Applied Mathematics & Information Sciences
An International Journal

An Improved Discriminative Tracking Algorithm Based on Spatial Information and Model Updating

Liangwei Jiang[1]*, Jiang Jiang*[1] *and Nong Sang*[1]

[1]Science & Technology on Multi-spectral Information Processing Laboratory, Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, 430074, P.R. China

Received: 4 Sep. 2012; Revised 17 Oct. 2012; Accepted 23 Nov. 2012
Published online: 1 Mar. 2013

Abstract: Tracking object occluded partially is a difficult problem in video surveillance. Many previous tracking methods fail to track occlusion objects robustly. In this paper, we propose an improved discriminative tracking algorithm based on bag of patches to cope with the partial occlusion as well as drift. In the proposed method, the spatial information is introduced to build the object appearance model and construct the confidence map from three different aspects, which directly determine the ultimate tracking effect. In addition, the context information of the small image patches is also applied. In order to adapt to the variance of the environment, an online model updating strategy is proposed. Contrasting experimental results on several real world scenarios show that our proposed approach can handle partial occlusion and recover from drift. Comparing with four stat-of-the-art tracking methods, our proposed method has better tracking performance.

Keywords: Bag of patches, discriminative tracking, appearance model, spatial information, model updating

1. Introduction

Visual tracking is a very important and hot issue. It has been applied in various fields in computer vision, such as security and surveillance, human-computer interaction and so on. Although a great deal of researches have been conducted and lots of tracking approaches have been developed for this topic in the past few decades, however, to track object in the complex real world is still a challenging task due to noise, occlusion, deformation, varying viewpoints, background clutter, illumination change, and so on.

Reviewing all related work about visual tracking is beyond the scope of this paper. In this paper, we only pay attention to two categories tracking methods which are distinguished according to modeling only the object or both the object and the background. As for the former, kernel based trackers [1–5] and particle filter based trackers [6–8] are two of the most popular methods. In addition, the covariance tracker [9] and incremental tracker [10] also belong to the former. An obvious disadvantage for these approaches is that they employ the information only from the object and lose sight of the background information so that these trackers cannot distinguish the object from the background well. In order to make up this shortfall, another type of tracking algorithms using both the object and background information were proposed firstly by Collins [11]. At the same time, Avidan [12] proposes the *ensemble tracker* which regards the tracking as a binary classification problem where many weak classifiers are trained online and then they are combined into a strong classifier by Adaboost. In the training phase, the positive samples are sampled near to the object and the negative ones are sampled from the surrounding background region. Due to good performance of this type of approaches, many learning based tracking algorithms [13–17] have been developed in the last several years. They try to improve the tracking performance by introducing different learning technology into this framework such as semi-supervised learning [14, 15, 17] and multiple instances learning [16, 17]. However, these learning based trackers considerate the problem only from the classification point of view and the bios will be introduced into the model inevitably so that the descriptive ability of the model is inadequate [18, 19]. Different from the Collins's method and learning based approaches, Wang [20] segments the object and background

* Corresponding author: e-mail: nsang@hust.edu.cn

regions into supper pixels and utilizes dictionary learning algorithm to construct object model. Then it will take too much time to collect train frames and learning the dictionary via segmentation and clustering. Tang [18, 19] models the object appearance by means of a bag of small image patches with two different scales from object and background region. The purities of these image patches belonging to the object are estimated using K-NN classifier. After that, these model image patches are utilized to calculate the confidence map to locate the object position. Representing the model using small image patches makes the model more robust to occlusion. However, this model built above is still inadequate in descriptive ability. The tracking performance of this algorithm can be improved by making full use of the spatial information between image patches and their context when building appearance model and constructing the confidence map, which is very important for improving the performance of the appearance model.

In this paper, we propose an improved discriminative tracking algorithm based on bag of patches. In our proposed method, two main contributions have been made. When building the discriminative appearance model, the calculation of the image patch purities belonging to the object take the spatial information between image patches and context information (we will explain it in the following section) into consideration. At the stage of constructing the confidence map, certain an object image patch in the appearance model acts on all its K-NN test image patches from candidate object region, but with different degrees. Therefore, we treat these test image patches differently via taking this diversity into consideration using spatial information weighting. Moreover, the context information is also employed. In addition, in order to make our proposed method more robust to occlusion and drift, an online model updating strategy is proposed.

The rest of this paper is organized as follows. In section 2, we present our proposed discriminative tracking method in details, which includes the tracking framework, building appearance model, locating object position and online model updating. Section 3 presents the implementation details of our proposed method and performs experimental comparison with the state-of-the-art. In section 4 we draw the conclusions.

2. The proposed method

We present the details of the proposed object model and tracking algorithm in this section.

2.1 The workflow of the tracking method

Our proposed algorithm includes three inter-related parts, namely, building appearance model, constructing confidence map to locate the position and online updating model. The workflow of our proposed is illustrated in Figure 1.

At the first frame, we initialize the object status by boxing out the object to be tracked by hand with rectangular box. Then two bags of image patches are sampled respectively from the object region and its surrounding background region. Finally, these two bags of patches are used to calculate the purity of each patch from the bag of object patches with K-NN classifier and build the object appearance model so that we can construct the confidence map in the next part.

When a new frame comes, a bag of candidate test image patches are sampled in a search window around the predicted position. Then employ the appearance model built above to evaluate the confidence of each test image patch. After that the confidence of each pixel in the search window is obtained. Final a mode seek algorithm is acted on the confidence map to get the optimal position.

After obtaining the optimal status in current frame, the appearance model is updated using the information from current frame so that the tracking method can adapt to various variance caused by partial occlusion, noise, deformation, illumination change and so on.

The details of them will be described in the following subsections respectively.

2.2 Building Appearance Model

We represent the object with a rectangular box $\boldsymbol{x}_t$. At the first frame ($t = 1$). $\boldsymbol{x}_1$ is boxed out by hand or detected by detector trained in advance. An appearance model $\boldsymbol{\Omega}_t = \{(p_i^o, \alpha_i^o)\}_{i=1}^{N_1}$ is modeled by a bag of image patches sampled inside $\boldsymbol{x}_1$ as well as the probabilities belong to the object, which are denoted with $\boldsymbol{P}_o = \{p_i^o\}_{i=1}^{N_1}$ and $\{\alpha_i^o\}_{i=1}^{N_1}$ respectively. In order to determine the model, a key step that we have to do is how to calculate the probability α_i^o for each object image patch p_i^o. This probability is left to three factors: (1) the target-background purity based on the object region and its surrounding background region; (2) the distance to the center of the object; (3) the purities of these image patches around p_i^o. We denote these three factors as $\beta_{i,1}$, $\beta_{i,2}$ and $\beta_{i,3}$ respectively. For each of them, the spatial information is utilized to construct the corresponding probabilities by different ways. The details will be presented in the following.

For the first factor, a bag of background image patches $\boldsymbol{P}_b = \{p_j^b\}_{j=1}^{N_2}$ are sampled outside $\boldsymbol{x}_1$ and in its surrounding region firstly. Then K nearest neighbor patches $\boldsymbol{R}_i = \{p_{i1}, p_{i2}, \cdots, p_{iK}\}$ are selected from $\boldsymbol{P}_o \cup \boldsymbol{P}_b / p_i^o$ using K-NN Euclidean distance metrics in feature space for each $p_i^o \in \boldsymbol{P}_o = \{p_i^o\}_{i=1}^{N}$. Finally, the probability is calculated as follows:

$$\begin{cases} \beta_{i,1} = \frac{1}{Z_i} \sum\limits_{j=1}^{K} \gamma_{ij} \cdot I(p_{ij} \in \boldsymbol{P}_o), \ p_{ij} \in \boldsymbol{R}_i. \\ \gamma_{ij} = \omega_c(c_i^o, c_{ij}) \cdot \omega_f(f_i^o, f_{ij}). \\ I(p_{ij} \in \boldsymbol{P}_o) = \begin{cases} 1, & if \quad p_{ij} \in \boldsymbol{P}_o. \\ 0, & otherwise. \end{cases} \end{cases} \tag{1}$$

where Z_i is a normalization factor ensuring $\sum_{j=1}^{K} \gamma_{ij} = 1$. By introducing weight term $\omega_c(c_i^o, c_{ij})$ in the original image space and weight term $\omega_f(f_i^o, f_{ij})$ in feature space we

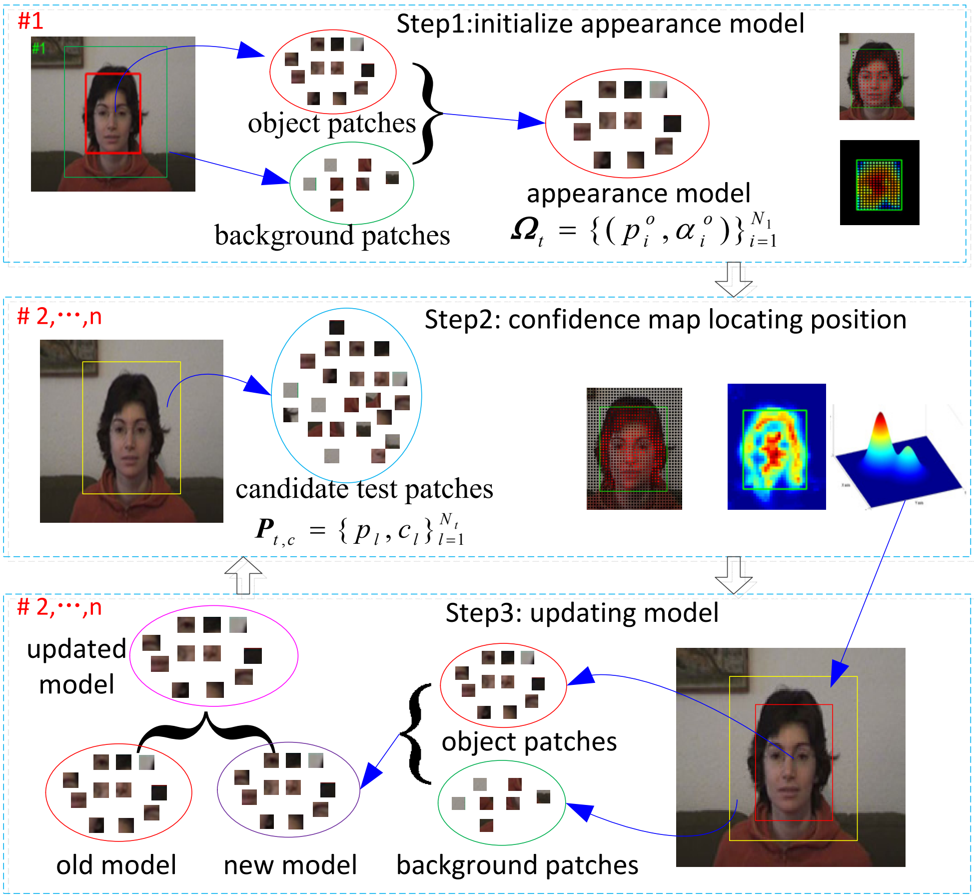


Figure 1 The framework of our proposed tracking algorithm.

emphasize that each image patches in $\boldsymbol{R}_i$ does not make an equal contribution to calculate the purity of image patch p_i^o . $\omega_c(c_i^o, c_{ij})$ descripts the distance between image patch center c_i^o and c_{ij} of p_i^o and p_{ij} respectively. $\omega_f(f_i^o, f_{ij})$ denotes the weighting term based on the distance between feature f_i^o and f_{ij} of image patch p_i^o and p_{ij} respectively. For these two weight terms, we design them with Gaussian function as formula (2).

$$\begin{cases} \omega_f(f_i^o, f_{ij}) = \exp(-\lambda_f \|f_i^o - f_{ij}\|^2) \\ \omega_c(c_i^o, c_{ij}) = \exp(-\lambda_c \|c_i^o - c_{ij}\|^2) \end{cases} \tag{2}$$

This two weighting terms indicate that the closer and more similar image patches in $\boldsymbol{R}_i$ will make greater contribution, rather than in reference [18, 19] which accumulates directly with 1 for all patches without difference, which is unreasonable.

At the same time, it is more scientific to distribute different probabilities to different patches sampled inside the object rectangular box sine they lay from the center of the object with different distances. Larger probability should be distributed to the patches which are closer to the center of the object rectangular box; vice versa, the patches further from the center of object rectangular box must be distributed with smaller probabilities due to these marginal patches may be background noise. Based on this fact, the second factor $\beta_{i,2}$ is designed as formula (3) via taking the spatial distance into account.

$$\beta_{i2} = D(c_i, c_o) = \exp(-\lambda_d \|(c_i - c_o)/h\|^2) \tag{3}$$

where c_o and c_i are the centers of the object rectangular box $\boldsymbol{x}_1$ and small image patch p_i^o respectively. h denotes the radius of the object rectangular box. Up to now, we can obtain a rough appearance model $\{p_i^o, \tilde{\alpha}_i^o\}$ via adding $\beta_{i,1}$ and $\beta_{i,2}$ to obtain $\hat{\alpha}_i^o$

It is an indisputable fact that a certain image patch is not independent on these image patches around it, but related closely. A certain image patch must belong to the object with larger probability, if the other image patches around it belong to the object with larger probability. Based on this spatial context relationship, we design the third probability term $\beta_{i,3}$ as formula (4).

$$\begin{cases} \beta_{i3} = \exp(\lambda_{sc} S(i,8)). \\ S(i,8) = 0.125 \cdot \sum_{k=1}^{8} \hat{\alpha}_k^o,\ p_k^o \in \boldsymbol{P}_o \cap N(p_i^o, 8). \end{cases} \tag{4}$$

where $N(p_i^o, 8)$ denotes the eight domain image patch set of the patch p_i^o. λ_{sc} is a normalization factor.

Therefore, the ultimate probability α_i^o of image patch p_i^o is set as $(\sum_{j=1}^{3} \beta_{i,j})/3$. This probability takes spatial information and the spatial context into account to improve the object appearance model. Comparing with the appearance model proposed in [18, 19], this appearance is more reasonable and scientific than (see Figure 2). The appearance models obtained by our proposed method are shown in subfigure (g) and (i), the ones obtained by [18, 19] are shown in subfigure (m) and (o). The (g) and (m) are the initial appearance models for the first frame, (i) and (o) are the updated appearance model for the 21th frame. From these comparing appearance models, we can see that both appearance models can suppress the background information on the edges effectively. However, the appearance model obtained by our proposed method make these purities of the image patches in the appearance model more diverse and reasonable than the ones obtained by [18, 19].

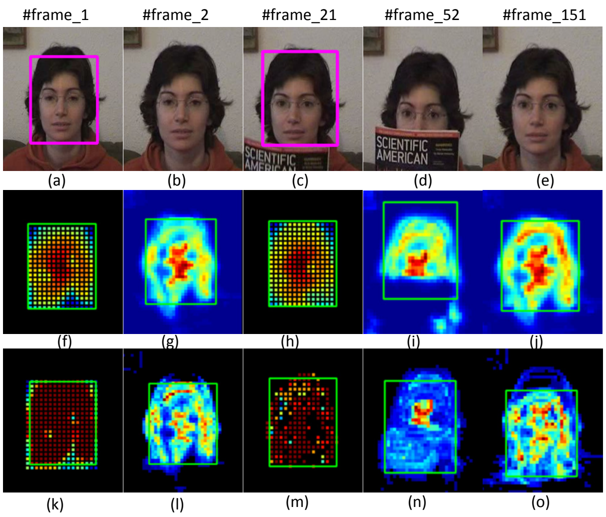


Figure 2 The appearance models and confidence maps obtained respectively by [18, 19] and our method. The (a) - (e) are the original images. The (f) and (h) are the initial appearance model at the first frame and updated appearance model at 21th frame for our proposed method. The (k) and (m) are the initial appearance model at first frame and the updated appearance model at 21th frame for the method proposed in [18, 19]. The (g) and (i)-(j) are the confidence maps for our proposed method. The (l) and (n)-(o) are the confidence maps for the method proposed in [18, 19].

2.3 Locating object position

When a new frame arrives, in order to locate the object position, we must construct a confidence map for current frame firstly. We extract a surrounding region of the object and sample N_t small candidate test image patches $\boldsymbol{P}_{t,c} = \{p_l\}_{l=1}^{N_t}$.For each object image patch from object appearance model $p_i^o \in \boldsymbol{P}_o$, we select the K nearest neighbor image patches from $\boldsymbol{P}_{t,c} = \{p_l\}_{l=1}^{N_t}$ and make up its K-NN lists, which is denoted as $\tilde{\boldsymbol{R}}_i = \{\tilde{p}_{i1}, \tilde{p}_{i2}, \cdots, \tilde{p}_{iK}\}$.In order to compute a confidence map for current frame, we evaluate every test image patch and compute its confidence measure c_l , which depends on the object image patches in the appearance model as well as their probabilities belonging to the object. The relation between them is that if a certain test image patch p_l belongs to the K-NN list $\tilde{\boldsymbol{R}}_i$ of an object image patch p_i^o in the feature space, then this object image patch as well as its probability of α_i^o will act on the test patch p_l . The confidence measure of each test image patch is calculated as follows:

$$\begin{cases} c_l = \sum_{i=1}^{N_1} \omega_f(f_l, f_i^o) \cdot \alpha_i^o \cdot I(p_l \in \tilde{\boldsymbol{R}}_i) \\ I(p_l \in \tilde{\boldsymbol{R}}_i) = \begin{cases} 1, & if \quad p_l \in \tilde{\boldsymbol{R}}_i \\ 0, & otherwise. \end{cases} \end{cases} \tag{5}$$

where $\omega_f(f_l, f_i^o)$ is a weighting term similar with formula (2), which takes the distance metric into account. The farther the object image patch p_i^o lies from the test image patch p_l, the less the object image patch makes contribution for the test image patch. However, the author in [18, 19] distributes all test image patches belonging to $\tilde{\boldsymbol{R}}_i$ with the same weights α_i^o indiscriminately; it's unreasonable since the similarity between them is different.

In order to locate the position of the object in current frame, the confidence of each pixel in current frame must be computed, which can be interpreted as an up-sampling of the per-element saliency. In this paper, we adopt an idea proposed in the context of range image up-sampling [22], which has been applied in saliency filter [23], and apply it to our framework. We define the confidence $\tilde{c}_i$ of a pixel as a weighted liner combination of the confidence c_l of its surrounding image patches. If the pixel is outside the search window, its confidence is zero. After obtaining the confidence map, any mode seek algorithm such as mean shift can be used to seek the position of the object and get its status $\boldsymbol{x}^*$ in current frame.

$$\tilde{c}_i = \sum_{l=1}^{N} w_{i,l} \cdot c_l, \; w_{i,l} = (\exp(-\lambda_{fin} \|c_i - c_l\|^2))/Z_i \tag{6}$$

Similar with [23], we chose the Gaussian function for the weight $w_{i,l}$.

2.4 Online model updating

To adapt to the variance of the object appearance caused by illumination, deformation, and occlusion and so on, the updating of the appearance model is imperative for a robust tracking algorithm. In this paper, the following online updating strategy is adopted.

For each object status $\boldsymbol{x}^*$, its confidence is the sum of the confidence of the pixels inside $\boldsymbol{x}^*$. Among the process of tracking, we record the templates whose confidences are larger than some threshold predefined. The difference ξ between the confidence of the current frame and the average of the confidences in the retained sequence (the last five frames) is calculated, namely,

$$\xi = |conf_t - ave\_cof| / s \tag{7}$$

where $conf_t$ is the confidence of current frame, ave_cof is the average of the confidences in the retained sequence, is the area of the object so that $\xi \in [0,1]$. If $\xi \in [\xi_l, \xi_u]$, $\xi_l < \xi_u$, we update the appearance model with the following our proposed method, this means that we update the model when the object appearance has changed but not too much. If $\xi > \xi_o$, $\xi_o > \xi_u > \xi_l$, which means the object appearance has changed too much(e.g. occluded heavily), the object status of last frame is set as the current status.

When update the appearance model, rather than re-sampling object and background image patches as in [18, 19], we group the test image patches set $\boldsymbol{P}_{t,c} = \{p_l\}_{l=1}^{N_t}$ directly into two subsets, denoted as $\tilde{\boldsymbol{p}}_o$ and $\tilde{\boldsymbol{p}}_b$ respectively, according to they are inside or outside the rectangular box $\boldsymbol{x}^*$. The intuitive advantage of this is that it saves time for re-sampling image patches and computing their feature. Based on $\tilde{\boldsymbol{p}}_o$ and $\tilde{\boldsymbol{p}}_b$, the similar method as previous section is employed to build a new appearance model $\tilde{\boldsymbol{\Omega}}_t = \{(\tilde{p}_i^o, \tilde{\alpha}_i^o)\}_{i=1}^{\tilde{N}_1}$. Next, the patches in $\tilde{\boldsymbol{\Omega}}_t$ with confidence lower than threshold λ_l or larger than threshold λ_u are got rid of. The aim of this is to retain the patches that have changed to some degree but not too much. In order to prevent the size of model from increasing endlessly, the image patches in updated model are composed by resampling from the rest image patches in both old model $\boldsymbol{\Omega}_t$ and new model $\tilde{\boldsymbol{\Omega}}_t$ according to their probabilities belonging to the object, until the size of the updated model achieves the size of the old model.

Based on the detailed analysis of above, we summarize our tracking method in ***Algorithm 1***.

Algorithm 1: Our Proposed Tracker

Initialization: (for t=1)

1. box out the object rectangular box x_1 and its surrounding rectangular box x_b by hand.
2. sample object image patches $P_o = \{p_i^o\}_{i=1}^{N_1}$ inside x_1 and background image patches $P_b = \{p_j^b\}_{j=1}^{N_2}$ outside x_1 but inside x_b.
3. obtain the appearance model $\Omega_t = \{(p_i^o, \alpha_i^o)\}_{i=1}^{N_1}$ according to formula (1)-(4).

Tracking: (for t=2 to end)

1. predict the object position $\hat{x}_t$ using x_{t-1} ,box out search window $\hat{x}_s$ around $\hat{x}_t$.
2. sample a candidate test image patch set $P_{t,c} = \{p_l\}_{l=1}^{N_t}$ inside $\hat{x}_s$.
3. calculate the confidence c_l of each patch in $P_{t,c} = \{p_l\}_{l=1}^{N_t}$ according to formula (5).
4. compute the confidence map for current frame as formula (6).
5. seek the mode of the confidence map using mean shift algorithm and obtain the optimal status x^*.
6. updating the appearance model using our model updating strategy.

3. Experimental Results

In this section, we present experimental results to validate the effectiveness and efficiency of our proposed method. We also conduct a thorough comparison between our proposed method and state-of-the-art tracking methods where applicable.

3.1 Experiment Setup

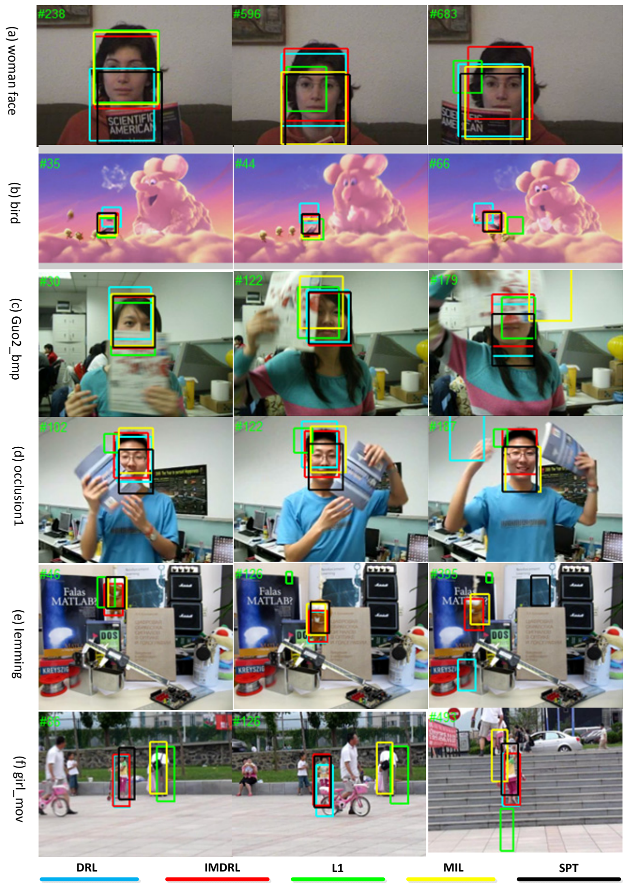


Figure 3 Tracking results of five trackers, namely, DRL, L1, MIL, SPT and our method (IMDRL) on six test image sequences, which are woman face(a), bird(b), guo2-BMP(c), occlusion1(d), lemming(e) and girl-mov sequence(f).

We implement our proposed algorithm with C++. In our experiments, the normalized histogram in RGB color space and histogram of gradient orientation are utilized as features for each image patch. The image patches are sampled according to grid. The initial status of the object is boxed out by hand. The normalization factors λ_f and λ_c in formula (2), λ_d in formula (3), λ_{sc} in formula (4) as well as λ_{fin} in formula (6) are set between 2 and 5 in our experiments. The thresholds λ_l and λ_u for removing the outliers are set as 0.45 and 0.95 respectively. K=16. The sample rate in object region is set to 0.3 in our experiments. The sample rate in background region is set to 0.18 to ensure the number of image patches from background region is not too much. The patch size is determined by the sample rate in object region.

To evaluate the proposed tracking method, six publicly available test image sequences are used, bird (S2) and girl-mov (S6) sequence from [20], woman face (S1) sequence from [21], lemming (S5) sequence from [24], Guo2-BMP (S3) and occlusion1 (S4) sequence from [25]. We compare our proposed method with four state-of-the-art trackers, namely, MIL [16], discriminative rank list tracker (DRL)

[18, 19], supper-pixel tracker (SPT) [20] and L1 [26]. Our proposed method is denoted as IMDRL. The DRL tracker is the base algorithm whose performance our tracker wants to improve. For all them, we employ the codes publicly available or provided by the original authors. All the parameters are set in accordance with the original papers. We compare the performance of these trackers qualitatively and quantitatively.

3.2 Qualitative Evaluation

Firstly, we test our proposed method on four sequences from indoor scene, which are *woman face, guo2-BMP, occlusion1 and lemming* are used. Figure 3(a) and Figure 3(c)-(e) show the sampling tracking results using different schemes on *woman face, guo2-BMP, and occlusion1 and lemming* sequence respectively. From the results of #683 in Figure 3(a), #179 in Figure 3(c), #187 in Figure 3(d) and #395 in Figure 3(e), we can see that when the object being tracked is occluded partially by the book or appearance changes too much, all the other trackers drift from the object a lot, and only our proposed method can track the object accurately.

In order to contrast the tracking performance further, all the trackers are tested on two test sequences from outdoor scene. The results for both sequences are shown in Figure 3(b) and Figure 3(f). As we can see from the results of #66 in Figure 3(b), when the shape of the bird changes, The DRL and L1 drift away from the object. Then the MIL, SPT and IMDRL can track the bird accurately. The results of #126 in Figure 3(f) show that when the moving girl reappears after being occluded by other people, all the other trackers fail to track the object and only our proposed tracker and SPT can still track the object accurately.

Although the MIL and SPT perform comparative with our proposed method on some test sequences, then for the other test sequences, our proposed method is more robust obviously than MIL and SPT. As for L1 and DRL, their performance is inferior obviously to our proposed method.

The reason that our method is better than L1 is that the L1 tracker represents object using information only in object region and loses sight of the information in background region so that it cannot distinguish the object from the surrounding background. Although the DRL tracker take advantage of the information in background region, then DRL tracker build the appearance model without taking the spatial distance and spatial relationship into consideration. Comparing with the MIL tracker which trains classifier using image patches filling the whole object, our method is based on small patches and it is more robust to dealing with occlusion.

3.3 Quantitative Evaluations

In order to evaluate the performance of the proposed tracker quantitatively, we utilize the labeled ground truth bounding box in each frame for six sequences provided by original authors. The difference between the centers of the ground truth bounding box and tracking result box is used as the evaluation indicator.

Table 1 Tracking results, the numbers denote the average errors of center location in pixels

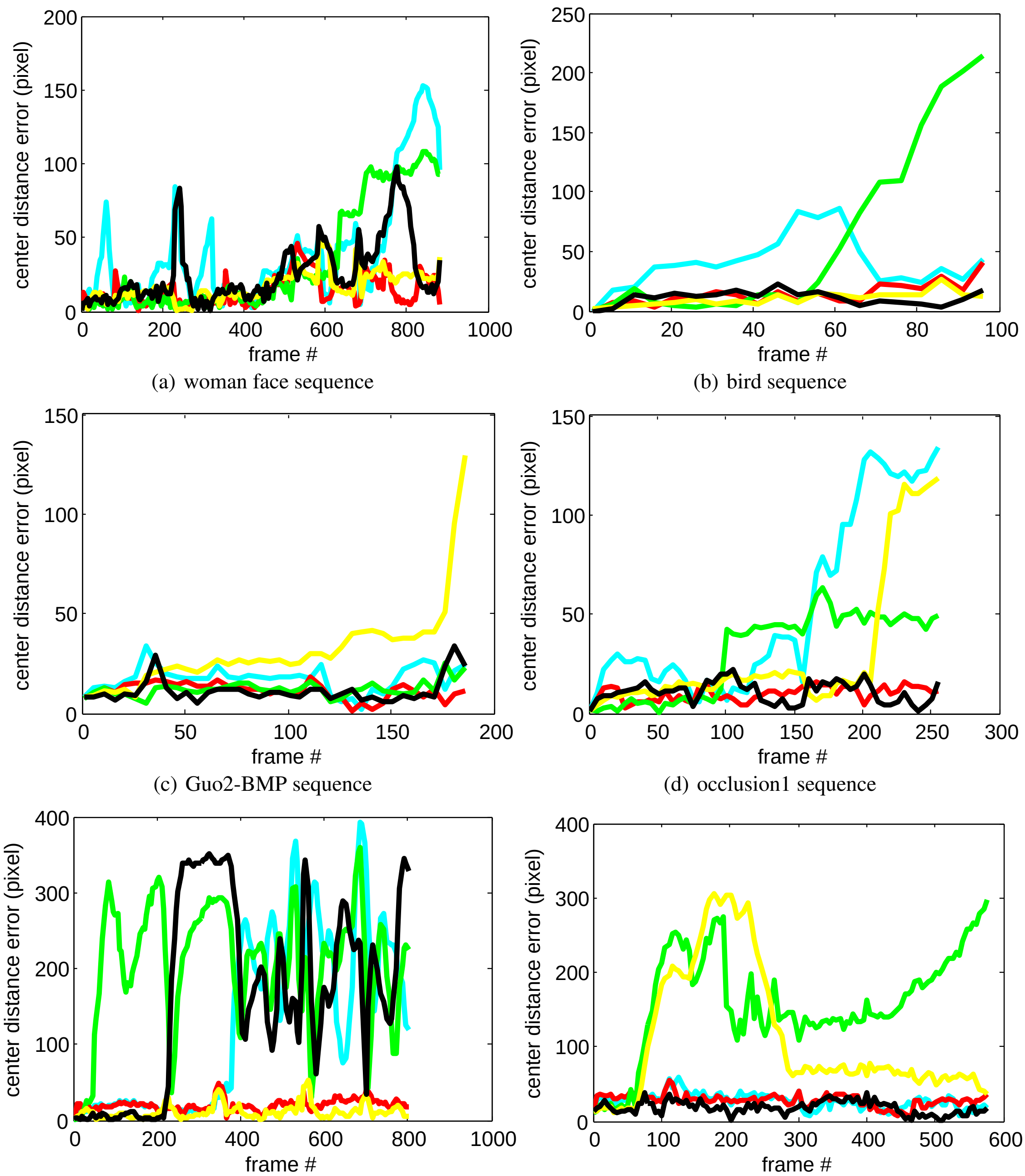
	DRL	IMDRL	L1	MIL	SPT
S1	40.3	**13.8**	33.2	**17.1**	25.7
S2	40.9	**12.1**	61.3	**10.2**	**10.8**
S3	17.3	**11.3**	12.2	31.5	**11.3**
S4	53.6	**9.9**	30.9	30.5	**10.6**
S5	126.9	**20.0**	204.7	**11.3**	163.0
S6	28.3	**27.9**	158.8	112.6	**18.2**
ave	51.2	**15.8**	83.5	**35.5**	39.9

Table 1 gives the average of the tracking errors for each approach on six test sequences. From the statistical results we can see that our proposed method (IMDRL) is much better than the original tracking method (DRL) and L1 tracker, the average errors of center location on all of the six test sequences of IMDRL are smaller than the ones of DRL and L1. Than for MIL tracker, our method is comparative with it on the S2 and S5 sequence. The average errors for our proposed method are 12.1 and 20.0, which are lager a litter than ours on these tow sequence. However, on the rest four test sequences, our proposed method is better than MIL obviously. The tracker whose performance is best comparable to our tracker is STP. Except for S1 and S5 sequence, the average errors of STP on the rest other sequences approximate to the ones of our proposed method, even smaller (S2 and S6). In summary, the performance of our proposed method has better performance than DRL, L1 and MIL in dealing with occlusion and drift, and is comparable to the ones of STP.

To illustrate the superiority of our method further, the tracking error polylines for each algorithm on each test image sequence are also given in Figure 4 respectively. Each subfigure corresponds to one test sequence, and in each subfigure five lines with different colors represent different trackers. From these subfigures, the same conclusion with Table 1 can be obtained.

4. Conclusions

In this paper, we proposed an improved discriminative visual tracking method based on bag of image patches. The proposed approach constructs the object appearance model using the spatial information from three aspects and the context of the image patch, which has been looked down upon by the other authors. In addition, the spatial information and context are also utilized when construct the confidence map to locate the object to be tracked. When updating the model, a novel updating strategy is designed to make it more robust to the various variances. Comparing with several state-of-the-art tracking algorithms, the proposed tracker is superior or competitive.



(a) woman face sequence

(b) bird sequence

(c) Guo2-BMP sequence

(d) occlusion1 sequence

(e) lemming sequence

(f) girl-moving sequence

Figure 4 Center distance error polylines between ground truth boxes and tracking results by DRL, L1, MIL, SPT and our method (IMDRL) on six image sequences.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No.61105014 and No.61271328). We are grateful to all the researchers for their codes and data sets. We are also grateful to the referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] D. Comaniciu, V, Ramesh and P. Meer. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**, 564-577(2003).

[2] R.T. Collins. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition **2**, 234-240(2003).

[3] J. Fang, J. Yang, H. Liu, J Lv and Y. Zhou. Optical Engineering **49**, 087201(2010).
[4] Z. Li, J. Gao, Q. Tang and N. Sang. Optical Engineering **47**, 086402(2008).
[5] C. Shen, J. Kim, and H. Wang. IEEE Transactions on Circuits and Systems for Video Technology **20**, 119-130(2010).
[6] M, Isard and A. Blake. International Journal of Computer Vision **29**, 5-28(1998).
[7] Y. Wang, Y. Tan and J. Tian. Optical Engineering **49**, 077004(2010).
[8] Y. Liu, S. Lai, B. Wang, M. Zhang and W. Wang. Optical Engineering **51**, 047203(2012).
[9] F. Porikli, O. Tuzel and P. Meer. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 728-735(2006).
[10] D. Ross, J. Lim, R. Lin and M. Yang. International Journal of Computer Vision **77**, 125-141(2008).
[11] R. Collins, Y. Liu, and M. Leordeanu. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**, 1631-1643(2005).
[12] S. Avidan. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition **2**, 494-501(2005).
[13] H. Grabner, H. Bischof. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition **1**, 260-267(2006).
[14] H. Grabner, C. Leistner, and H. Bischof. in The 10th European Conference on Computer Vision, 234-247(2008).
[15] S. Stalder, H. Grabner, and L. Van Gool. IEEE 12th International Conference on Computer Vision Workshops, 1409-1416(2009).
[16] B. Babenko, M.H. Yang and S. Belong. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 983-990(2009).
[17] B. Zeisl, C. Leistner, A. Saffari and H. Bischof. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 1879-1879(2010).
[18] M. Tang, X. Peng and D. Chen. Proceedings of the 10th Asian conference on Computer vision-Volume Part I, 283-295(2010).
[19] M. Tang, X. Peng and D. Chen. IEEE Trans. on Image Processing **21**, 3273-3281(2012).
[20] S. Wang, H. Lu, F. Yang and M.H. Yang. In Proceedings of 13th International Conference on Computer Vision, 1323 -1330(2011).
[21] A. Adam, E. Rivlin, and I. Shimshoni. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition **1**, 798-805(2006).
[22] J. Dolson, J. Baek, C. Plagemann, and S. Thrun. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition **5**, 1141-1148(2010).
[23] F. Perazzi, P. Kr?henbl, Y. Pritch and A.Hornung. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 733-740(2012).
[24] J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof. IEEE Conf. on Computer Vision and Pattern Recognition, 723-730(2010).
[25] F.Yang, H. Lu and Y.W. Chen. In Proc. International Conference on Pattern Recognition , 153-156(2010).
[26] X. Mei and H. Ling. IEEE International Conference on Computer Vision , 1436-1443(2009).

Liangwei Jiang received the MS degree in probability and statistics from Huazhong University of Science and Technology in 2008, He current is a candidate PhD student in pattern recognition and artificial intelligence in Zhuazhong University of Science and Technology. His research interests include object detection, recognition and tracking.

Jiang Jiang received the bachelor degree from Huazhong University of Science and Technology in 2005. He current is a candidate PhD student in pattern recognition and artificial intelligence in Zhuazhong University of Science and Technology. His research interests include image representation, object recognition and tracking.

Nong Sang graduated from Huazhong University of Science and Technology and received his BE degree in computer science and engineering in 1990, MS degree in pattern recognition and intelligent control in 1993, and PhD degree in pattern recognition and intelligent control in 2000. He is currently a professor at the Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China. His research interests include pattern recognition, computer vision, and neural networks.