distributed image, as in (4), (5) and (6). Iterate CAMSHIFT algorithm until the position coordinates don't evidently change through (7) and (8) or maximum times is reached. T is maximum times.

step 8: Compute an interesting area according to coordinates of the tracked target again and put them be an initial position and a tracking area of rear frames and return step 7.

$$M=M_1\times a_1+M_2\times a_2+M_3\times a_3 \quad (23)$$

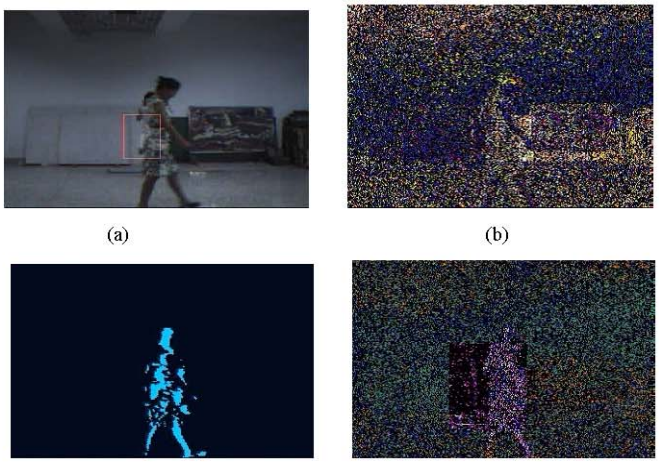
where $a_1+a_2+a_3$=1.

III. Experiment And Discussion

To validate CMF algorithm in this paper CMF algorithm and CAMSHIFT algorithm and CMST algorithm are applied to

video sequence 1, video sequence 2, video sequence 3 and video sequence 4, shown as Fig. 1. Setting T=15, $a_1 = a_2 = a_3$=1/3. A rectangle of a tracking window is used in this paper. The running condition of a tracking algorithm is CPU (P4 2.8G), memory (512M), hard disk (80G), operating system (windows XP), tool (Matlab7.1)

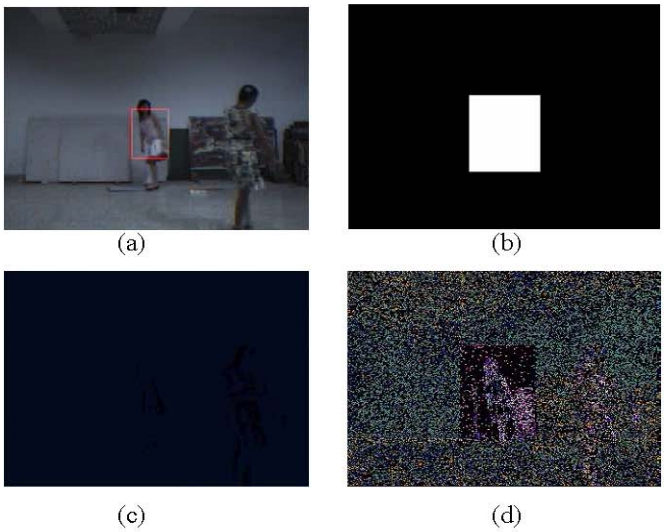
Video sequence1 (180 frames) is the condition that background and foreground are similar. In video sequence 2 (190 frames) target A and target B are tracked (target A is occluded for a long time). There is not evident occlusion in video sequence 3 (140 frames). A target is occluded for a short time in video sequences 4 (150 frames). Resolution of video sequence 1 is 320*256. Resolution of others is 640*512.



(a) (b) (c) (d)

(a)A tracked target image (b) A probability image of color feature (c) A probability image of feature of target position prediction (d) A probability image of multi-cue fusion

Fig. 2 Results of CMF (frame 96)

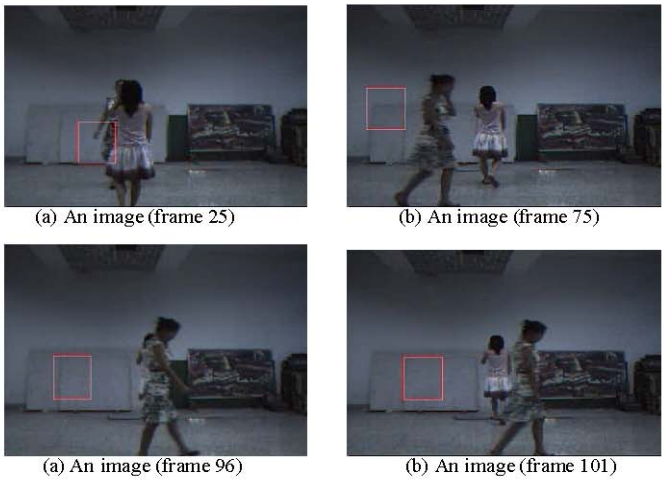


(a) (b) (c) (d)

(a) A tracked target image (b) A probability image of feature of motion continuity (c) A probability image of feature of target position prediction (d) A probability image of multi-cue fusion

Fig. 3 Results of CMF (frame 121)

For video sequence 1, results of CMF algorithm are Fig. 2-Fig. 3. When a target is occluded results are good, shown as Fig. 2. When a cue (feature of target position prediction) fails results of frame difference are 0. Tracking is realized according to other cues, shown as Fig. 3.

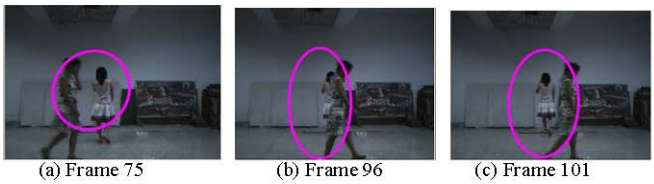


(a) An image (frame 25) (b) An image (frame 75)

(a) An image (frame 96) (b) An image (frame 101)

Fig. 4 Results of CAMSHIFT

Fig. 4 is experimental results of CAMSHIFT algorithm based on color feature. CAMSHIFT algorithm only considers color feature. When foreground and background is similar tracking results are not good. After frame 25 it can not track a target.



(a) Frame 75 (b) Frame 96 (c) Frame 101

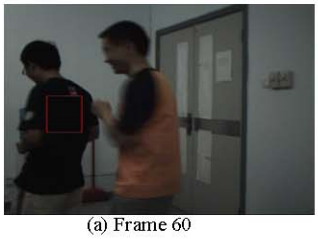
Fig. 5 Results of CMST

Fig. 5 is experimental results of CMST algorithm for video sequence 1. Because it uses a background subtraction way to get a target position and simulates background with a single Gaussian model the results of tracking are not precise.

(a) Frame 60 (d) Frame 165

Fig. 6 Results of CMF

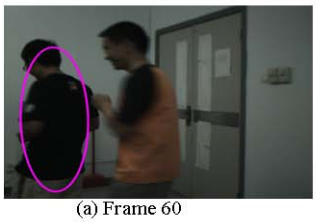
Fig. 6-Fig. 8 are experimental results of 3 algorithms for target A in video sequence 2. CMF algorithm works best. When target A is occluded for a long time it still works well. However, CAMSHIFT and CMST algorithms work badly.



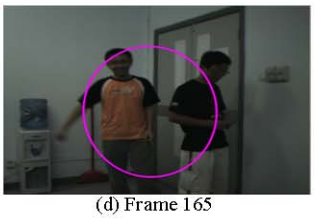
(a) Frame 60

(d) Frame 165

Fig. 7 Results of CAMSHIFT



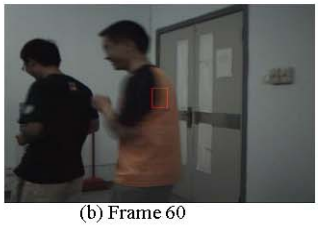
(a) Frame 60



(d) Frame 165

Fig. 8 Results of CMST

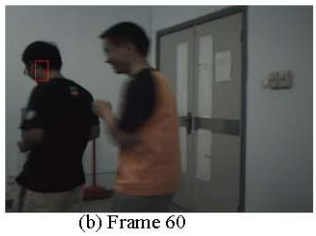
(a) Frame 47



(b) Frame 60

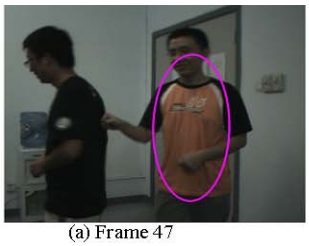
Fig. 9 Results of CMF

(a) Frame 47

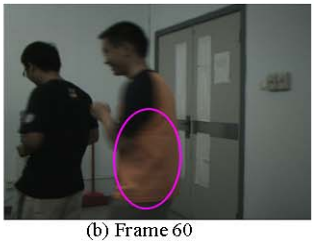


(b) Frame 60

Fig. 10 Results of CAMSHIFT



(a) Frame 47

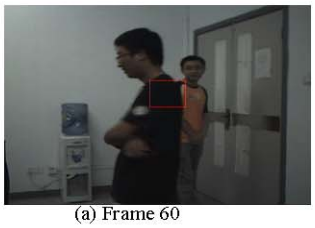


(b) Frame 60

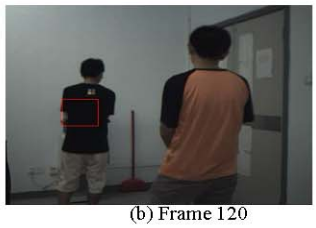
Fig. 11 Results of CMST

Fig. 9-Fig. 11 are experimental results of 3 algorithms for target B in video sequence 2. CMF algorithm works best and CMST algorithm works better for video sequence 2. When target B is occluded the algorithms still work well. CAMSHIFT algorithm fails when target B is occluded.

Fig. 12-Fig. 14 are experimental results of 3 algorithms for video sequence 3. 3 algorithms all work well because a target isn't obviously occluded.



(a) Frame 60



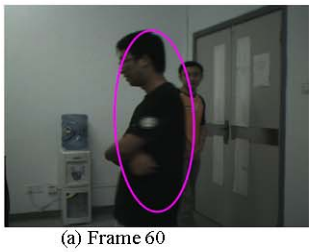
(b) Frame 120

Fig. 12 Results of CMF

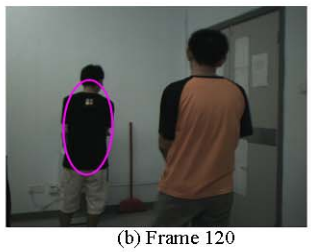
(a) Frame 60

(b) Frame 120

Fig. 13 Results of CAMSHIFT



(a) Frame 60



(b) Frame 120

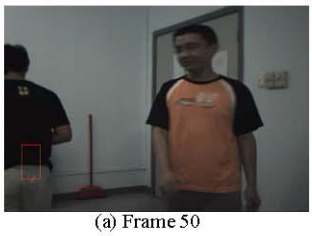
Fig. 14 Results of CMST

Fig. 15-Fig. 17 are experimental results of 3 algorithms for video sequence 4. CMF algorithm works best and CMST algorithm works better for video sequence 4. When a target is occluded the algorithms still work well. CAMSHIFT algorithm fails when a target is occluded.

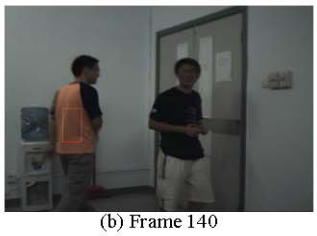
(a) Frame 50

(b) Frame 140

Fig.15 Results of CMF



(a) Frame 50



(b) Frame 140

Fig.16 Results of CAMSHIFT

(a) Frame 50

(b) Frame 140

Fig. 17 Results of CMST

Table I denotes the comparison of tracking success rate of 3 algorithms. Suppose that it fails that a tracking window includes other human targets or doesn't include a target in 1/2 of a tracking window. CMF algorithm works well in handling occlusion, especially the condition that color of background and foreground is similar.

Table I Comparison of tracking success rate of 3 algorithms

Video sequences	Algorithms	Successful tracked frames (total frames)	successful rate
video sequence 1	CMF algorithm	$167_{(180)}$	92.7%
video sequence 1	CAMSHIFT algorithm	$18_{(180)}$	10%
video sequence 1	CMST algorithm	$72_{(180)}$	40%
Target is A in video sequence 2	CMF algorithm	$133_{(190)}$	70%
Target is A in video sequence 2	CAMSHIFT algorithm	$76_{(190)}$	40%
Target is A in video sequence 2	CMST algorithm	$114_{(190)}$	60%
Target is B in video sequence 2	CMF algorithm	$161_{(190)}$	84.7%
Target is B in video sequence 2	CAMSHIFT algorithm	$114_{(190)}$	60%
Target is B in video sequence 2	CMST algorithm	$152_{(190)}$	80%
video sequence 3	CMF algorithm	$140_{(140)}$	100%
video sequence 3	CAMSHIFT algorithm	$140_{(140)}$	100%
video sequence 3	CMST algorithm	$140_{(140)}$	100%
video sequence 4	CMF algorithm	$141_{(150)}$	94%
video sequence 4	CAMSHIFT algorithm	$105_{(150)}$	70%
video sequence 4	CMST algorithm	$125_{(150)}$	83.3%

IV. CONCLUSIONS

A tracking algorithm is proposed based on CAMSHIFT and multi-cue for human motion analysis in this paper. It combines color feature, feature of target position prediction and feature of motion continuity. It can handle occlusion problems of a target. It need not suppose a background model and train video sequences without a target before. It works well even though the color feature between background and foreground is similar. Computation cost of CMF algorithm is not high. It can reach requirement of real time. Next work is that a dynamic system of cues selection should be designed.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (NSFC, No. 60675025 and No. 60975050) and National High Technology Research and Development Program of China (863 Program, No. 2006AA04Z247). Shenzhen Bureau of Science Technology and Information.

REFERENCES

[1] S. Avidan, "Ensemble Tracking", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, issue 2, pp 261-271, Feb. 2007.

[2] D. Comaniciu, V. Ramesh, P. Meer, "Kernel-Based Object Tracking", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, issue 5, pp 564-577, May 2003.

[3] Liang Dawei, Huang Qingming, Jiang Shuqiang, Yao Hongxun, Gao Wen, "Mean-Shift Blob Tracking with Adaptive Feature Selection and Scale Adaptation", IEEE International Conference on Image Processing, vol. 3, pp 369-372, Sept. 2007.

[4] Yao Anbang, Wang Guijin, Lin Xinggang, Wang Hao, "Kernel based articulated object tracking with scale adaptation and model update", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), pp 945-948, Mar. 2008.

[5] O.-D. Nouar, G. Ali, C. Raphael, "Improved Object Tracking With Camshift Algorithm", IEEE International Conference on Acoustics, Speech and Signal Processing (2006 ICASSP), vol. 2, pp 657-660, May 2006.

[6] Y Wu, T S. Huang, "Robust visual tracking by integrating multiple cues based on co-inference learning", International Journal of Computer Vision, vol. 58, issue 1, pp 55−71, 2004.

[7] Wang Tao, Diao Qian, Zhang Yimin, Song Gang, Lai Chunrong, G. Bradski, "A dynamic Bayesian network approach to multi-cue based visual tracking", The 17th International Conference on Pattern Recognition (ICPR 2004), vol. 2, pp 167-170, Aug. 2004.

[8] Cheng Ming-Yang, Wang Chun-Kai, "Dynamic Visual Tracking Based On Multi-Cue Matching", The 4th IEEE International Conference on Mechatronics (ICM2007), pp 1-6, May 2007.

[9] Hong Liu, Lin Zhang, Ze Yu, Hongbin Zha, Shi Ying, "Collaborative Mean Shift Tracking Based on Multi-Cue Integration and Auxiliary Objects", 14th IEEE International Conference on Image Processing (ICIP 2007), San Antonio, Texas on, pp 217-220, September, 2007.

[10] P´erez P, Vermaak J, Blake A, "Data fusion for visual tracking with particles", Proceedings of the IEEE, vol. 92, issue 3, pp 495−513, 2004.

[11] Wei Kaiping, Zhang Tao, Shen Xianjun, Liu Jingnan, "An Improved Threshold Selection Algorithm Based on Particle Swarm Optimization for Image Segmentation", Third International Conference on Natural Computation (ICNC 2007). vol. 5, pp 591-594, Aug. 2007.

[12] Wei Kaiping, Zhang Tao, He Bin, "Detection of Sand and Dust Storms from MERIS Image Using FE-Otsu Alogrithm", The 2nd International Conference on Bioinformatics and Biomedical Engineering(ICBBE 2008), pp 3852-3855, May 2008.

[13] Chen Mingyang, Wang Chunkai, "Dynamic Visual Tracking Based on Multi-cue Matching", International Conference on Mechatronics, Kumamoto Japan. pp 1-6, May 2007.

[14] N. Habili, Cheng Chew Lim, A. Moini, "Segmentation of the face and hands in sign language video sequences using color and motion cues", IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, issue 8, pp 1086-1097, Aug. 2004.