The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Collaboration of Spatial and Feature Attention for Visual Tracking

Hong Liu, Weiwei Wan and Ying Shi

Abstract—**Although primates can facilely maintain long-duration tracking of an object without infection of occlusion or other near similar distracters, it remains a challenge for computer vision system. Studies in psychology suggest that the ability of primates to focus selective attention on the spatial properties of an object is necessary to observe object quickly and efficiently while focus selective attention on the feature properties of object is necessary to render it more prominent from the distracters. In this paper, we propose a novel spatial-feature attentional visual tracking (SFAVT) algorithm to encode both. In SFAVT, tracking is treated as an on-line binary classification problem where spatial attention is employed in early selective procedure to construct foreground/background appearance model by identifying image patches with good localization properties, and in late selective procedure to update models by maintaining image patches with good discrimitive motion properties. Meanwhile, feature attention works in mode seeking procedure to help select feature spaces that best separate a target from background. The on-line tuned adaptive appearance models by those selected feature spaces are used to train a classifier for target localization, then. Experiments under various real-world conditions show that this algorithm is able to track an object influenced by dramatic distracters while is of comparable time efficiency with meanshift.**

I. Introduction

Visual tracking is crucial to machine vision applications such as surveillance, driver assistance, autonomous robot system and many others that require video analysis. However, continuous tracking of an object in cluttered environment is still a challenge, because it may suffer from unpredictable target visual appearance (e.g. complex object shape and motion, non-rigid or architected nature of object, partial or full occlusion.) and unconstrained environment (e.g. varying scene illumination condition, unpredictable various background and distracters).

A lot of outstanding visual tracking algorithms have been generated by the research of computer vision over the past decades. And most of the state-of-the-art models of object tracking mainly devote their efforts to two aspects [1]: Object representation and mode seeking. Object representation describes the basic criteria for mode seeking and hence helps to locate the candidate target in consequent video frames.

In the beginning of research on tracking, because of limited computational ability, the object model is simple and constant, and mode seeking algorithm is relatively efficient. For simple points or tiny regions, object representation [2] is a set of points or the centroid of them, and target localization algorithm based on it is also very simple. When it comes to some bigger rigid objects, point representation is no longer proper for tracking. Primitive sliding window [3] does well by describing object motion in way of translation, affine or projective transformation, which increases computational complexity of later seeking algorithm but not much. However, with the increasing demands for complex video analysis (e.g. tracking those objects with irregular shapes or non-rigid nature), primitive window method inevitably adds some background information to object representation and then the seeking algorithm may be influenced and the result will degrade. In order to solve this problem, many high-level representations have been proposed, such as spatiogram model to encode spatial information [4], skeletal and articulated shape model to capture articulation of the object [5], generative model for on-line adaptation [6,7], and so on. Also, some kernel methods have been employed to reduce background information by computing a weighted feature density histogram [8]. However, the mode seeking algorithms based on those solutions are of high computation complexity for their high correlation with high-level recognition procedure.

It seems that there is a dilemma between effective tracking and efficient computation in traditional models of visual tracking. Low computation complexity of mode seeking algorithm should be based on simple object representation which, however, could not fully capture the unpredictable object appearance. High-level object representation could handle those variances very well; yet followed by complex localization algorithm. On the contrary, as we know, it's only a basic ability for the primates to easily maintain a long-duration tracking of object even in cluttered environment. According to the study in psychology and cognitive neuroscience, selective visual attention is nature's answer to computational dilemma [9] that it acts like a filter to select active information out of the deluge information of the image.

Hong Liu is with the Key Lab of Machine Perception and Intelligence and the Key Lab of Integrated Micro-System, Shenzhen Graduate School, Peking University, China. hongliu@pku.edu.cn

Weiwei Wan and Ying Shi is with the Key Lab of Machine Perception and Intelligence, Peking University, China. wanweiwei@cis.pku.edu.cn

It's interesting and meaningful to apply these studies to artificial visual tracking and some novel visual tracking algorithms have been proposed based on them.

Researches show that spatial-based attention shifts across salient image regions and helps observe the target efficiently. One attentional visual tracking algorithm proposed by M. Yang [10] successfully employed spatial selective attention in both early selective procedure to extract a pool of attentional regions for object representation and late selective procedure to identify a subset of discriminative attentional regions for model updating (e.g. when appearance changes). Indeed, spatial attention in this algorithm helps to focus its computational resources to more informative regions. However, the seeking algorithm based on such object models like the locality-sensitive hashing technique overlooks the contribution due to knowing the distracting background.

It is implied that feature-based attention shifts in feature spaces by selectively enhancing features that could render target more salient from the distracters and helps to speed up the efficient detection of a target in cluttered environment [11]. Y. Liu [12] proposed a feature selecting method for adjusting the set of features used in mode seeking algorithm to improve tracking performance. However, a bunch of observation for best object representation should be displayed in advance.

Considering these factors, it could be benefitable to combine these two attentions for visual tracking in both object modeling/updating procedure and mode seeking procedure. In this paper, we propose a novel spatial-feature attentional visual tracking (SFAVT) algorithm that adaptively connects both. Specifically, tracking is treated as an online binary classification problem where spatial attention is introduced in early selective procedure to construct target/background appearance model by using image patches with good localization properties and in late selective procedure to efficiently update these models by using image patches with discrimitive motion properties. Meanwhile, feature attention works in a mode seeking procedure to help select feature spaces that best separate targets from background. And tuned by these selected feature spaces on line, the models become adaptive. Then, they are used to train a classifier for target localization and hence speed up the mode seeking procedure.

The proposed SFAVT algorithm can substantially handle the dilemma between effective tracking and efficient computation. To be exact, the collaboration of spatial and feature attention could help to extract selective data carrying enough active information in both the modeling procedure and the mode seeking procedure. Consequently, the effciency and robustness of the tracker can be promised: As the size of these selective data is small, fewer computational resources are required for processing them.

The remainder of this paper is organized as following. In section II, we discuss previous works on online binary classification based tracking and address the advantages of our work by comparison. The SFAVT algorithm is then described in section III. And section IV gives extensive experiments under various real-world conditions. Finally we come to conclusions of the paper and discuss several open issues and possible extensions in section V.

II. Related Work

In this paper, we address tracking as a binary classification problem [13, 14] which employs the collaboration of spatial attention [10] and feature attention to acquire both effective tracking and efficient computation.

Ensemble tracking [14] maintains the binary foreground and background appearance model through an ensemble of simple weak classifiers trained online from a specific frame. The updating training of them is in a predefined range of recent frames but according to the real situation of environments. In this way, the tracker will fail when an extended occlusion happens and should be recovered with the help of particle filtering or other temporal filtering methods. Non-parametric tracking [13] could handle this problem very well by constructing temporal appearance model directly on fine-grained data samples, ie, pools of simple color-texture features of image patches. However, sampling rate should be predefined to reduce the number of foreground/background training set for the following classifier, or the computational complexity will be high. Although it is announced that final tracking performance will not be influenced by low sampling rate, it is reasonable to doubt that a lot of useful information for separating target from background will be lost in such way. And efficient model updating should be maintained by a bidirectional consistency check which is of high computational complexity. What's more, when there is a large group of similar distracters in the background, the tracker will unsurprisingly be confused in the learning procedure and finally go wrong. Different from these algorithms, by introducing spatial selective attention in the sampling procedure, our SFAVT algorithm will effectively obtain useful information from the image and construct proper temporal appearance model with relative limited patches. Meanwhile, feature attention in learning procedure strengthens those features that discriminate target from background and restrains those features that confuse target with background. The indirect learning by tuning the patch models with feature attentional spaces could reduce the confusion of similar distracters as much as possible and meanwhile requires only a simple model updating procedure with low computational complexity.

III. SFAVT ALGORITHM

Tracking is difficult when objects and background change their appearance, and therefore binary classification tracking has been a focus of recent work for its good performance on this problem by maintaining temporal integration. However, robust tracking based on accurate model updating is of high computational complexity, which cumbers its application. Selective attention plays a crucial role by selecting information for prioritized process from a huge amount of information contained in visual scenes. Of which spatial selection directs the attention to a restricted region of the visual field and enhances efficiency of representing the scene, while feature selection makes attention strengthen active feature signal and boosts the contrast sensitivity between target and background. Introducing attention into the binary classification tracking can greatly solve the bothering problem of high computational complexity.

Fig.1 shows the framework of the SFAVT algorithm, where *F* denotes the target (foreground) and *B* denotes the background while subscript *i* shows the index of current frame. Ω is the early training set generated by sampling the foreground and background image, which are denoted by superscipted *F* and *B* respectively. And then, the later training set *T* is tuned by the weight *W* calculated from foreground/background histograms. *P* denotes the newly sampled testing set, and finally the classifier is denoted by *C*. The whole algorithm is devided into two stages, the initilization phase and the updating phrase, and they are denoted in the figure with light grey and dark grey seperately. The introduction of spatial attention and feature attention in bold box and dashed box, play important roles in parts of the whole algorithm to help focus computational resource on efficient model initializating and updating, as well as mode seeking.

Take a certain frame *I* for example. At the beginning of the updating phrase, Ω_i and *F/B* of the previous frame, *I-1*, have been generated. Then, the algorithm will update W_i and that will be introduced to tuneΩ_i into T_i. After these steps, a new classifier will be prepared for this *Ith* frame. At the same time, testing set P_i shall be sampled from the background image of B_i according to spatial attention. Next, C_i is going to tell foreground patches of the testing set from background ones, leading to F_{i+1} and B_{i+1}. And finally, we come to a simple tracker such as meanshift [14] to identify the peek of the classification result (confidence map), namely the location of the target. At last, spatial attention will be employed to update the foreground/background model and prepare a new early set for the $(I+1)^{th}$ frame.

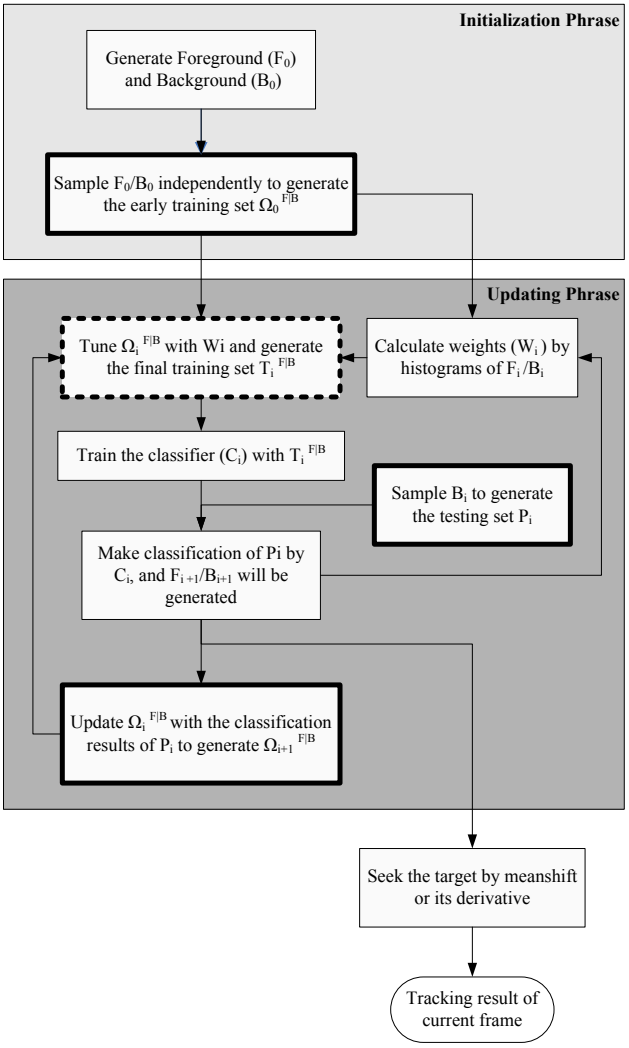


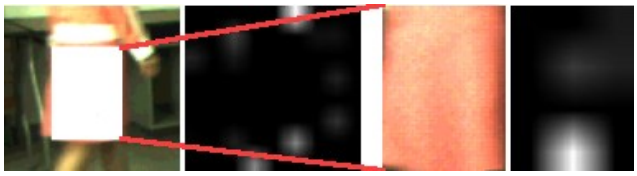
Fig.1 Framework for the SFAVT algorithm
(dashed box denotes a feature-attention driven processing; bold box denotes a spatial-attention driven processing)

A. Early Modeling

Attention is used by primates to selectively focus on some aspects of environment they are interested in and so could help to reduce computation of the brain. Early spatial attention is usually driven by external stimuli (e.g. bright light and color, distinctive shape or motion) sensitive to low-level visual layer, such as retina, LGN (the lateral geniculate nucleus) and V1 (the primary visual cortex). This procedure has little relationship with high-level vision driven by task solving and is efficient to help observe and catch the main characteristics of an unknown object or a new scene.

This section describes a spatial selective method for constructing initial foreground and background models. Both models are built by attentional patches from a fixed foreground window and a surrounding "context window" as

shown in Fig.2 (c) and (a). Early spatial selective visual attention process performs as the preliminary filter based on innate principles of human perception that finds the patches that are more likely to attract human visual attention and serves as a direction for the global searching. It's not difficult to sense that moving objects are generally easier to be caught in one's eyes; also, the attention is usually sensitive to the regions of color and intensity different from the environment around it. Those patches selected by early spatial attention would be best reflecting the natural characteristics of those two parts and hence save the computation to process redundant information.



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

Fig.2 salient map of foreground and background
((a) and (c) are the background and foreground windows sampled for initial models. (d) and (b) are the salient maps of those two windows according to the visual attnetion model [8])

In view of this, we choose to construct "visual bags" models of foreground and background $\Omega^{F|B}$ directed by Itti and Koch's saliency-based visual attention model [15], which computes three center-surround features (image intensity contrast, red/green and blue/yellow double opponent channels, motion) using Center-surround operations and then combines a salient map to define the salient image location that spatial visual attention would shift to as shown in Figure 2. According to the salient map, those regions with high salience will gain more attention and hence high sampling rate while those with low salience even no salience will gain low sampling rate. In this way, the foreground and background model $\Omega^{F|B}$ with attentoinal patches will be the most appropriate for later learning procedure.

B. Tuning

Attention not only shifts in spatial location but also in multiple feature spaces. The later shift is driven by task and could bring potential benefits to visual search and visual tracking. As we know, the efficiency of tracking depends on the salience ratio between the target and the background. The tracking of a man in red cloth under the sun is very easy for its salient color. However, it's very difficult to keep tracking of him when he walks into a shadow. At this time, red is not salient anymore, and we should shift our attention from red color to distinctive shape in feature spaces. Therefore, feature selective attention plays a more important role in the learning procedure of visual tracking to find best feature spaces that could efficiently discriminate the target from the background.

In our approach, the contribution of one feature to distinguish a target from the background is calculated by the variance ratio of the log likelihood function [12]. The discrete probability distributions of one stimulus feature $p(i)$ in target and $q(i)$ in background are separately estimated by normalizing their feature histograms $H_T(i)$ and $H_B(i)$ obtained from target and background windows with the number of pixels n_T and n_B in it,

$$p(i) = H_T(i) / n_T \tag{1}$$

$$q(i) = H_B(i) / n_B \tag{2}$$

where index i ranges from 1 to 2^b indicating the patches, b is the number of histogram buckets.

The *log* likelihood of the feature value i is then given by

$$L(i) = \log \frac{\max\{p(i), \sigma\}}{\max\{q(i), \sigma\}} \tag{3}$$

where σ is a small value like 0.001 that prevents dividing by zero or taking the log of zero.

The variance ratio $VR(L;p,q)$ of $L(i)$ is calculated below to quantify the feature's contribution to distinguish the target from the background. Given a discrete probability density function $d(i)$, the variance of $L(i)$ with respect to d is calculated as following:

$$\mathrm{var}(L;d) = \sum_i d(i)L^2(i) - [\sum_i d(i)L(i)]^2 \tag{4}$$

The variance ratio of the *log* likelihood function L can now be defined as:

$$VR(L;p,q) \equiv \frac{\mathrm{var}(L;(p+q)/2)}{[\mathrm{var}(L;p) + \mathrm{var}(L,q)]} \tag{5}$$

As proved in [16], a feature is relevant and receives high weight if it renders that target more salient than the distracters in the background. Updating this weight is shifting attention in feature spaces and is highly decided by the constantly changing appearances of foreground and background. So, in our approach, the variance ratio of the log likelihood function of each feature f is calculated and normalized to determine the weight of each feature, ω_i:

$$\omega_i = \frac{VR_i}{\sum_{i=n} VR_i} \tag{6}$$

The features of those image patches in foreground and background model will be tuned in each frame based on the weights gained above as shown in Fig.3. Then they will be used in the step of learning as foreground and background model $T^{F|B}$. The RGB part of the tuned feature spaces is also demonstrated in the second row of Fig.3. Features (e.g. RGB, textures, and orientation) chosen in our approach will be

combined by changing their weights online, and in this way, it is not necessary to set a constant feature space as [12] before coming to the next step. Consequently, it is unnecessary to learn a bunch of observation of several displays in advance.

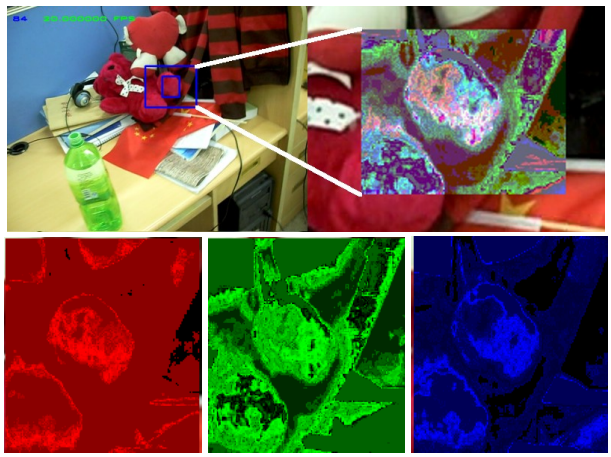


Fig. 3 The tuned feature spaces and their backproject

C. Learning and Tracking

Given a target/background model $T_t^{F|B}$ constructed by spatial selective attention and then tuned by feature selective attention at time t, we train a target/background binary classifier C_t, and use this classify patch samples P_{t+1} obtained by spatial attention at next frame and generate a likelihood map l_P^F for location estimation. Quite different from [13], our tuned model $T_t^{F|B}$ not only contains all appearance model history, but also maximizes the real-time seperation between target and background, in which way the munutal consistence and on-line appearance changing will be naturally maintained in one model. Meanwhile, because no hard decisions over P_{t+1} are made, any classification algorithm with reasonable performance can be employed as [13]. Then, each patch foreground likelihood value l_P^F will be mapped onto an image coordinates as in [14] to create a confidence map. Meanshift algorithm [8] is used from L_t to locate the mode of this map and assign it as the object position L_{t+1}.

D. Model Updating

Model updating is a crucial procedure to maintain the on-line appearance of target and background and hence enhance the robustness and efficiency of the tracker. However, neither the predefined weak classifier updating of ensemble tracking [14] nor the bidirectional consistency check in non-parametric tracking [13] solved well the confusion between foreground and background. The reason is that their foreground and background models have too high correlation with the learning procedure. In this way, once the updating of those models does not quite catch the changing of target and environment, the confusion between them will directly deteriorate the tracking result.

In our algorithm, the updated foreground and background patch bags will not directly be used in the learning procedure. In each frame, those patches will be tuned by adaptive feature spaces selected by feature attention according to the real-time situation and then used in the learning procedure of the classifier. So, the tracking result will depend on both the collected patch models and the adaptive accommodation of feature spaces by feature selective attention. Hence the updating of patch models can have more attention on keeping mutually consistent.

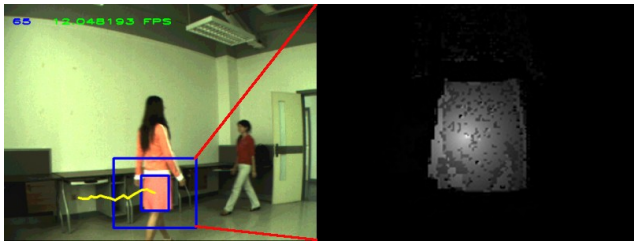


Fig.4 Demonstration of spatial attention
(The yellow line represents the motion information of the target and the light point in the left figure is the center of spatial attention)

In addition to feature attention employed for adaptive weights, spatical attention is taken into account in updating the early Ω. Prominent spatial information could be the peek of the foreground samples (that is, the representative patch of the target's motion), and attention deteriorates as distances to the centroid become longer as shown in Fig.4. Nevertheless, imprudently updating patches with less spatial information, usually the marginal ones, may lead to local minima. In this case, perturbation is necessary. In our algorithm, marginal patches will be updated by wrongly classified testing samples, and randomly, substitution for interior ones will also be employed to avoid local minima problems.

IV. Experiments

Based on naturally employment of spatial attention and feature attention in our binary classification based tracking, extensive experiments are carried out on an ordinary personal computer. Although the single-core central processing unit with 2.8GHz frequency is almost out of date, our algorithm still maintains a smart result. Details about time efficiency will be shown in the first section. Then, we will discuss the performance of SFAVT comparing with non-parametric algorithm and meanshift algorithm. Due to the collaboration

of the two attentions employed in the algorithm, simple feature, even RGB, performs well in the experiments. Other features, such as different color spaces, textures, orientations, etc, are also tested for the final scrutiny. Nevertheless, there is virtually little improvement achieved. Such results may coincidently prove that attention rarely depends on feature selection. Another opening of SFAVT is the learning algorithm, namely, the classifier. Of all the classifier tested (different kernels of SVM, K-near and Boosting), RBF-based SVM gained the best efficiency due to fewer times of iteration. After all, machine learning is still a challenging field and it is impossible to specify a best classifier. For this reason, we

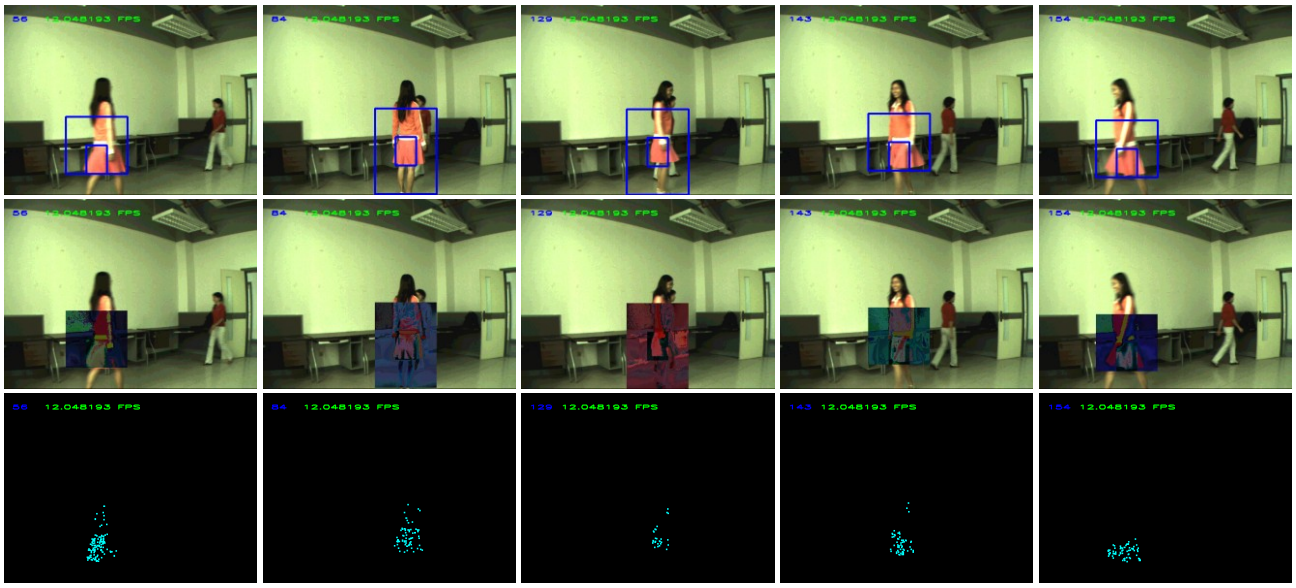


Fig.5 Tracking for frames #56, 84, 129, 143, 154 during changing illumination
((1st row) SFAVT tracker (2nd row) tuned feature backproject (3rd row) confidence map)

Fig.6 Tracking for frames#52, 107, 118, 145, 173 across dramatic distracters
((1st row) SFAVT tracker (2nd row) Non-parametric tracker (100/100 traning set) (3rd row) Meanshift algorithm)

Fig.7 Tracking for frames#31, 57, 100, 129, 152 in cluttered environment
((1st row) SFAVT tracker (2nd row) Non-parametric tracker (100/100 training set) (3rd row) Meanshift algorithm)

leave our machine learning part opened and it is not restrained to any fix mode.

A. *Time Efficiency*

Usually, algorithms for tracking come slow in order to guarantee an online learning phase. The more samples there are, the more accurate the tracking algorithm will be. To our disgust, as accuracy improves, the classifier takes more time to be trained. Table1 shows the time efficiency of SFAVT, non-parametric algorithm and meanshift..

Table 1 Comparison of time efficiency and performance

Algorithm	SFAVT			Non-parametric			Mean shift
TS	100	300	500	100	300	500	
Exp1/ms	72	164	321	78	167	302	78
Exp2/ms	81	153	313	76	167	289	93
Exp3/ms	86	164	313	79	167	288	94
Exp4/ms	81	168	305	83	143	289	78
Exp5/ms	77	172	314	80	172	309	94
Accuracy	Steady	Steady	Steady	Lost	Tracked	Steady	Lost

Five experiments are carried out respectively with 3 different sizes of training sets. As shown in Table1, Exp1~Exp5 denotes the average time required of these three tracking tests. TS denotes the size of the training set. For the number 100, the size of the positive set and the negative will be of the same magnitude, 100. And accuracy in the last row denotes the steadiness of the central tracked target. As shown in Fig. 4, the yellow line denotes the steadiness of the tracker. A "steady" tracker will generate a regular line while a "tracked" tracker just ends up with an irregular one. If the tracker failed to keep up with the target, we call it "lost".

Although the random updated non-parametric tracking algorithm can achieve a steady central point trajectory, it costed too much before taking spatial/feature attention into consideration. Most importantly, the SFAVT algorithm can even attain a comparable efficiency with meanshift with relatively small training sets (the active information) yet attain much better performance (steady trajectory in this case but a lost target).

B. *Changing and Abnormal Illumination*

A girl is walking in Fig.5 with light changing in different places. The target, the girl's pink skirt, is also influenced by her quite similar blouse. The first row shows SFAVT tracking results, the second row shows the tuned feature backproject and the third row shows classification confidence map. Despite of the light changing condition of the skirt, SFAVT could efficiently locate the skirt of the girl by spatial selection's updating of foreground/background model and tuning feature spaces as shown in the second row which renders the target more salient from the background with little influtions of the changing illumination.

C. *Moving across Dramatic Distracters*

Moving across dramatic distracters with similar appearance would be a great challenge for tracking. Few trackers could handle this problem well for its poor object model maintaining procedure and low consideration of relationship between the target and background. In the second and third row of Fig.6, the tracking of a girl in black overcoat by non-parametric

algorithm and meanshift algorithm is remarkably distracted by the guys in black coat walking around her. The same situation also happens in the second and third row of Fig. 7 of the two algorithms for tracking a bear doll's right feet with similar objects around it, including the bear doll's the other feet, its body and some other complex distracters. However, SFAVT shows robust performance in the first row due to the model updating with spatial attention (motion information, distance deterioration) and prohibition of the effect of color features by approriate tuning (see Fig.8 to fix the idea) of feature attention. The variation of the weights employed for tuning is demonstrated in Fig.8 (for the bear doll video, merely RGB feature).

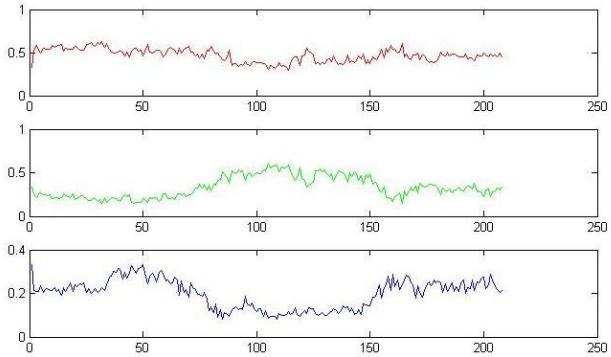


Fig.8 Normalized weights variation (red, green, blue for RGB)

V. CONCLUSIONS

In this paper we propose a novel binary classification based tracking algorithm and introduce the collaboration of spatial and feature attention to enhance the efficiency and robustness of tracking performance and meanwhile reduce its computational complexity. First of all, representing a target with spatial attentional image patches bags of foreground and background could capture all two-class image variations throughout the video volume, which helps to tolerate target appearance variations due to partial occlusions, small deformation, and similar distracters in cluttered background. Also, feature attentional spaces are combined by those weighted features that render target more salient from background, and thus allow SFAVT to save its computational complexity on learning framework and discriminate the target from background more quickly. In addition, the training sets of foreground and background patches are tuned by feature attentional spaces before used by the classifier, which reduces the relationship between accurate model updating and robust visual tracking. Consequently, an efficient updating procedure with low computational complexity could maintain the temporally changing appearance models of foreground and background and enhance the robustness of the tracker under various conditions.

Researches on neuroscience show that there are many other kinds of attentional models that help human being gain information more efficiently. Object-based attention may selectively enhance an object even if there is another object which is spatially superimposed [18]. The large shape variation or full occlusion in tracking may be perfectly solved by combining this kind of attention and our future work will mainly focus on it.

ACKNOWLEDGEMENT

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

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah. *Object Tracking: A Survey*, ACM Computing Surveys. Volume 38, No. 4, pages 1-45, 2006

[2] C. Veenman, M. Reinders, and E. Backer. *Resolving motion correspondence for densely moving points,* IEEE Transaction on Pattern Analysis and Machine Intelligence. Volume 23, No. 1, pages 54-72, 2001

[3] H. Schweitzer, J. W. Bell, and F. Wu. *Very fast template matching.* European Conference on Computer Vision. pages 358-372, 2002

[4] D. Comaniciu. *Bayesian kernel tracking*, Annual Conference of the German Society for Pattern Recognition. pages 438-445. 2002

[5] A. Ali and J. Aggarwal. *Segmentation and recognition of continuous human activity*, IEEE Workshop on Detection and Recognition of Events in Video. pages 28-35. 2001

[6] A. Jepson, D. Fleet, and T. El-Maraghi. *Robust online appearance models for visual tracking*, Computer Vision and Pattern Recognition. Volume 1, pages 415-422, 2001

[7] K. Toyama and A. Blake. *Probabilistic tracking in a metric space*, International Conference on Computer Vision. Volume 2, pages 50-57, 2001

[8] D. Comaniciu, V. Ramesh and P. Meer. *Kernel-based object tracking*, IEEE Transaction on Pattern Analysis Machine Intelligence. Volume 25, pages 564-575, 2003

[9] W. James, *The Principles of Psychology*, Henry Holt, New York. 1890

[10] M. Yang, J. Yuan and Y. Wu. *Spatial selection for attentional visual tracking*, IEEE Conference on Computer Vision and Pattern Recognition. Volume 2, pages 1-8, 2007

[11] E. Blaser, Z. Pylyshyn, and A. O. Holcombe. *Tracking an object through feature-space*, Nature. Volume 408, pages 196-199, 2000

[12] R. T. Collins and Y. Liu. *On-line selection of discriminative tracking features*, International Conference on Computer Vision. Volume 1, pages 346-352, 2003

[13] L. Lu and G. D. Hager. *A nonparametric treatment for location/segmentation based visual tracking*, Computer Vision and Pattern Recognition. pages 1-8, 2007

[14] S. Avidan. *Ensemble tracking*, Computer Vision and Pattern Recognition. Volume 2, pages 494-501, 2005

[15] L. Itti and C. Koch. *Feature Combination Strategies for Saliency-Based Visual Attention Systems*, Journal of Electronic Imaging. Volume 10, No. 1, pages 161-169, 2001

[16] V. Navalpakkam and L. Itti. *An Integrated Model of Top-down and Bottom-up Attention for Optimal Object Detection*, Computer Vision and Pattern Recognition, pages 2049-2056, 2006

[17] H. Grabner and H. Bischof. *On-line boosting and vision*, Computer Vision and Pattern Recognition, Volume 1, pages 260-267. 2006

[18] G. A. Alvarez, T. S. Horowitz, H. C. Arsenio and J. S. DiMase. *Human Perception and Performance,* Journal of Experimental Psychology. Volume 31, No. 4, pages 643–667, 2005