2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

A Salient Feature and Scene Semantics based Attention Model for Human Tracking on Mobile Robots

Hong Liu*, Huijun He

***Abstract*—It is a great challenge to perform robust tracking for a mobile robot owing to dynamic environments. Also, fast motion or abrupt jerk of the robotic camera poses a severe threat for continuous tracking. To address these problems, a novel attention model is proposed motivated by human attention mechanism which consists of low level salient feature and high level scene semantics. The low level layer extracts color and motion feature to obtain combined feature probability map. In semantic level, the *ADM(attention distribution map)* is computed by applying an attenuation function on the combined feature map which is motivated by human's foveal vision. The object position is found using CAMSHIFT algorithm in *ADM*. And this layer also generates a region-based *SSG(scene semantics graph)*. When robot moves abnormally, the model detects candidate regions in color saliency map and then attention shifts from one region to the next and check it by elastically matching *SSG* until the target is recovered. Experiments in several kinds of environments give promising results and show that this model is robust for mobile robotic tracking. When camera moves steadily, a little fast or even jerks very abruptly, it can keep continuous tracking.**

I. INTRODUCTION

Mobile robots have been a topic of intense research recently because of their various applications in home service, intelligent vehicle, military and so on. Owing to its non-contact character and friendly interface, camera is an important sensor in the interaction of mobile robot system and the external environments. Therefore robotic vision plays a crucial role in robot perception. In vision, human tracking is a key task for mobile robot.

There exit several difficulties for mobile robotic tracking. For example the environment is dynamic which causes static feature to fail easily. Especially, there is a noticeable problem: camera's abnormal motion. When the robot is moving, turning or stops suddenly, the camera mounted on the robot has an unpredictable motion way such as *fast motion* and *strong jerk*. In this paper *abnormal motion* mainly refers to them. The difference of these two problems is the intensity of camera's motion. For fast motion, it means that camera's motion is relatively fast, but not extremely strong. The main influence is motion blur in this case. For strong jerk, it means that camera jerk is so abruptly and strongly that the view field changes totally. The duration of jerk is usually very short (typical time is about 5~8 frames) and then camera returns at once. The main influence is the frame content changes dramatically during the time. These two problems are shown in Fig.1. The target's trajectory changes dramatically, which will have undesirable influence and affects continuous tracking severely.

a) Fast motion. Camera turns quickly and it causes motion blur. (Frame #317, #318)

b) Strong jerk. Camera is bumped strongly and recovers at once. The frame content changes dramatically. (Frame #40, #41, #43, #46)

Fig.1. Example frames of fast motion and strong jerk.

Up to now, researchers have proposed several methods to address fast motion problem. H. Jin et al. in[1] match regions by blurring images instead of deblurring them. The deformation parameters are estimated by minimizing a cost function. The work done by S. Dai et al. in [2] is concerned with local blurs. The method integrates signal processing and statistical learning. The estimated blurs are used to reduce the search range by providing strong motion predictions and to localize the best match accurately by modifying the measurement models. C. Mei and I. Reid in[3] extend the problem and addresses tracking in presence of spatially

Manuscript received October 9, 2009-9-15. This work was supported by National Natural Science Foundation of China (NSFC, No.60675025, 60875050) and National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247). Shenzhen Scientific and Technological Plan and Basic Research program (No.JC200903160369A), Natural Science Foundation of Guangdong (No.9151806001000025).

Hong Liu is with the Key Laboratory of Machine Perception and Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, CO 518055 CHINA. (e-mail: hongliu @ pku.edu.cn).

Huijun He is with the Key Laboratory of Machine Perception and Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, CO 518055 CHINA. (e-mail: hehj @ cis.pku.edu.cn).

variant motion blur generated by a planar template. A. Pretto in [4] discusses both the feature extraction and tracking approaches under motion blur on humanoid robots. It estimates the point spread function (PSF) of the motion blur individually for image patches obtained via a clustering technique. However, the obvious disadvantage of these works is that they are only concerned with the motion blur caused by camera's fast but *regular* motion. They don't discuss the *abrupt and strong jerk* problem which is also common for robotic tracking. For example, when the robot is an intelligent vehicle running on a tough road, the camera will jerk so abruptly that the tracked target is usually out of the view field. This problem influences continuous tracking severely.

Several visual attention models have been proposed to explain and simulate human visual attention. There exist two ways by which information can be used to direct attention in literatures (see [5, 6] for reviews). One approach uses bottom-up[7] information including basic features such as color, orientation, motion, depth, conjunctions of features such as objects in 2D or 3D space and even learned features. A great number of models make use of "saliency"[8] to direct attention. However, saliency cannot always capture attention in a purely bottom-up fashion if attention is focused or directed elsewhere in advance. Thus it is necessary to recognize the importance of how attention is also controlled by top-down[9] information relevant to current visual behaviors.

From these researches, it can be found that human's visual system not only extracts low level feature but also forms scene semantic knowledge. Motivated by this mechanism, a novel attention model which consists of low level salient feature and high level scene semantics is proposed for robotic tracking in this paper. The features extracted in lower layer are the primal stage and basic stimuli of attention model. Meanwhile, the higher semantics level decides the extraction and shift of low level features. In semantics level, *ADM(attention distribution map)* which denotes the object position and *SSG(scene semantics graph)* which represents the scene semantic knowledge are obtained. When camera moves steadily, object position is found in *ADM*. When camera moves abnormally, although the low level features are distorted, the semantics doesn't changes nearly. Semantics information will help recover the target position and the model keeps continuous tracking.

Compared with the aforementioned attention models, our model emphasizes higher semantic information more and therefore is more robust when extreme condition happens. The main contributions of this paper are summarized as follows:

1) Visual attention mechanism is introduced to object tracking and one general attention model is proposed for robotic tracking when moving steadily or abnormally.

2) This model combines not only basic image features such as color, motion but also contains semantic spatial information described by *SSG*.

3) This model can recover the lost target position by elastic graph matching when camera jerks strongly, which is the first one, to our best knowledge.

The remainder of this paper is organized as follows. Section 2 gives a brief introduction of the general structure. In section 3, 4 the details of the two levels of attention model are described, respectively. Section 5 discusses how to handle abnormal motion with this model. And then section 6 shows the experimental results and comparisons with the other methods. At last conclusions are made in section 7.

II. GENERAL STRUCTURE

The overall architecture of our proposed attention model which consists of low level salient feature and high level scene semantics is illustrated in Fig.2.

For moving robot, the environment illumination will vary along with the camera position, and this will influence the target appearance. The color channel with the biggest discriminating measurement is adaptively selected from a predefined feature set. And color probability map $P(x,y)$ is computed under this channel by back projection. KLT interest points are detected and the points' correspondence relationship between consecutive frames is also obtained. It can be supposed that the camera's ego-motion is affine transformation which is determined by four parameters. Therefore, RANSAC algorithm is used to estimate the transformation matrix and then the background of last frame is registered with current frame. Frame difference is implemented and motion regions $R(x,y)$ are detected. Moreover, the motional objects O_0, O_1, O_2,…,O_n are segmented and they are used to generate scene semantic graph.

Afterwards, in semantics level, an exponential spatial attenuation function $f(x,y)$ which is motivated by the foveal vision is applied on $P'(x, y)$ to produce *ADM(attention distribution map)*. This mechanism can focus attention on the tracked object and suppress the influence of distracters remarkably. And then mode seeking algorithm such as CAMSHIFT[10] is applied to find the object position in *ADM*. What's more, this level infers scene spatial relation and generates *SSG(scene semantics graph)* denoted by $G(v, e)$.

One attention evaluation coefficient is set up to decide whether abnormal motion happens. If the coefficient is smaller than one threshold, attention will skim the whole image to detect salient regions according to color similarity. After this, attention shifts from one region to the next and checks whether it is the target by matching spatial scene with the generated *SSG*. Once the target position is recovered, it re-initializes the model and tracking continues.

III. LOW LEVEL SALIENT FEATURE

A. Color Feature

According to the studies in psychology and cognitive neuroscience, selective visual attention[11] acts like a filter to select discriminative information from the massive information in the field of view.

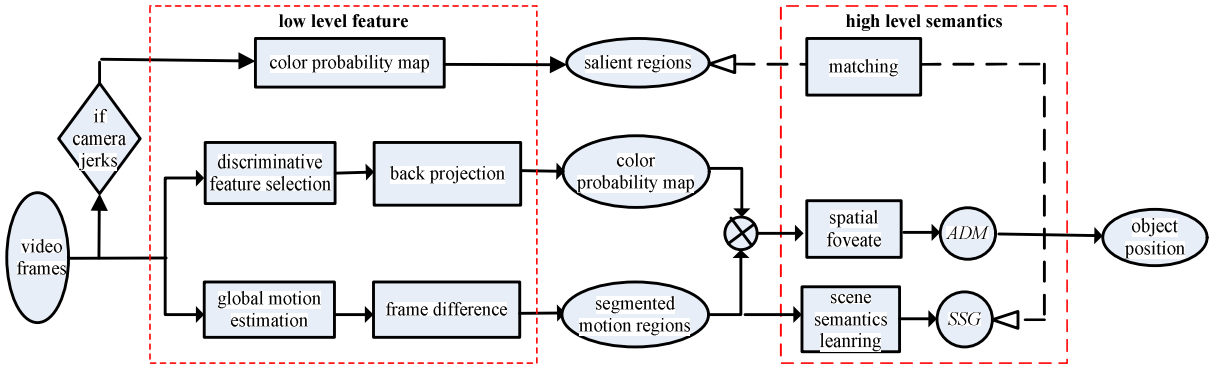

Fig.2. The general framework of our proposed attention model.

The color feature pool is predefined as nine channels {*R, G, B, H, S, V, r, g, b*}, which is the decomposition of the three color spaces: *RGB, HSV, rgb*. For each frame, one color channel is selected according to an ad hoc criterion from the feature pool. When the RGB space is converted to HSV space, if pixels' color has a low saturation near to zero, RGB channels will have similar values and the hue channel is not well defined or inaccurate. Therefore two thresholds are set and pixels with too low saturation and too brightness are discarded.

The normalized *rgb* color model consists of the *r*, *g* and *b* channels[12]. They are obtained by dividing the *R, G* and *B* values by their total sum. The formal definition is written as follows,

$$r=\frac{R}{R+G+B}\quad g=\frac{G}{R+G+B}\quad b=\frac{B}{R+G+B} \tag{1}$$

where $r+g+b=1$. The normalized *rgb* color model defined above possesses photometric invariant features, which is insensitive to surface orientation, illumination direction and intensity.

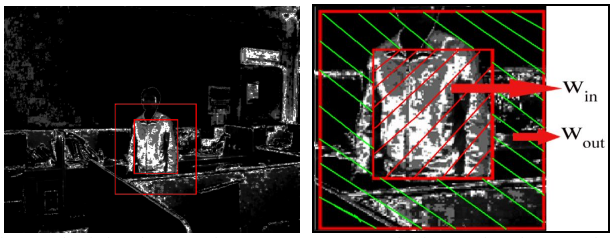

Fig.3. Principle of computing measurement r. The internal and external regions are denoted by the two red rectangles

Under different illuminative conditions, the discriminating ability of various color channels differs. As to how to measure the quality of the feature, the measurement in [13] is adopted. We draw two rectangles as showed by Fig.3. The internal one denotes target region and external one denotes background region. The measurement is written as follows.

$$r=\frac{\sum_{(i,j)\in W_{in}} p(i,j)^2}{\sum_{(i,j)\in W_{out}} p(i,j)}\bullet\frac{1}{|W_{in}|} \tag{2}$$

The nine channel histograms of target template are stored at first. In each frame the best color feature is selected, and then color probability map $P(x, y)$ is computed through back projection.

B. Motion Feature

Usually frame differencing is a commonly used technique to get the motional regions. However, when the camera moves (eg. when it is mounted on a mobile robot), frame differencing is infeasible because an obvious change is generated by the moving camera. The idea of our model to detect the motion regions is that ego-motion of the camera can be estimated by tracking features between images. The motion between two consecutive frames can be assumed to be affine transformation which means that the motion is decomposed to scale, rotation and translation. Let (x, y) and (x', y') denote the pixel coordinates in frame I_{t-1} and I_t, respectively. The relationship of the coordinates in two frames can be described as follows:

$$\begin{pmatrix}x'\\y'\end{pmatrix}=s\bullet\begin{pmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{pmatrix}\bullet\begin{pmatrix}x\\y\end{pmatrix}+\begin{pmatrix}dx\\dy\end{pmatrix} \tag{3}$$

where s denotes scaling factor, α is the rotation angle and (dx, dy) is the translation displacement. Formula (3) can also be written as follows,

$$\begin{bmatrix}x & -y & 1 & 0\\ y & x & 0 & 1\end{bmatrix}\begin{bmatrix}a\\b\\dx\\dy\end{bmatrix}=\begin{bmatrix}x'\\y'\end{bmatrix} \tag{4}$$

As to the interest point, the KLT detection method is adopted[14], which can be implemented efficiently using multi-solution pyramid scheme.

From formula(4) it can be found that if two precise pairs of points are known, the four parameters which describe the motion between consecutive two frames are totally determined. In practice there is noise, and the points tracked wrongly or from foreground which is so-called outliers, thus RANSAC (random sample consensus) algorithm[15] is applied to estimate the motion parameter. RANSAC is a robust model parameter estimation algorithm which can even work when the portion of outliers is close to 50%. Suppose $map(I_{t-1})$ is the transformation of frame I_{t-1}, the frame differencing is computed as,

$$d(I_t, I_{t-1}) = | I_t - map(I_{t-1}) | \quad (5)$$

After frame differencing there is another problem which will be encountered. It is the so-called foreground aperture[16] shown as Fig.4, which means the result is always the boundary but not the whole motion region. A horizontal and vertical projection based method is proposed to overcome this problem. The difference image is projected to horizontal direction to find the left and right endpoints, and then project to vertical directions to find the upper and low endpoints. In this way, the rectangle denoting the motion region can be determined. This method is illustrated in Fig.4. The result of motion feature is a binary function *R(x,y),*

$$R(x,y) = \begin{cases} 1, & (x,y) \in motion\ region \\ 0, & others \end{cases} \quad (6)$$

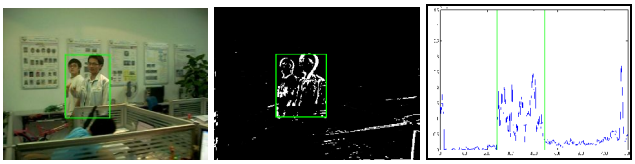

Fig4. Method of detecting motion attention region.

The color probability map is only considered in the detected motion regions and therefore irrelevant information can be reduced remarkably. The combined feature probability map *P'(x, y)* is

$$P'(x,y) = P(x,y) \bullet R(x,y) \quad (7)$$

IV. High Level Scene Semantics

After low level features are extracted, high level semantics calculates *ADM* through the spatial attenuation function motivated by foeavte vision and generates *SSG* by using the information of the other detected moving objects.

A. Foveal vision

The retina of human's vision system has a non-uniform structure[17]. The retina's resolution is high in the central area and this resolution continuously drops as you go into the peripheral area. In vision task, the fovea corresponds to the focusing point where the highest density of attention is distributed and the farther it is from the center, the less attention there is. The relative acuity of human eye is illustrated in Fig.5.[18]

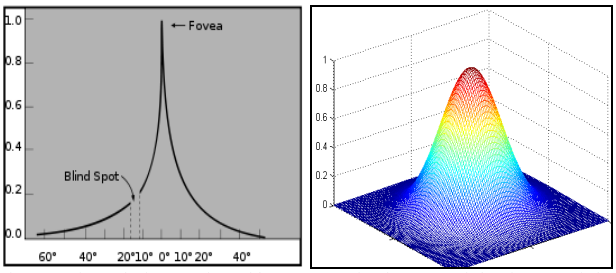

Fig.5. Left: Relative acuity of human eye in degrees of the fovea. Right: the attention distribution map with attenuation function.

During tracking, because eyes are gazing the tracked object, the object position is the fixation point which attracts the densest attention and that of the region surrounding it is less and less. Motivated by this idea, when computing the attention distribution, an exponential attenuation function which decreases with respect to the distance to the object position is introduced. It can focus attention on the tracked region and suppress the influence of distracters passing-by. The analytic form of the function is written as follows,

$$f(x,y) = \exp(-\frac{(x-x_0)^2 + (y-y_0)^2}{c \bullet w \bullet h}) \quad (8)$$

where *(x, y)* is one pixel coordinate in the frame, (x_0, y_0) is the fixation point which is the tracked result in last frame. Note that the smaller the tracked region is, the more extent the attenuation is. Therefore *w* and *h* which mean the width and height of the object rectangle region respectively and constant *c* controls the attenuation degree.

Therefore the *ADM(Attention Distribution Map) A(x,y)* is computed as follows

$$A(x,y) = P'(x,y) \bullet f(x,y) \quad (9)$$

B. SSG Generation

The segmented objects denoted by $O_0, O_1, \ldots, O_n$ in motion feature extraction are used to generate *SSG*. This graph contains *RGB* color histogram, size and relative position of each object which represents not only appearance information but also spatial semantics.

Of the objects $O_0, O_1, \ldots, O_n$, O_0 denotes the tracked target and $O_1, O_2, \ldots, O_n$ denote the other moving objects. Denote the graph as *G=(V, E)*, *V* is vertex set and *E* is edge set. Each vertex v_i represents the object O_i and it records its information *{hist, p, s}*, in which *hist* means the *RGB* histogram of O_i, *p* means the object's position, i.e. the horizontal and vertical coordinates in current frame and *s* is the object size i.e. the width and height of object. Each edge e_i is a vector representing the relative position from the target O_0 to object O_i. Fig.6 shows two examples of generated *SSG* in two scenarios. In this graph, the red eclipse denotes the tracked target while the white rectangles denote the moving objects detected in motion feature extraction. Two edges connect the target with the other objects respectively.

Fig.6.The generated SSG in two videos.

When the robot moves normally, the attention model generates *ADM* and *SSG*. Object position is found in *ADM* using CAMSHIFT mode seeking algorithm. *SSG* records the appearance information and scene spatial semantics in the

current frame. If extreme condition happens such as fast motion and strong jerk which may cause tracking drift away, *SSG* can help recover target position by elastically matching scene semantics.

V. CAMERA'S ABNORMAL MOTION HANDLING

A. Abnormal Motion Detection

One attention evaluation coefficient is set up to decide whether attention should focus on the target. If it is smaller than the threshold, it means that abnormal motion happens and the target position changes dramatically.

The Bhattacharyya distance ρ is used to measure the similarity of objects represented by two histograms. Suppose that the histogram of tracked region is p, and the one of target template is q, the bins of the histograms are p_u, q_u. The distance of p and q is written as follows,

$$\rho = \sum_{u=1}^{m} \sqrt{p_u q_u} \tag{10}$$

If this score ρ is big enough, it means that the template matches well and tracking continues. Otherwise, it means camera abnormal motion happens and attention model will make use of the generated scene semantics to recover target position.

B. SSG Elastic Matching and Target Position Recovering

When abnormal motion happens, the content of frame changes dramatically and attention will not focus on the target but skim the whole frame to detect salient regions. Because the object may move a little, when checking the detected region, an elastic similarity function $S(G_0, G_1)$ considering object appearance, size and scene semantics is proposed to decide which region is the target.

The 3D *RGB* histograms of each object in the last frame before abnormal motion are calculated and stored in *SSG* which is denoted by G_0. When abnormal motion is detected target color saliency map $C(x,y)$ of current frame is calculated by back projecting to $hist_0$ stored in G_0. In $C(x,y)$, the gray value of each pixel means its probability of belonging to target appearance. Salient regions are detected in $C(x,y)$ through projection-based region segmentation method. These regions are stored in a list and then attention model will shift and check from one region to the next.

When checking one salient region R_k, the model first computes the predicted *SSG* denoted by G_1 of this region. The relative position of each segmented object $O_1, O_2, \ldots, O_n$ represented by edges of G_0 determine predicted search windows. The n edges determine n predicted windows. Use them as initial search windows of CAMSHIFT and denote the resulting rectangles as $Rect_1$, $Rect_2$, .., $Rect_n$. Each resulting rectangle $Rect_i$ is the vertex v_i of G_1 and the vector from the checked region R_k to $Rect_i$ is the edge e_i.

Suppose that the sizes, the *RGB* color histograms and the centers of resulting region of vertex i of the G_0 and G_1 are $s_i, \hat{s}_i, h_i, \hat{h}_i, P_i(x,y), \hat{P}_i(x,y)$. Suppose the widths and heights of the two sizes are $\hat{w}_i$, $\hat{h}_i$, w_i, h_i. The distance of the two sizes c_i is defined as follows,

$$c_i = \left|\frac{\hat{w}_i}{w_i} - 1\right| + \left|\frac{\hat{h}_i}{h_i} - 1\right| \tag{11}$$

The similarity of two histograms ρ_i is the Bhattacharyya distance as formula (11). The distance of two centers d_i is Euclidean distance of the center coordinates $(\hat{x}_i, \hat{y}_i), (x_i, y_i)$

$$\begin{aligned} d_i &= \| \hat{p}_i(x,y) - p_i(x,y) \| \\ &= \sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2} \end{aligned} \tag{12}$$

The elastic similarity function $S(G_0, G_1)$ of the two graphs is defined as the ration of appearance similarity to the sum of size and center distance. The formal definition is expressed as follows,

$$\begin{aligned} S(G_0, G_1) &= \frac{\sum_{i=1}^{n} \rho_i}{\alpha \sum_{i=1}^{n} c_i + \beta \sum_{i=1}^{n} d_i} \\ &= \frac{\sum_{i=1}^{n} \sum_{u=1}^{m} \sqrt{p_{iu} q_{iu}}}{\alpha \sum_{i=1}^{n} \left(\left|\frac{\hat{w}_i}{w_i} - 1\right| + \left|\frac{\hat{h}_i}{h_i} - 1\right|\right) + \beta \sum_{i=1}^{n} \sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2}} \end{aligned} \tag{13}$$

In this expression, ρ_i, c_i, and d_i account for the object appearance change, size variation and scene structure deformation, α and β are the weights of size and position distance. The overall value measures the similarity of not only the local region but also global scene. The region with the biggest similarity is selected to re-initialize CAMSHIFT and tracking continues.

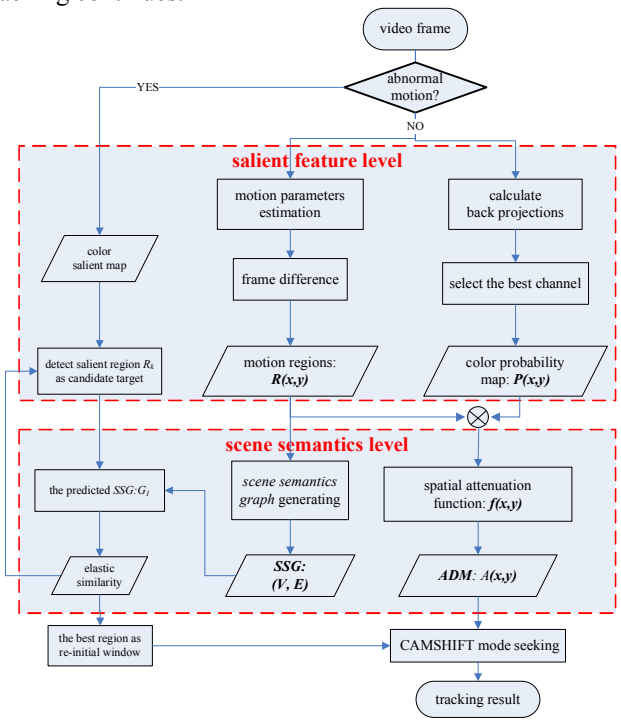

Fig.7. The flow chart of our proposed two level attention model.

In summary, Fig.7 shows the flow chart of the attention model based tracking with abnormal motion.

VI. EXPERIMENTS

A. Experiment Settings and General Performance

To evaluate the effectiveness of the proposed model, an experimental system is set up. The PC platform is 1.6 GHz AMD Turion Dual CPU and 1GB memory. And the mobile robot for capturing video called "Pengpeng Ⅱ" is a HRI (Human Robot Interaction) oriented mobile robot shown in fig.8. The testing sequences are captured by this robot or directly by a hand held camera. The capturing rate is 20 fps and the image size is 640*480. The testing video database includes 11 sequences which have over 10800 frames and they include several kinds of environments and camera's motion on mobile robot, for example steady moving, fast motion, strong jerk. And meantime, in the video there are distracters with similar color clothes passing by.

Nine color channels *{R, G, B, H, S, V, r, g, b}* and a 1D histogram with 60 bins are used. The width and height of external rectangles are two times of those of internal ones. In each frame 100 KLT interest points are detected. In semantics level, the constant c of attenuation function is set to be 0.6~0.8 in different videos. The weights α and β is set to be 0.2 and 0.015, respectively. The comparing algorithms are mean-shift in[19], its variant CAMSHIFT in[12] and the improved mean-shift integrating motion blur handler proposed in [2]. When the resulting ellipse which denotes tracked region converges to a wrong position or is 1.5 times larger than right size, it is defined as tracking failure. The rates are computed at frame level. The performance means successfully tracking rate which is the number of successful frames dividing by the total numbers

Fig.8. The "Pengpeng Ⅱ" HRI oriented mobile robot and the camera mounted on it.

The general performance and comparisons are listed in table Ⅰ. When there is no abnormal motion, the simple mean-shift performs not badly. When the camera moves fast or even jerks, performance is very vulnerable to strong jerk and degrades severely. Because it assumes that target moves smoothly and doesn't have mechanism to recover. Therefore once jerk happens, the target position is lost and the results afterwards are random. Although the improved mean-shift is enhanced obviously, it is still influenced by strong jerk dramatically, because it doesn't consider this situation. Our model can perform quite well both in normal and abnormal cases in that it contains scene semantics information and recovers the position when the target is lost caused by abnormal motion.

B. Tracking under normal motion

The tested videos are S1-S3 which includes 2900 frames captured in indoor or outdoor environments. Of them S3 is explained thoroughly.

There are four persons walking on the platform, the tracked target is the girl. The other three are distracters with similar coats. Fig.9 shows the results of our model and the comparing mean shift algorithm. Because it is in outdoor environments, the illumination changes with the camera angle. Moreover, the other three distracters' similar color also influences tracking. Our model can adaptively select the most salient color channel in each frame and delete the impact of static background by motion feature. Therefore, it can keep robust tracking. By contrast, the method in[2] which only uses color information performs quite badly because of the influence of similar color distracters and background. The results drift easily to the wrong place after a few frames and never recover.

C. Tracking under fast motion

The attention model for handling fast motion is tested using four videos S5-S8 which include 2800 frames and 34 times of fast motion. The fast motion is produced by the camera's quick left and right turns or sudden translational motion.

Fig.10 shows the tracking results of S6. In this sequence there are four persons who wear similar blue t-shirts. This video is captured in outdoor environments. The tracked target is the girl with blue stripes and the distracters are the three boys who are walking randomly behind. Red ellipses denote tracked regions and green dotted rectangles are detected salient regions when there are fast motions.

Table Ⅰ. The general performance and comparison with the other two methods.

Method	Abnormal motion type			Capturing environment		Number of objects		Average	Speed (fps)
	No	Fast motion	Strong jerk	Indoor	Outdoor	3	4		
Our proposed model	92%	89%	86%	92%	88%	91%	84%	89%	14
Improved mean shift	89%	82%	68%	83%	76%	80%	72%	74%	18
Mean shift	84%	59%	37%	61%	42%	64%	48%	55%	23

It can be seen that in frame 715, when camera turns left and right quickly the trajectory of the target is not smooth and changes suddenly. Because the camera's sudden motion is not linear and uniform speed strictly which is the basic assumption of parameter based method. Moreover, this kind of method only uses the low level feature of object. Consequently, method 2 can't handle this case and the results converge to the distracter. The results of mean shift are even worse because it doesn't have any mechanism to cope with this situation. On the contrary, with our model, when sudden motion happens, it detects salient regions at first and then recovers the target's position using semantics. It can keep robust tracking when camera moves quickly.

D. Tracking under strong jerk

The attention model for handling strong jerk is tested using four videos S8-S11 which include 5200 frames. In these videos, S8-S10 include strong jerk only and S11 includes jerk and fast motion both. The time of jerks in all videos is 61.

Fig.11 shows the comparisons of our proposed model with the simple mean shift and improved one with motion blur handler. In this sequence there are three persons with red t-shirts and one person with dark blue. The tracked target is the boy with red t-shirt. It is captured in outdoor environments. The crowd walk along the hallway and the camera jerks for many times. In the result images green dot rectangles denote detected *SRs* when jerk is detected while white ellipses mean tracking results. It can be seen from the comparison that our model can keep robust tracking even when the jerk is very abrupt and frequent because it detect salient regions when jerk happens and then check by matching generated *SSG* in high level. On the contrary, the algorithm without jerk handler fails totally even when the camera jerks for only one time. Afterwards, it proceeds completely in wrong position. And the improved mean-shift although is relatively better than simple version, it tracks the wrong target with similar color because it only uses low level features.

VII. CONCLUSION

This paper proposes a novel visual attention model for tracking on mobile robot which consists of low level salient feature and high level scene semantics. The low level extracts salient color and motion feature to construct combined feature probability map, and then in semantics level, *ADM* which denotes the distribution of attention and *SSG* which expresses the scene semantic are obtained. When robot moves steadily, the tracked position is found in *ADM*. When camera's abnormal motion happens salient regions are detected in color saliency map. And then attention shifts on each region and elastically matches *SSG* to check until the target position is recovered. Demonstrated by extensive experiments, compared with the simple and improved mean-shift this model performs quite well in some challenging situations for example camera moves suddenly or jerks strongly, it can also keep robust tracking.

ACKNOWLEDGEMENT

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.JC200903160369A), Natural Science Foundation of Guangdong(No.9151806001000025).

REFERENCES

[1] H. Jin, P. Favaro , R. Cipolla, "Visual Tracking in the Presence of Motion Blur", Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp:18-25, Volume 2, 2005

[2] S. Dai, M. Yang, Y. Wu, A K. Katsaggelos, "Tracking Motion-Blurred Targets in Video", IEEE Intl. Conf. on Image Processing , pp:2389-2392, 2006.

[3] C. Mei and I. D. Reid, "Modeling and Generating Complex Motion Blur for Real-time Tracking", IEEE Intl. Conf. on Computer Vision and Pattern Recognition. pp:1-8, Volume 2, 2008

[4] A. Pretto, E. Menegatti, M. Bennewitz, W. Burgard, and E. Pagello, "A Visual Odometry Framework Robust to Motion Blur", proc IEEE Intl. Conf. on Robotics and Automation, 2009.

[5] Timothy J Buschman, Earl K Miller, "Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices", Science, Vol. 315, No. 5820, pp. 1860-1862, 2007.

[6] CE Connor, HE Egeth, S Yantis, "Visual attention: bottom-up versus top-down", Curr Biol, Vol. 14, No. 19, 2004

[7] L. Itti, "Models of Bottom-Up Attention and Saliency", Neurobiology of Attention, pp. 576-582, 2005.

[8] L. Itti, C. Koch, "Feature Combination Strategies for Saliency-Based Visual Attention Systems", [J].Electronic Imaging, Vol. 10, No. 1, pp. 161-169,2001

[9] Oliva, A., Torralba, A., Castelhano, M.S., Henderson, J.M, "Top-down control of visual attention in object detection", IEEE Intl. Conf. on Image Processing ,vol.1: 253-256, 2003

[10] G. R. Bradski, "Computer Vision Face Tracking for Use in a Perceptual User Interface," Intel Technology Journal, 2nd Quarter, pp. 1-20, 1998

[11] Wolfe, J. M, "Guided search 2.0: a revised model of visual search." Psychonomic Bulletin Review 1: 202-238, 1994

[12] CHEN Bai-sheng, CHEN Duan-sheng. "Normalized rgb color model based shadow detection", [J].Computer Applications, 26(8): 1879-1881, 2006.

[13] H. Stern and B. Efros, "Adaptive Color Space Switching for Face Tracking in Multi-Colored Lighting Environments," The Fifth IEEE Intl. Conf. on Automatic Face and Gesture Recognition, pp. 20-21, 2002.

[14] C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.

[15] M. A. Fischler, R. C. Bolles. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", Comm. of the ACM, Vol 24, pp 381-395, 1981.

[16] Shuguang Zhao, et al. "Moving Object Detecting Using Gradient Information, Three-Frame-Differencing and Connectivity Testing". AI, LNAI 4304, pp. 510 – 518, 2006

[17] N. Oshiro and A. Nishikawa, N. Maru, F. Miyazaki, "Foveated vision for scene exploration," Proc. Asian Conf. on Computer Vision, pp. 256-263, vol. 1351, 1998.

[18] Hans-Werner Hunziker, "Im Auge des Lesers: foveale und periphere Wahrnehmung - vom Buchstabieren zur Lesefreude" Transmedia Stäubli Verlag Zürich, 2006.

[19] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.564-577, Vol. 25, 2003.

Fig. 9. The comparison with mean-shift algorithm in[20] when robot moves steadily in out door environment. (frame#74, #82, #98, #153, #179)
Top: our model. Down: mean shift algorithm

Fig.10. The comparison of tracking results when camera has fast motion and it cause motion blur. (Frame: # 715, #720, #722, #725, #726)
1st row: our model. 2nd: mean-shift 3rd: improved mean-shift for motion blur.

Fig.11. The comparison of tracking results when camera has strong jerks and the frame content changes dramatically.(frame: # 118, #120, #122, #125, #127)
1st row: our model. 2nd: mean-shift 3rd: improved mean-shift for motion blur.