Abstract—A novel visual tracking algorithm based on visual attention and multiple cues fusion for human motion analysis is proposed in this paper. An auxiliary object is selected through visual attention mechanism. Feature of target position prediction and feature of motion continuity and color feature are used to determine the location of a target and an auxiliary object. Candidate color feature includes feature of hue and saturation, features of R (Red) channel, G (Green) channel, B (Blue) channel and linear combination of R, G and B when a target is tracked. Candidate color features with high reliability are dynamically selected and weight values of the color features are dynamically adjusted according to the change of a scene. Candidate color feature includes feature of hue and saturation when an auxiliary object is tracked. Combining with CAMSHIFT (Continuously Adaptive Mean Shift) technique, experimental results show that this new algorithm is more robust than the traditional static non-adaptive algorithm and gets better tracking effect than CMET (Collaborative Mean Shift Tracking). It can handle the situation that a target is occluded and that a target and its background color are similar. It can also track a cross-border target.

Keywords—Visual Tracking; Visual Attention; Multi-cue Fusion; Human Motion.

I. INTRODUCTION

With the development of information technology and intelligent science, computer vision has been the front of IT and high technology domain. Computer vision can be applied to HRI (Human Robot Interaction), intelligent control, virtual reality, image coding based on a model, content search of streaming media and so on. Visual tracking is one core content of computer vision. It is to find the place where a moving target is in image sequences. Namely, the tracked target is marked in different frames of a video. It is necessary and urgent to research it.

Many researchers study the tracking algorithms based on multi-cue fusion [1-11], including three classes: visual tracking algorithms based on adaptively adjusting weight values of cues [1-4, 11]; visual tracking algorithms based on dynamically choosing cues [5-9]; visual tracking algorithms based on adaptively adjusting weight values of cues and dynamically choosing cues [10]. A traditional tracking algorithm based on a single cue only considers one feature. Tracking results are not ideal. When situation changes the target can't be tracked. For example, the biggest shortcoming of a tracking algorithm based on color is that tracking will fail when color feature of non-interesting objects and interesting objects is similar [4]. Therefore, color feature, features of target position prediction and motion continuity are considered in this paper.

Wu et al. [1] proposed a visual tracking algorithm based on cooperative study inference combined with multi-clue. It used a sequential Monte Carlo way. Multi-clue included shape feature and color feature. Cooperative study inference was that the inference in high dimension space could be inferred through an iterative method from low dimension space. But it was much complex.

To increase reliability of visual tracking, Wang et al. [2] proposed a dynamic Bayesian network way for multi-clue. Multi-clue included color of skin, ellipse shape and face detection. It combined multi-clue with concealed motion states. Wang et al. [2] used approximate inference based on a particle to estimate practice motion states. Simple linear Gauss distribution need not be hypothesized. However it only adapted a tracking way of a face.

Cheng et al. [3] proposed a match way based on multi-clue of data fusion for a target movement and illumination change. Multi-clue included color feature, contour feature and target position prediction. It could dynamically adjust a weight of every clue. 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.

Perez et al. [4] proposed a particle filter tracking way based on color feature, motion feature and sound feature. It implemented a tracking way that used sound locating cue. It also realized a tracking way that combined color feature with motion locating cue. Color feature used RGB (Red, Green and Blue) channels. It defined the color likelihood model that consulted Bhattacharyya comparability coefficient. Motion feature was obtained through computing frames difference. Sound feature was obtained through measuring the time delay of audio frequency signal that got to 2 microphones. But its computation amount is large.

Collins et al. [5] proposed an online feature choosing mechanism. In tracking process it often estimated and adjusted features. It ranked every feature according to the capacity of
distinguishing a target sample distribution and background sample distribution. It selected the fittest color feature from 3 different color channels. Tracking was realized through combining a feature estimating mechanism with Mean Shift algorithm. But it only considered color feature. It didn't consider other features and was lack of robustness.

Triesch et al. [6] proposed a tracking system based on multi-feature. It combined luminance feature, color feature, movement continuity feature and feature of shape with contrast field feature. These cues used a democracy fusion way. Namely, the tracking system tried to obtain the most consistent among different cues. Every cue must be adaptively adjusted. But its computation amount is too large.

Yin et al. [7] transformed a tracking problem into a classification problem. Yin et al. [7] considered space information and time information. It segmented an image into smaller regions using space segmentation and occupying ways. It distinguished foreground and background through choosing different features. It combined motion estimation of foreground with motion segmentation. Every weight value image of space regions and motion information was combined into a final joint weight image. It realized tracking by Mean Shift algorithm. But it didn't resolve drift problems.

Wang et al. [8] extended standard Mean Shift algorithm. It chose reliable features according to discriminating capability of color feature and shape-texture feature. Shape-texture feature was described by direction histogram. It defined likelihood of a feature. It chose the feature whose variance was the biggest as a tracking cue. A target model was changed according to the similarity between initial and current models. But it didn't handle the situation that a target was occluded.

Wang et al. [9] proposed an online selective feature way from some feature space. It increased robustness of the algorithm. It combined particle filter. It sorted every feature using Fisher discrimination way. Yeh et al. [10] transformed a feature choosing problem into finding good feature subset problem. This feature subset composed a mixture likelihood image. But it is complex and is difficult to realize them.

Liu et al. [11] proposed CMET (Collaborative Mean Shift Tracking) algorithm. It combined color feature, position feature and prediction feature. It could update a weight value of every clue according to the background. It used Mean Shift technology and auxiliary objects. But it supposed that a model of background obeyed a single Gauss model and need train video sequences without moving targets before. This limited its application. Probability distribution of mean shift technology was based on static distribution and was not dynamically adjusted. This paper proposes a visual tracking algorithm based on CAMSHIFT and multi-clue fusion for human motion analysis. It need not suppose a model of background and need not train video sequences without a moving target before. It considers color feature, feature of target position prediction and feature of motion continuity. It uses CAMSHIFT technology and can dynamically adjust its probability distribution.

Visual attention is a key task in human vision research field. It is to study what human pay attention to when human watch an image. In a physical computer system resource is limit and a lot of information need be processed per second in a visual tracking system. Both are incompatible. It makes computation be bottleneck. The way that solves the problem is to process firstly part of an input image. Processing focus transfers from a place to another. Namely, prior processing part of an image and processing sequence are chosen through visual attention. Visual attention is a sequence of policy to reduce computation cost in search process of visual field [12].

There are two kinds in visual attention [13]-[20]: 1) data/stimulus-driven, independent-task bottom-up attention. Color, luminance and motion are normal stimulus (features). It only considers the sensory information. It doesn't consider the impact of special task or target. At the situation the results of pure bottom-up model don't match with users' actions. 2) dependent-task top-down attention. Priori knowledge, memory and target are normal factors. It is not sufficient that sensory attention is stimulated to pure top-down process in visual reality. A pure top-down model doesn't consider the reaction for virtual person to sensory stimulus. In recent years, some ways of visual attention mechanism are applied to image processing. It gradually becomes one of research hotspots.

Traditional visual attention is based on spatial location hypothesis. Sun et al. [20] proposed object-driven and feature-driven visual attention. Computer vision uses visual attention mechanism to decrease important information about visual behavior or visual tasks. This will handle the balance among computation resource, time cost and visual tasks in different normal situation. Sun et al. [20] analysed the three problems that need be handled. 1) How can the visual system know what information is important enough to catch attention? There are 2 ways to resolve: bottom-up information and top-down information. Bottom-up information includes basic features, for example color, orientation, motion, depth, conjunctions of features and so on. At that time salient features are used to attract visual attention. If visual attention is attracted to other places purely bottom-up way doesn't resolve this situation. Top-down information about current visual behavior is used to attract visual attention. 2) How does the visual system know when and how to direct attention and choose important information rather than doing so at random times and by random selection? 3) Where is (are) the next potential target(s) of visual attention shifts? That is, how does attention know where to go and what to do next? There are two hypotheses: space-based attention principle and object-based attention principle. The space-based attention principle allocates attention to a space region. It handles everything in the region. The object-based attention principle allocates attention to an object or kinds of objects. It handles everything in the object, not a space region. The different between both is that basic unit of visual attention. The object-based attention principle discretely chooses an object, rather than a space region. They use different ways to build and combine a low feature image and to simulate control mechanism of visual attention transfer.

Current research focuses on simulating visual attention mechanism of natural environment and uses a good model to simulate it. A saliency region is extracted through visual attention mechanism in this paper. To increase robustness of the tracking algorithm the objects that have correlation with a
target are chosen from a saliency region as auxiliary objects and multi-feature is used in this paper. This paper proposes visual tracking algorithm based on visual attention and multi-cue fusion for human motion analysis. It uses feature of target position prediction and feature of motion continuity and color feature. In 14 candidate color features reliable features are dynamically chosen and their weight values are adaptively adjusted. New color reliable function and feature of target position prediction and feature of motion continuity are defined in this paper. It doesn't suppose a model of background and doesn't train a video sequence without moving targets.

Section I analyzes other visual tracking algorithms, especially their shortcoming. Section II describes visual attention mechanism including bottom-up visual attention and top-down visual attention. Section III describes VTA algorithm (Visual Tracking Algorithm based on Visual Attention and Multi-cue Fusion for Human Motion Analysis) based on CAMSHIFT (Continuously Adaptive Mean Shift) and multi-cue fusion. VTA algorithm, CAMSHIFT algorithm and CMET algorithm are analyzed through experiments in section IV. Results of experiments show that VTA algorithm is better than other algorithms even if background and so on into a single saliency map [21].

In this paper bottom-up visual attention mechanism extracts the regions that are different from other regions all round them as candidate auxiliary objects. Top-down visual attention mechanism extracts compatible auxiliary objects from candidate auxiliary objects through prior knowledge. Namely, a target and its auxiliary objects appear at the same time. The auxiliary objects have high correlation with the target. The auxiliary objects are easy to be tracked [22]. This may track a target and its auxiliary objects and increase the robustness of this algorithm. Auxiliary objects and correlation between a target and auxiliary objects are used when a target is difficult to be tracked, shown as Fig. 1.

II. VISUAL ATTENTION MECHANISM

Visual saliency region is the region which is prominent in an image. Saliency map is the two-dimension image that indicates visual salient of every region in an image. It combines low image features (color, luminance and movement and so on) into a single saliency map [21].

In this paper bottom-up visual attention mechanism extracts the regions that are different from other regions all round them as candidate auxiliary objects. Top-down visual attention mechanism extracts compatible auxiliary objects from candidate auxiliary objects through prior knowledge. Namely, a target and its auxiliary objects appear at the same time. The auxiliary objects have high correlation with the target. The auxiliary objects are easy to be tracked [22]. This may track a target and its auxiliary objects and increase the robustness of this algorithm. Auxiliary objects and correlation between a target and auxiliary objects are used when a target is difficult to be tracked, shown as Fig. 1.

A. Bottom-up visual attention mechanism

A color image is made into Gaussian pyramids whose scales are $\sigma = [0, \ldots, 8]$ through constantly convolving with Gaussian filter and double sampling process. Resolution of the image whose scale is $\sigma$ is $1/2^\sigma$ of the original image. A Gaussian pyramid of every scale of the original image is made into a series of feature images. For every scale a Gaussian pyramid is made into three color channels, an intensity channel and four orientations [23].
III. MULTI-CUE FUSION ALGORITHM

A. Feature of target position prediction

This paper uses frame difference and computes difference values of every point between a prior frame and a rear frame, then judges which one is a moving point through setting a threshold [24]. 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 [25] [26]. Basic idea of Otsu algorithm is that an appropriate \( F \) 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 (1), \( B \) denotes pixels of foreground, 0 denotes pixels of background.

\[
B = \begin{cases} 
1 & I(t) - I(t-2) > F \\
0 & I(t) - I(t-2) \leq F 
\end{cases} 
\]  

(1)

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 (2). The probability that a pixel belongs to class \( M \) is \( P_M \), as in (3). The probability that a pixel belongs to class \( Q \) is \( P_Q \), as in (4). The average value of class \( M \) is \( \mu_M \), as in (5). The average value of class \( Q \) is \( \mu_Q \), as in (6). The average value of an image of frame difference is \( \mu \), as in (7). The variance of the two divided classes is \( \sigma^2 \), as in (8). \( F \) is set to make \( \sigma^2 \) be largest.

\[
P_i = \frac{m_i}{\sum_{i=1}^{N} m_i} 
\]

(2)

\[
P_M = \sum_{i=1}^{F} P_i 
\]

(3)

\[
P_Q = \sum_{i=F+1}^{N} P_i 
\]

(4)

\[
\mu_M = \sum_{i=1}^{F} i P_i 
\]

(5)

\[
\mu_Q = \sum_{i=F+1}^{N} i P_i 
\]

(6)

\[
\mu = \sum_{i=1}^{N} i P_i 
\]

(7)

\[
\sigma^2 = P_M (\mu_M - \mu)^2 + P_Q (\mu_Q - \mu)^2 
\]

(8)

B. 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 [27]. Velocity and the current position of a target are estimated through prior frames.

\[
X(t, row) = X(t-1, row) \pm (X(t-1, row) - X(t-2, row)) 
\]

(9)

\[
X(t, col) = X(t-1, col) \pm (X(t-1, col) - X(t-2, col)) 
\]

(10)

Suppose that \( X(t, row) \) is row coordinate of current position of a target at time \( t \), as in (9). \( X(t, col) \) is column coordinate of current position of a target at time \( t \), as in (10).

Suppose that row width of a tracking window is \( width \) and column width is \( length \). Row coordinate of a target is (11) and column coordinate of a target is (12). Namely a target is in the rectangle area, \( rows \) is the largest row number and \( cols \) is the largest column number in an image.

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

(11)

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

(12)

C. 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. Candidate color features include hue and saturation feature, red channel feature (R), green channel feature (G), blue channel feature (B) and their line combination, as in (1). Some combination is redundant, for example R+B+G and -R-G-B. The results of both are same. Only one is reserved. There are 14 combinations in candidate color features, as in (13).

\[
F = \{a \times R + a_2 \times G + a_3 \times B\} \quad a \in \{-1, 0, 1\} 
\]

(13)

A histogram uses \( m \) bins in this paper. There are \( n \) pixels. Their position and corresponding values in the histogram are \( \{x_i\}_{i=1, \ldots, m} \times \{q_u\}_{u=1, \ldots, m} \) (R channel feature, G channel feature, B channel feature and linear combination) or \( \{q_u\}_{u=1, \ldots, m} \), \( \{v_u\}_{v=1, \ldots, m} \) (hue and saturation feature). In a histogram, the \( u \)th color area corresponds to a value.

Define a function, as in (14). The function denotes the discrete area value corresponding to every pixel. In a histogram, the \( n \)th color area corresponds to a value, as in (15) or (15’) and (16) or (16’).

\[
b : R^2 \rightarrow \{1, \ldots, m\} 
\]

(14)

\[
q_u = \sum_{i=1}^{n} \delta[b(x_i) - u] 
\]

(15)

\[
q_{u(v)} = \sum_{i=1}^{n} \delta[b(x_i) - u(v)] 
\]

(15’)

\[
p_u = \min \left( \frac{255}{\max(q_u)}, 255 \right) 
\]

(16)

\[
p_{u(v)} = \min \left( \frac{255}{\max(q_{u(v)})}, 255 \right) 
\]

(16’)

In a color probability distributed image zero moment of a window area is (17).

\[
M_{00} = \sum_x \sum_y p_u(x, y) 
\]

(17)
\[ M_{10} = \sum_{x} \sum_{y} x^2 p_u(x, y) \]  
\[ M_{01} = \sum_{x} \sum_{y} y^2 p_u(x, y) \]
\[ \frac{x = \frac{M_{10}}{M_{00}}}{y = \frac{M_{01}}{M_{00}}} \]

One moment of a window area is (18) and (19). Coordinates of a tracking point are (20) and (21).

D. Multi-cue fusion

Value \( i \) is the value of feature \( k \), suppose that \( H_1^k(i) \) denotes a histogram of feature values for pixels in a target region. \( H_2^k(i) \) denotes a histogram of feature values for pixels in background regions. \( p_k(i) \) is discrete probability distribution of a target region. \( q_k(i) \) is discrete probability distribution of background regions. \( L_i^k \) is log likelihood ratio of feature \( k \), as in (22), \( \delta > 0 \) and \( \delta \) is a very small value. \( \text{var} \ (L_i^k; p_k) \) is variance of \( L_i^k \) for target class distribution \( p_k(i) \), as in (23). \( \text{var} \ (L_i^k; q_k) \) is variance of \( L_i^k \) for background class distribution \( q_k(i) \), as in (24). \( \text{var} \ (L_i^k; R_k) \) is variance of \( L_i^k \) for target and background distribution, as in (25). \( V(L_i^k; p_k, q_k) \) is variance of \( L_i^k \), as in (26). \( \text{var} \ (L_i^k; R_k) \) denotes separable ability between a target and its background. The larger \( \text{var} \ (L_i^k; p_k, q_k) \) is, the more easily feature \( k \) can separate a target from background. At that time feature \( k \) is a suitable feature to track the target.

\[ L_i^k = \log \frac{\max \{ p_k(i), \delta \}}{\max \{ q_k(i), \delta \}} \]

\[ \text{var} \ (L_i^k; p_k) = \text{E}[L_i^k * L_i^k] - \left( \text{E}[L_i^k] \right)^2 = \sum_i p_k(i) * L_i^k * L_i^k \]
\[ \sum_i p_k(i)^2 L_i^k \]  
\[ \left( \sum_i p_k(i) * L_i^k \right)^2 \]  
\[ \text{var} \ (L_i^k; q_k) = \text{E}[L_i^k * L_i^k] - \left( \text{E}[L_i^k] \right)^2 = \sum_i q_k(i) * L_i^k * L_i^k \]
\[ \sum_i q_k(i)^2 L_i^k \]  
\[ \left( \sum_i q_k(i) * L_i^k \right)^2 \]  
\[ \text{var} \ (L_i^k; R_k) = \text{E}[L_i^k * L_i^k] - \left( \text{E}[L_i^k] \right)^2 = \sum_i R_k(i) * L_i^k * L_i^k \]
\[ \sum_i R_k(i)^2 L_i^k \]  
\[ \left( \sum_i R_k(i) * L_i^k \right)^2 \]

where \( R_k(i) = \frac{[p_k(i) + q_k(i)]}{2} \)

\[ V(L_i^k; p_k, q_k) = \frac{\text{var}(L_i^k; R_k)}{\text{var}(L_i^k; p_k) + \text{var}(L_i^k; q_k)} \]

VTA algorithm computes, tests and judges hue and saturation feature, R channel feature, G channel feature, B channel feature and reliability of linear combination of R, G and B. When their reliabilities change they realign according to their values. \( W \) is discrete probability whose \( V(L_i^k; p_k, q_k) \) are the biggest act as tracking color features [28]. Multi-cue fusion is applied to a visual tracking system. Supposing that \( P_k \) is probability distribution of a pixel (row, colu) at time \( t \) through feature \( k \). It denotes the probability that a pixel (row, colu) belongs to a target region under feature \( k \). \( P \) denotes final probability distribution of \( W \) features (W color features and a feature of target position prediction and a feature of motion continuity) fusion at time \( t \). It denotes the probability that a pixel (row, colu) belongs to a target region, as in (27). The normalized reliability of the chosen color feature \( k (k \in [1, W]) \) is \( V_k \), as in (28).

\[ P(\text{row, colu}, t) = \sum_{k=1}^{W} r_k * P_k(\text{row, colu}, t) \]

where \( r_k \) is weight value of feature \( k \). \( r_1, r_2, ..., r_w \) are weight values of chosen color features. \( r_{w+1} \) is weight value of feature of target position prediction. \( r_{w+2} \) is weight value of feature of motion continuity. \( \sum_{k=1}^{W} r_k = 1 \).

Normalizing reliability of the chosen color feature \( k \): \( V_k = \frac{V(L_i^k; p_k, q_k)}{\sum_{z=1}^{W} V(L_i^k; p_z, q_z)} \)

\[ \bar{V}_k = V_k \rightarrow r \]

\( \bar{V}_k = V_k - r_k \)  

\( k \in [1, W] \)  

\( \tau \) \( \tau \rightarrow r \)

\( \bar{V}_k \) denotes the reliability of color feature \( k \) of previous frame when weight value of color feature \( k \) need be changed, as in (29). \( \tau \) is the time constant that controls modifying rate of weight value, as in (30). Considering frame rate of every video is different, \( r_k \) is adjusted to \( \bar{V}_k \) during S frames in this paper. \( \bar{r} \) denotes adjusting range of color feature \( k \) at each time during \( S \) frames. Weight value \( r_k \) of color feature \( k \) is adjusted according to the return reliability. When \( \bar{V}_k \leftarrow r_k r_k \) increases. When \( \bar{V}_k \rightarrow r_k r_k \) decrease. 14 color features adjust
their weight value according to their reliability of the previous frame. When reliability of a feature is the highest it is dominant status of color features in a visual tracking system and it supplies more information to the tracking system. When its reliability decreases its information will decrease or be neglected, shown as Fig. 3.

The fusion of color feature and feature of target position prediction and feature of motion continuity is applied to this visual tracking system in this paper. $W$ color features are chosen every $S'$ frames according to their reliability. The chosen $W$ color features adjust their weight values during $S(S'>S)$ frames according to their reliability. Even if a cue doesn’t work this tracking algorithm still works well.

Realizing step of VTA algorithm for a target region:
step 1: Set an interesting area $I_1$ 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 $G_1$.
step 3: Set the most reliability $W$ color features and adjust their weight values.
step 4: Compute a probability distributed image $M_1$ of $W$ color features.
step 5: Compute a probability distributed image $M_2$ of feature of target position prediction.
step 6: Compute a probability distributed image $M_3$ of feature of motion continuity.
step 7: Set weight value $r_k$ of the three probability image $(M_1, M_2, M_3)$ and make a final probability distributed image $M$.
step 8: Compute zero moment and one moment of a window area in the probability distributed image. Iterate CAMSHIFT algorithm until the position coordinates don't evidently change or maximum times is reached. $T$ is maximum times.
step 9: Compute an interesting area $I_1$ 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 3.

Realizing step of VTA algorithm for an auxiliary object:
step 1: Set an interesting area $I_2$ 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 auxiliary target $G_2$.
step 3: Compute a probability distributed image $M_4$ of color features.
step 4: Compute a probability distributed image $M_5$ of feature of target position prediction.
step 5: Compute a probability distributed image $M_6$ of feature of motion continuity.
step 6: Set weight value $a_1, a_2, a_3$ of the three probability image $(M_4, M_5, M_6)$ and make a final probability distributed image $M, a_1+a_2+a_3=1$.
step 7: Compute zero moment and one moment of a window area in the probability distributed image. Iterate CAMSHIFT algorithm until the position coordinates don't evidently change or maximum times is reached. $T$ is maximum times.
step 8: Compute an interesting area $I_2$ according to coordinates of the tracked auxiliary target again and put them be an initial position and a tracking area of rear frames and return to step 3. When other auxiliary objects are chosen it is similar, shown as Fig. 4.

Locating of a target region and locating of an auxiliary object are cooperation and competition to determine a final target region for human motion in VTA algorithm. Supposing that coordinates of a target region is $(x_i, y_i)$ and coordinates of auxiliary object $n (n \geq 1)$ is $(x_{n+1}, y_{n+1})$. When a target region is close to an auxiliary object, namely $|x_j-x_1|<\alpha, |y_j-y_1|<\beta, ..., |x_j-x_{n+1}|<\alpha, |y_j-y_{n+1}|<\beta$, tracking results of the target region (coordinates and scope of the region) are final human target region. When a target region is far from an auxiliary object, namely $|x_j-x_2|\geq\alpha, |y_j-y_2|\geq\beta, ..., |x_j-x_{n+1}|\geq\alpha, |y_j-y_{n+1}|\geq\beta$, average values of coordinates of a target region and a auxiliary are coordinates of a final human target region. Scope of a final
IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

To test the results of the algorithm in this paper, VTA algorithm and CAMSHIFT algorithm and CMET algorithm are applied to video sequence 1-10, shown as table II. Resolution of sequence 1-3, sequence 5 and sequence 7-9 is 640*440. Resolution of sequence 4 and sequence 6 is 640*480. Resolution of sequence 10 is 320*256. Setting $T=15$, $n=1$, $\delta=0.00001$, visual attention control code of top-down is "01", $\alpha=60$, $\beta=0$, $m=255$, $S'=10$, $S=5$, a rectangle tracking window, $w=2$ when a target region is located, weight value of total color features is $r=3/7$, weight value of feature of target position prediction is $r_x=3/7$, weight value of feature of motion continuity is $r_z=1/7$. Only one auxiliary object is chosen in this paper. When auxiliary object 1 is located $a_x=3/7$, $a_z=3/7$, $a_x=1/7$. The running condition of this tracking algorithm is CPU (P4 2.8G), memory (512M), hard disk (80G), operating system (windows XP), tool (Matlab7.1).

<table>
<thead>
<tr>
<th>video sequences</th>
<th>People number</th>
<th>video sequences characteristic</th>
<th>frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>4</td>
<td>The target is occluded and cross-border. The target and background color are similar.</td>
<td>762 frames</td>
</tr>
<tr>
<td>S2</td>
<td>4</td>
<td>The target is occluded and cross-border. The target and other foreground color are similar.</td>
<td>512 frames</td>
</tr>
<tr>
<td>S3</td>
<td>4</td>
<td>The target is cross-border and isn’t occluded.</td>
<td>352 frames</td>
</tr>
<tr>
<td>S4</td>
<td>4</td>
<td>The target is occluded and cross-border. The target and background color are similar.</td>
<td>559 frames</td>
</tr>
<tr>
<td>S5</td>
<td>3</td>
<td>The target is occluded. The target and other foreground color are similar. Color saturation is low.</td>
<td>280 frames</td>
</tr>
<tr>
<td>S6</td>
<td>3</td>
<td>The target is occluded and cross-border. The target and other foreground color are similar. Color saturation is low.</td>
<td>263 frames</td>
</tr>
<tr>
<td>S7</td>
<td>3</td>
<td>The target is occluded and cross-border. The target and other foreground color are similar. Color saturation is low.</td>
<td>594 frames</td>
</tr>
<tr>
<td>S8</td>
<td>2</td>
<td>The target and background color are similar. The target is cross-border.</td>
<td>687 frames</td>
</tr>
<tr>
<td>S9</td>
<td>2</td>
<td>The target is occluded and cross-border. The target and background color are similar.</td>
<td>400 frames</td>
</tr>
<tr>
<td>S10</td>
<td>2</td>
<td>The target is occluded. The target and background color are similar.</td>
<td>272 frames</td>
</tr>
</tbody>
</table>
Fig. 8 denotes that for video sequence S2 the experimental results of VTA algorithm and CMET algorithm and CAMSHIFT algorithm. The target and the other human with similar color oppositely move. VTA algorithm considers the movement of the target and the auxiliary object and multiple color choice mechanism. It gets a good tracking result. CAMSHIFT algorithm doesn’t work well when the target enters the field of view of the camera again because cross-border time is too long. The result of CMET algorithm is better than that of CAMSHIFT algorithm and is worse than that of VTA algorithm. For video sequence S2 color features that VTA algorithm chooses are shown as Fig. 9.

Fig. 10, Fig. 11 and Fig. 12 denote that for video sequence S5 the experimental results of VTA algorithm and CMET algorithm and CAMSHIFT algorithm. When low saturation happens and color between the target and another human is similar the result of CAMSHIFT algorithm is the worst. The results of VTA algorithm is the best. For video sequence S5 color features that VTA algorithm chooses are shown as Fig. 13.
Fig. 14 denotes that for video sequence S6 experimental results of VTA algorithm and CAMSHIFT algorithm and CMET algorithm.

Fig. 15 denotes that for video sequence S8 the experimental results of VTA algorithm, CMET algorithm and CAMSHIFT algorithm. During tracking process the target uncovers red shirt. Therefore successful tracking rate of three algorithms decreases sharply. When the target uncovers the red shirt to some extent VTA algorithm can still track the target through the auxiliary object. Successful tracking rate of VTA algorithm is the highest. CAMSHIFT algorithm only considers color feature. The target is disturbed by the red cloth in the background. Its successful tracking rate is the lowest. Once the target uncovers red shirt CMET algorithm loses the target. For video sequence S8 color features that VTA algorithm chooses are shown as Fig. 16.

Table III denotes the comparison of tracking successful rate for 3 algorithms. Suppose that it will fail when a tracking window includes other human targets or 1/2 tracking window doesn't include the target. VTA algorithm works the best.

Fig. 17 is the results of video sequence S10. VTA algorithm works the best. CAMSHIFT algorithm works the worst. In video sequence S10 background and foreground color are similar. Successful rate of VTA algorithm is much higher than the other two algorithms. VTA algorithm uses multi-cue and visual attention mechanism. It fits the situation that colors of background and foreground are similar.
The successful key to a visual tracking system is that the tracking method is rapid, steady, precise and low cost. A novel visual tracking algorithm based on visual attention and multiple cues fusion for human motion analysis is proposed in this paper. Compatible auxiliary objects are chosen by visual attention mechanism. The robustness of the algorithm is increased through tracking a target and auxiliary objects. This algorithm combines color feature, feature of target position prediction and feature of motion continuity. Color feature considers hue, saturation, red channel, green channel and blue channel and so on. The finest information that can distinguish a target from background is chosen as color feature. Color feature and their weight value can adjust dynamically according to real environment. It can handle the situation that a target is occluded. The model of background need not be supposed. It need not train a video sequence without a moving target. It gets better tracking results. Next step is to consider other cues and increase the robustness of this algorithm.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Shenzhen Scientific and Technological Plan and Basic Research program (No.JC20090316369A), Natural Science Foundation of Guangdong(No.9151806001000025).

REFERENCES

[16] Huang-Chia Shih, Jenq-Neng Hwang, Chung-Lin Huang. ”Content-Based Attention Ranking Using Visual and Contextual Attention Model

<table>
<thead>
<tr>
<th>Video sequences</th>
<th>Algorithms</th>
<th>Successful tracked frames (total frames)</th>
<th>Successful rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>VTA algorithm</td>
<td>684(262)</td>
<td>89.8%</td>
</tr>
<tr>
<td>S1</td>
<td>CAMSHIFT algorithm</td>
<td>414(172)</td>
<td>54.3%</td>
</tr>
<tr>
<td>S1</td>
<td>CMET algorithm</td>
<td>639(262)</td>
<td>83.9%</td>
</tr>
<tr>
<td>S2</td>
<td>VTA algorithm</td>
<td>462(132)</td>
<td>90.2%</td>
</tr>
<tr>
<td>S2</td>
<td>CAMSHIFT algorithm</td>
<td>233(132)</td>
<td>45.5%</td>
</tr>
<tr>
<td>S2</td>
<td>CMET algorithm</td>
<td>448(132)</td>
<td>87.5%</td>
</tr>
<tr>
<td>S3</td>
<td>VTA algorithm</td>
<td>239(132)</td>
<td>67.9%</td>
</tr>
<tr>
<td>S3</td>
<td>CAMSHIFT algorithm</td>
<td>161(132)</td>
<td>45.7%</td>
</tr>
<tr>
<td>S3</td>
<td>CMET algorithm</td>
<td>166(132)</td>
<td>47.2%</td>
</tr>
<tr>
<td>S4</td>
<td>VTA algorithm</td>
<td>241(109)</td>
<td>43.0%</td>
</tr>
<tr>
<td>S4</td>
<td>CAMSHIFT algorithm</td>
<td>134(109)</td>
<td>24.0%</td>
</tr>
<tr>
<td>S4</td>
<td>CMET algorithm</td>
<td>225(109)</td>
<td>40.3%</td>
</tr>
<tr>
<td>S5</td>
<td>VTA algorithm</td>
<td>101(109)</td>
<td>36.1%</td>
</tr>
<tr>
<td>S5</td>
<td>CAMSHIFT algorithm</td>
<td>49(109)</td>
<td>17.5%</td>
</tr>
<tr>
<td>S5</td>
<td>CMET algorithm</td>
<td>91(109)</td>
<td>32.5%</td>
</tr>
<tr>
<td>S6</td>
<td>VTA algorithm</td>
<td>7(233)</td>
<td>28.1%</td>
</tr>
<tr>
<td>S6</td>
<td>CAMSHIFT algorithm</td>
<td>51(233)</td>
<td>19.4%</td>
</tr>
<tr>
<td>S6</td>
<td>CMET algorithm</td>
<td>58(233)</td>
<td>22.1%</td>
</tr>
<tr>
<td>S7</td>
<td>VTA algorithm</td>
<td>425(234)</td>
<td>71.5%</td>
</tr>
<tr>
<td>S7</td>
<td>CAMSHIFT algorithm</td>
<td>286(234)</td>
<td>48.1%</td>
</tr>
<tr>
<td>S7</td>
<td>CMET algorithm</td>
<td>416(234)</td>
<td>70.0%</td>
</tr>
<tr>
<td>S8</td>
<td>VTA algorithm</td>
<td>477(267)</td>
<td>69.4%</td>
</tr>
<tr>
<td>S8</td>
<td>CAMSHIFT algorithm</td>
<td>269(267)</td>
<td>39.2%</td>
</tr>
<tr>
<td>S8</td>
<td>CMET algorithm</td>
<td>353(267)</td>
<td>51.4%</td>
</tr>
<tr>
<td>S9</td>
<td>VTA algorithm</td>
<td>217(400)</td>
<td>54.3%</td>
</tr>
<tr>
<td>S9</td>
<td>CAMSHIFT algorithm</td>
<td>190(400)</td>
<td>49%</td>
</tr>
<tr>
<td>S9</td>
<td>CMET algorithm</td>
<td>207(400)</td>
<td>51.8%</td>
</tr>
<tr>
<td>S10</td>
<td>VTA algorithm</td>
<td>161(272)</td>
<td>59.2%</td>
</tr>
<tr>
<td>S10</td>
<td>CAMSHIFT algorithm</td>
<td>9(272)</td>
<td>3.3%</td>
</tr>
<tr>
<td>S10</td>
<td>CMET algorithm</td>
<td>21(272)</td>
<td>7.7%</td>
</tr>
</tbody>
</table>


