Multi-Modality Tracker Aggregation: From Generative to Discriminative

Xiaoqin Zhang[1], **Wei Li**[2], **Mingyu Fan**[1], **Di Wang**[1], **Xiuzi Ye**[1]
[1]College of Mathematics & Information Science, Wenzhou University, Zhejiang, China
[2]Taobao(China) Software Company Limited, Zhejiang, China
zhangxiaoqinnan@gmail.com, yichan.lw@taobao.com, {fanmingyu, wangdi, yexiuzi}@wzu.edu.cn

Abstract

Visual tracking is an important research topic in computer vision community. Although there are numerous tracking algorithms in the literature, no one performs better than the others under all circumstances, and the best algorithm for a particular dataset may not be known a priori. This motivates a fundamental problem-the necessity of an ensemble learning of different tracking algorithms to overcome their drawbacks and to increase the generalization ability. This paper proposes a multi-modality ranking aggregation framework for fusion of multiple tracking algorithms. In our work, each tracker is viewed as a 'ranker' which outputs a rank list of the candidate image patches based on its own appearance model in a particular modality. Then the proposed algorithm aggregates the rankings of different rankers to produce a joint ranking. Moreover, the level of expertise for each 'ranker' based on the historical ranking results is also effectively used in our model. The proposed model not only provides a general framework for fusing multiple tracking algorithms on multiple modalities, but also provides a natural way to combine the advantages of the generative model based trackers and the the discriminative model based trackers. It does not need to directly compare the output results obtained by different trackers, and such a comparison is usually heuristic. Extensive experiments demonstrate the effectiveness of our work.

1 Introduction

Visual tracking is a key component in numerous video analysis applications, such as visual surveillance, vision-based control, human-computer interfaces, intelligent transportation, and augmented reality. It strives to infer the motion states of a target, *e.g.*, location, scale, and velocity, from the observations in the each frame with the assumption that the appearance of the target is temporally consistent. This assumption may be occasionally violated due to the large appearance variations of a target induced by changing illumination, viewpoint, pose, occlusion, and cluttered background, which leaves visual tracking a quite challenging task after decades of intensive study.

In terms of how to construct the appearance models, the tracking algorithms can be roughly categorized into generative model or discriminative model based methods. The generative model based methods usually build a model to describe the visual appearance of a target, *e.g.*, by a distribution or subspace of intensities or features. Then the tracking problem may reduce to search for the optimal motion state of a target that yields the most similar object appearance in a maximum likelihood formulation. An image patch model [Hager and Belhumeur, 1998], which takes the set of pixels in the target region as the model representation, is a direct way to model the target, but it loses the discriminative information that is contained in the pixel values. The color histogram [Comaniciu *et al.*, 2003] provides global statistical information about the target region which is robust to noise, but it is sensitive to illumination changes and distracters with similar colors. [Stauffer and Grimson, 1999] first employ a Gaussian mixture model (GMM) to represent and recover the appearance changes in consecutive frames. Later, a number of more elaborate Gaussian mixture models are proposed for visual tracking [Jepson *et al.*, 2003; Zhou *et al.*, 2004; Wang *et al.*, 2007]. In [Porikli *et al.*, 2006], the object to be tracked is represented by a covariance descriptor which enables efficient fusion of different types of features and modalities. Another category of appearance models is based on subspace learning. In [Black and Jepson, 1998], a view-based eigenbasis representation of the object is learned off-line, and applied for matching successive views of the object. However, it is very difficult to collect training samples that cover all possible viewing conditions. To deal this problem, subspace learning based tracking algorithms [Lim *et al.*, 2004; Zhang *et al.*, 2010; 2014] are proposed to effectively learn the variations of both appearance and illumination in an incremental way. Generative models generally seek a compact object description and do not take advantage of the background information. Therefore, the generative model based trackers are apt to be distracted by background regions with similar appearances during tracking.

To address the background distraction problem, discriminative models are adopted and trained online during tracking, which concentrate on maximizing the difference between a target and the background. Hence, visual tracking is viewed

as a binary classification problem that finds the optimal decision boundary to distinguish a target from the background. [Collins *et al.*, 2005] firstly note the importance of background information for object tracking, and formulate tracking as a binary classification problem between the tracked object and its surrounding background. In [Lin *et al.*, 2004], a two-class FDA (Fisher Discriminant Analysis) based model is proposed to learn a discriminative subspace to separate the object from the background. In [Avidan, 2007], an ensemble of online weak classifiers are combined into a strong classifier. [Grabner *et al.*, 2006] adopt online boosting to select discriminative local tracking features. [Saffari *et al.*, 2009] introduce a novel on-line random forest algorithm for feature selection that allows for on-line building of decision trees. [Babenko *et al.*, 2011] use multiple instance learning instead of traditional supervised learning to learn the weak classifiers. This strategy is more robust to the drifting problem. Apparently the classification performance of discriminative models largely depends on the correctly labeled training samples. The common choice is to regard the current tracked object as the positive sample and sample its neighborhood locations to select negative samples. The *self-training* nature , that is, the classification results are directly utilized to update the classifier, will lead to overfitting.

Although there are numerous tracking algorithms in the literature, no mater generative based or discriminative based, no one algorithm performs better than the others under all circumstances, and the best algorithm for a particular dataset may not be known a priori. This motivates a fundamental problem-the necessity of an ensemble learning of different tracking algorithm to overcome their drawbacks and to increase the generalization ability. Besides, a natural step forward to cope with the *self-training* problem suffered by the discriminative appearance models is to introduce generative appearance models to supervise the sample selection and training of the discriminative model in a co-training manner [Yu *et al.*, 2008], since the generative and discriminative models have the complimentary advantages in modeling the appearance of the object. However, the major challenge of combining different tracking algorithms is how to measure the performance of different trackers in the absence of groundtruth? When trackers employ different features from multiple modalities, it is hard to directly evaluate their performance [Wu *et al.*, 2013].

To address the above issues, we propose a multi-modality ranking aggregation framework for fusion of multiple tracking algorithms. In our work, each tracker is viewed as a 'ranker' which outputs a rank list of the candidate image patches based on its own appearance model in a particular modality. Then, the proposed algorithm aggregates the rankings of different rankers to produce a joint ranking. Moreover, the level of expertise for each 'ranker' based on the historical ranking results (tracking results before current video frame) is also effectively used in our model. The main features of our work are two-fold: (1) Our work provides a general framework for fusing multiple tracking algorithms employing different features from multiple modalities. It does not need to directly compare the output results obtained by different trackers, and such a comparison is usually heuristic and is only feasible in some specific conditions. (2) Our work provides a natural way to combine the advantages of the generative model based trackers and the the discriminative model based trackers. To our best knowledge, such a framework for fusion of multiple tracking algorithms has not been addressed in the literature.

The rest of the paper is organized as follows. In section 2, the multi-modality ranking aggregation framework is introduced. In section 3, the proposed tracking algorithm is detailed. The experimental results to validate our method are presented in section 4. Some concluding remarks are made in section 5.

2 Multi-Modality Ranking Aggregation Framework

2.1 Ranking Aggregation Model

In this part, we will briefly introduce the theory of ranking aggregation. Let $O = \{o_1, o_2, .., o_N\}$ be a set of object candidates to be ranked by different rankers, and let $r_i \in \mathbb{S}_N$ be the ranking list of ranker i for these candidates where $\mathbb{S}_N$ is a pool of all the possible rankings. Assuming that a set of ranking $R = \{r_1, r_2, ..., r_K\}$ from K individual rankers is available, the probability of assigning a true ranking ξ to the candidates O can be defined using the extended Mallows model [Lebanon and Lafferty, 2002]:

$$p(\xi|\alpha, R) = \frac{1}{Z(\xi, \alpha)} p(\xi) \exp(\sum_{i=1}^{K} \alpha_i d(\xi, r_i)) \qquad (1)$$

where $p(\xi)$ is a prior of ξ, and $Z(\xi, \alpha) = \Sigma_{\xi \in \mathbb{S}_N} p(\xi) \exp(\sum_{i=1}^{K} \alpha_i d(\xi, r_i))$ is a normalizing constant. $d(.,.)$ is the distance measure between two ranking lists, and $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_K\}$ is the expertise of all the rankers which fulfils that $\alpha_i < 0$. The closer of α_i to zero, the less influence of the i-th ranker on the assignment of the probability. When α_i trends to negative infinity, the corresponding ranker trends to be the true rank. So the Eq.(1) calculates the probability that the true ranking is ξ, given the rankings from different rankers R and the degrees of their expertise α. In the learning process, only the rankings from different rankers R are available, based on which we need to infer ξ and α.

In the above model, the distance $d(.,.)$ between two rankings need to be right-invariant, which means the value of $d(.,.)$ does not depend on how the objects are indexed. More specially, if we re-index the object using τ, the distance between two rankings over the objects does not changes: $d(\xi, r) = d(\xi\tau, r\tau)$, where $\xi\tau$ is defined by $\xi\tau(i) = \xi(\tau(i))$. That means, the value of $d(.,.)$ does not change if we re-index the objects such that one ranking becomes $\xi\xi^{-1} = \mathrm{e} = (1, 2, ..., n)$ and the other $r\xi^{-1}$. *Kendall's tau distance* [Lebanon and Lafferty, 2002] is an example of common right-invariant distance function. The distance between ranking ξ and r is defined as the minimum number of pairwise adjacent transpositions needed to turn one ranking into the other:

$$d(\xi, r) = \sum_{i=1}^{N-1} \sum_{j>i} I(\xi r^{-1}(i) - \xi r^{-1}(j)) \qquad (2)$$

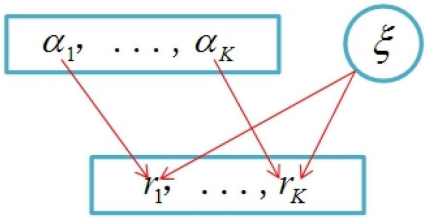


Figure 1: Generative process

[Klementiev *et al.*, 2008] proved that if the distance is right-invariant, the following generative process can be derived based on Eq. (2):

$$p(\xi, R|\alpha) = p(\xi)\prod_{i=1}^{K} p(r_i|\alpha_i, \xi) \tag{3}$$

This generative process can be described in Fig.1. From the generative model, we can find the observation is ranking r_i, ξ and α are the hidden parameters of the model. Based on the generative model, a set of samples can be sampled. These samples are used to approximate the distribution of the rankings.

The parameters α and ξ can be estimated using the EM algorithm [Klementiev *et al.*, 2008] given a set of rankings $R = \{r_1, r_2, ..., r_K\}$. In the E-step, the true ranking ξ is taken as the missing data. The expected value of the complete data log-likelihood with respect to missing data ξ, observed data R, and current expertise estimate α' is defined as follows:

$$Q(\alpha; \alpha') = E[\log p(R, \xi|, \alpha)|R, \alpha'] \tag{4}$$

The calculation of the expected value in above equation results in:

$$Q(\alpha; \alpha') = \sum_{\xi \in \mathbb{S}_N} L(\alpha)U(\alpha') \tag{5}$$

where $L(\alpha)$ is

$$L(\alpha) = \log \sum_{i=1}^{N} p(\xi^j) - N \log \sum_{i=1}^{K} Z(\alpha_i) + \sum_{j=1}^{N}\sum_{i=1}^{K} \alpha_i d(\xi^j, r_i^j) \tag{6}$$

where r_i^j is the ranking generated by the i-th ranker for the object candidate j, and $U(\alpha')$ is

$$U(\alpha') = \sum_{j=1}^{N} p(\xi^j|\alpha', r^j) \tag{7}$$

where r^j is the ranking generated by all rankers for the object candidate j.

In the M-step, Eq.(5) is maximized by α_i with the following derivation:

$$E_{\alpha_i}(d(\xi, r_i)) = \sum_{\xi \in \mathbb{S}_N} (\frac{1}{N}\sum_{j=1}^{N} d(\xi^j, r_i^j))p(\xi^j|\alpha', r^j) \tag{8}$$

For the *Kendall's tau distance*, $E_{\alpha_i}(d(\xi, r_i))$ is the expectation of Eq.(2) and can be expressed in the following form [Fligner and Verducci, 1986]:

$$E_{\alpha_i}(d(\xi, r_i)) = \frac{N\mathrm{e}^{\alpha_i}}{1-\mathrm{e}^{\alpha_i}} - \sum_{j=1}^{N} \frac{j\mathrm{e}^{\alpha_i j}}{1-\mathrm{e}^{\alpha_i j}} \tag{9}$$

At each iteration, a sampling method is introduced to obtain the approximate value of the RHS (right-hand side) of Eq. (8). Since $E_{\alpha_i}(d(\xi, r_i))$ is monotone decreasing, α_i can be easily obtained by a binary search approach.

2.2 Top-*k* Ranking Aggregation Model

As we can see, direct application of the above learning procedure is expensive, when the number of the candidates N is large. In tracking applications, it is reasonable that the groundtruth will be contained in the top-*k* candidates. To reduce to computational complexity, we do not need to do the rank aggregation on the whole list of the candidates.

Top-*k* lists are partial rankings indicating preferences over different (possibly, overlapping) subsets of *k* objects, where the elements not in the list are implicitly ranked below all of the list elements. Top-*k* ranking aggregation model means that the ranking aggregation is conducted on the top-*k* lists of all rankers. We find the above model can be easily extended to the top-*k* rank aggregation, the only difference is the $E_{\alpha_i}(.)$, which can be calculated as follows.

Definition 1: Let F_ξ and F_{r_i} be the top-*k* elements of ξ and r_i respectively, with $|F_\xi| = |F_{r_i}| = k$. $Z = F_\xi \cap F_{r_i}$ with $|Z| = z$. $P = F_\xi \setminus Z$, and $S = F_{r_i} \setminus Z$, with $|P| = |S| = k - z = l$.

Therefore,

$$E_{\alpha_i}(d(\xi, r_i)) = \frac{k\mathrm{e}^{\alpha_i}}{1-\mathrm{e}^{\alpha_i}} - \sum_{j=l+1}^{k} \frac{j\mathrm{e}^{\alpha_i j}}{1-\mathrm{e}^{\alpha_i j}} + \frac{l(l+1)}{2} - l(z+1)\frac{\mathrm{e}^{\alpha_i(z+1)}}{1-\mathrm{e}^{\alpha_i(z+1)}} \tag{10}$$

3 Multi-Modality Ranking Aggregation based Tracking Algorithm

Inspired by the above discussion, we propose a multi-modality ranking aggregation based tracking algorithm. In our work, we use three generative model based trackers and a discriminative model based tracker. The three generative model based trackers are: (1) the covariance matrix based tracker [Porikli *et al.*, 2006]; (2) the incremental subspace learning based tracker [Lim *et al.*, 2004]; (3) the spatial-color mixture of Gaussians based tracker [Wang *et al.*, 2007]. The discriminative model based tracker is the MIL (multiple instance learning) based tracker [Babenko *et al.*, 2011].

3.1 Algorithm Description

A flowchart of our algorithm is presented in Fig.2. The particle filtering tracking framework is adopted for all the trackers. In the prediction process, we first use the state transition to generate a set of candidate image patches corresponding to the object states $\{s_i\}_{i=1}^N$. In the tracking process, all the four trackers evaluate these candidate image patches according to their own appearance model, and output a set of ranking lists

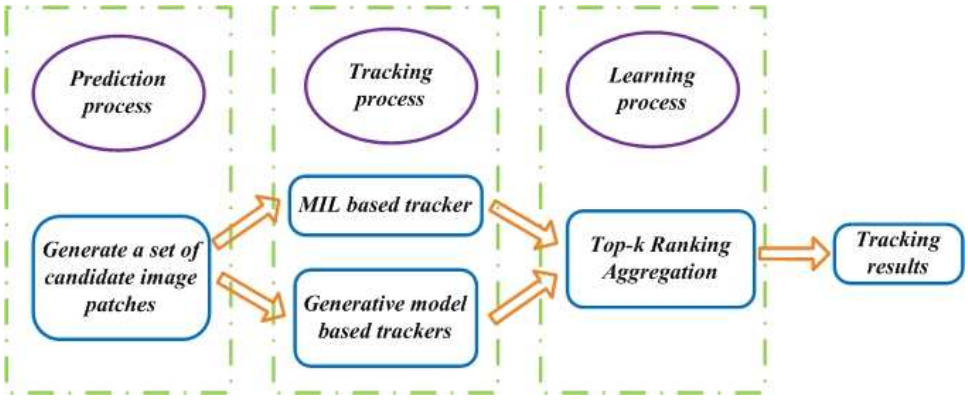


Figure 2: A flowchart of our proposed algorithm

based on their evaluation value. Then the top-k rankings are aggregated using the above model, and obtain the final ranking of the candidate image patches. The top-1 candidate image patch in the aggregated ranking is considered as the tracking result. All the four trackers are updated according to the aggregated ranking results respectively. In this way, the MIL tracker is supervised by the generative model based trackers and does not trained the classifiers using its own tracking result, therefore the *self-training* problem is avoided. Also in the instance selection process, even if the MIL tracker itself fails to locate the true object location during some frames, the other generative model based trackers may still help it select the correct positive bag.

3.2 Learning of the Ranker Expertise

The above ranking aggregation model obtains the expertise α_i of each ranker r_i using the EM algorithm in an iterative way. Then the true ranking ξ is deduced. It is an unsupervised ranking aggregation, and the expertise of each ranker is independent for different frames. However, it is not reasonable in tracking applications. In fact, the expertise of each ranker should meet the following two requirements: (1) the expertise of each ranker will vary little between two consecutive frames, and this consistency constraint should be considered; (2) the expertise of each ranker should reflect whether a ranker is a good tracker which needs to be adaptively evaluated in the tracking process.

After obtaining the aggregated ranking ξ, we can evaluate whether the ranker r_i is a good tracker by calculating the distance value $d(r_i, \xi)$. However, direct using this distance value for evaluation is not appropriate, for example, if r_i changes the top-2 candidates of ξ or the bottom-2 candidates of ξ, the distance value between them is the same. The former case means r_i do not give the same tracking results as ξ, while for the latter case, we think r_i give a very good tracking performance. To simplify this problem, we only focus on the top-1 candidate of the rankings. If the top-1 candidate of r_i and ξ is the same, we multiply the expertise α_i by a factor m ($m > 1$) which means r_i is reliable. While if the top-1 candidate of r_i and ξ is not the same, we multiply the expertise α_i by a factor m ($m < 1$). To take the historical ranking results and consistency constraint into consideration, the expertise parameter can be adaptively initialized in each frame as follow.

$$\alpha_i^t = w\alpha_i^{t-1} * m + (1-w)\alpha_i^{t-1} \tag{11}$$

where $w = 1 - \mathrm{e}^{-1/\sigma}$ acts as a forgotten factor and σ is a predefined constant. This warm start learning of the expertise α_i can accelerate the EM convergence process and make the results more accurate.

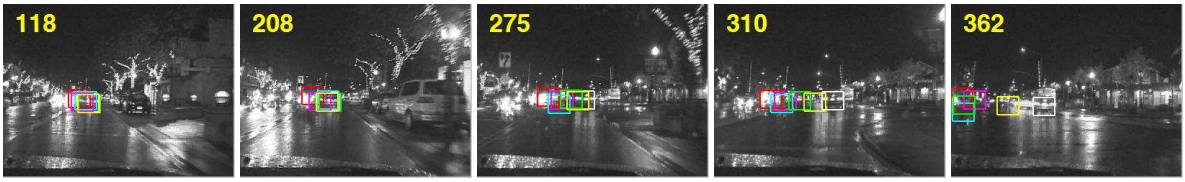
3.3 Implementation Detail

In this paper, the object is localized using a rectangular window and its state is $s_t = (t_x, t_y, w, h)$ where (t_x, t_y) denotes the center location of the bounding box and (w, h) are respective the width and height of the bounding box. Some important parameters are set as follows: $N = 600$, $k = 90$, $\alpha_i^0 = -1$, $m = 1.1$ or 0.9, $\sigma = 2$.

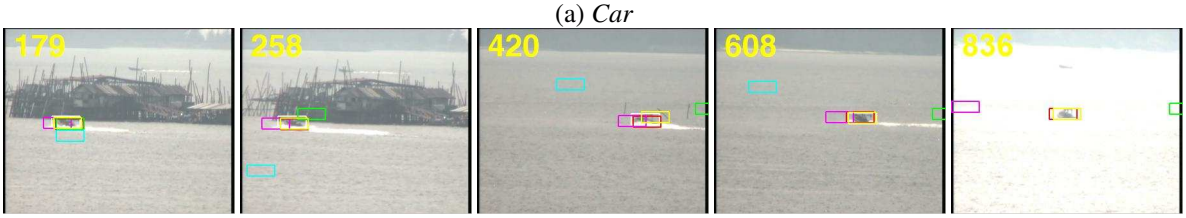
4 Experiment

In this section, we test the proposed multi-modality ranking aggregation based tracking algorithm (*MRAT*) on several challenging sequences. The difficulties of these sequences lie in that the background is noisy, the object undergoes large appearance changes, occlusion and the object is similar to the background. We compare our algorithm with five state-of-art algorithms. The first four algorithms are respectively multiple instance learning based tracker (*MIL*) [Babenko *et al.*, 2011], the L_1-regularized sparse template based tracker (*LRST*) [Mei and Ling, 2011], the visual decomposition tracker (*VDT*) [Kwon and Lee, 2010], and the online adaboost tracker (*OBT*) [Grabner *et al.*, 2006]. We also choose co-training multiple instance learning tracker (*CMIL*) [Lu *et al.*, 2011] as comparison[1]. In *CMIL*, it improves the *MIL* through combining HOG-classifier and RGB-classifier in a co-training way. While we improve the *MIL* tracker

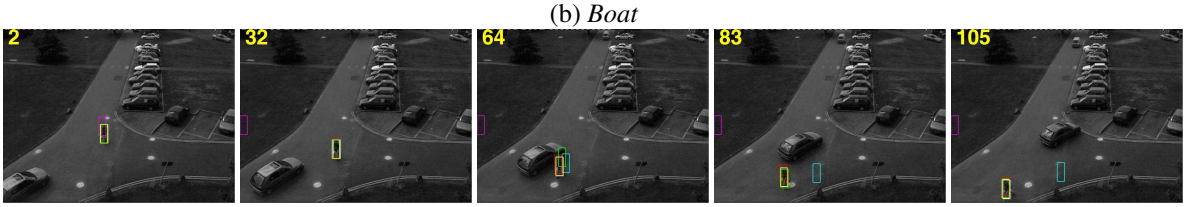
[1]The source codes of these trackers are available from:
http://vision.ucsd.edu/ bbabenko/project-miltrack.shtml
http://www.ist.temple.edu/ hbling/code/
http://cv.snu.ac.kr/research/ vtd/
http://www.vision.ee.ethz.ch/boostingTrackers/
http://ice.dlut.edu.cn/lu/Paper/FG2011/code/COMILcode.rar



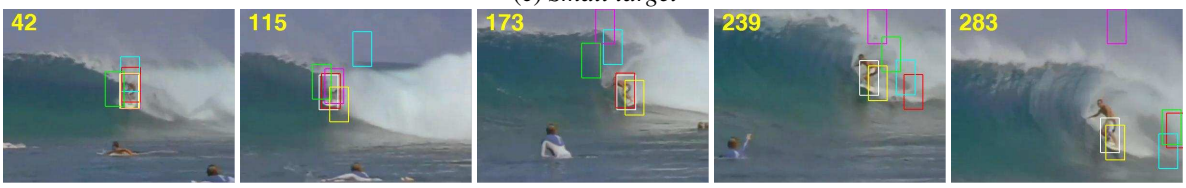
(a) *Car*



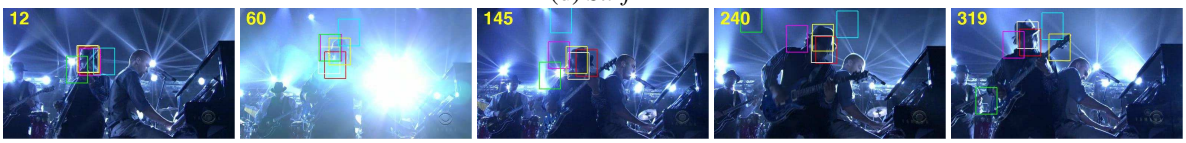
(b) *Boat*



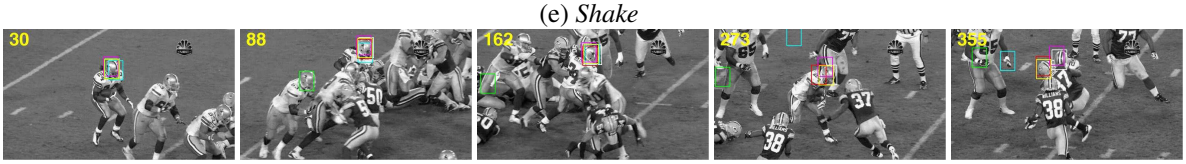
(c) *Small target*



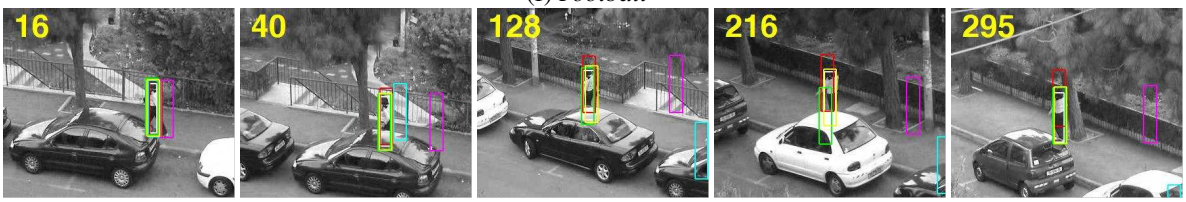
(d) *Surf*



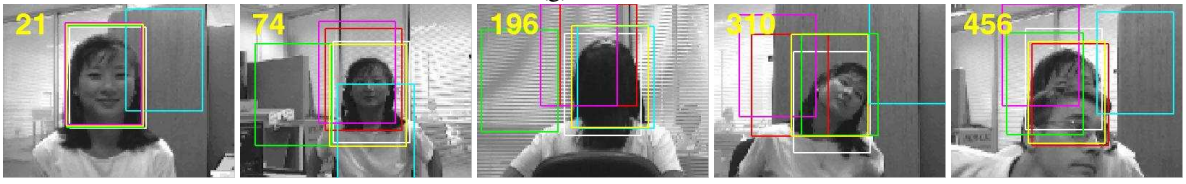
(e) *Shake*



(f) *Football*



(g) *Woman*



(h) *Girl*

Figure 3: The tracking results of §4.1.(Please refer to the PDF version and enlarge the figure for a clear view of the results)

through ranking aggregation instance selection, so it is interesting to compare these two algorithms. In the tracking results, *MRAT* is located using a white bounding box, *MIL* is located using a red bounding box, *LRST* is located using a green bounding box , *VDT* is located using a yellow bounding box, *OBT* is located using a green bounding box and *CMIL* is located using a light blue bounding box.

4.1 Qualitative Analysis

Car, Boat and SmallTarget

The *Car*, *Boat* and *Smalltarget* sequences are selected to test the performance of our algorithm against noisy and cluttered backgrounds. In the *Car* sequence, a car moves in a very noisy environment. The nearby background is so noisy that the car can not easily be located even by eyes. Some representative frames of the tracking results are shown in Fig.3(a). *MRAT* outperforms other trackers and successfully tracks all the frames. The other trackers fail to locate the object as the background becomes noisy.

In the *Boat* sequence, the surround background of the boat shares similar colors with it. Some representative frames of the tracking results are shown in Fig.3(b). *MRAT* and *MIL* are able to track the object. The *VDT* drifts away in some frames and can relocates the object. The other three trackers do not performs well in this sequence.

In the *Smalltarget* sequence, the object has a similar appearance to the background and there exists background distraction when the pedestrian passes by the car. Some representative frames of the tracking results are shown in Fig.3(c). *MRAT* and the *VDT* tracker can successfully track all the frames. The other trackers can not accurately locate the object when the background distraction exists.

Surf and Shake

The *Surf* and *Shake* are chosen to test the robustness of our algorithm against large appearance, illumination changes and cluttered backgrounds. In the *Surf* sequence, the surfer has a similar appearance with the wave and endures large appearance changes when he is covered by the wave. Some representative frames of the tracking results for this sequence are shown in Fig.3(d). *MRAT* achieves the best performance. It only fails to accurately locate the object as the scale of the object becomes so large that the bounding box cannot include the object. The other trackers cannot provide accurate results.

In the *Shake* sequence, the man changes his appearance by shaking his head. The background is cluttered and also undergoes large illumination changes. Some representative frames of the tracking results are shown in Fig.3(e). The *MIL* tracker successfully tracks all the frames. *MRAT* only fails to track the object around frame *frame* 60 and then quickly recovers to track the object. The other trackers do not perform well in this challenging sequence.

Football, Woman and Girl

The last three sequences are *Football*, *Woman* and *Girl*. We apply *MRAT* on these three sequences to test the robustness against occlusion and background distractions. In the *Football* sequence, severe occlusion happens and the target object shares a similar color with the other players. Some representative frames of the tracking results for this sequence are shown in Fig.3(f). The *MIL* and *VDT* tracker both begin to lose the object when severe occlusion and background distraction happen. *MRAT* successfully tracks the object in this challenging situation.

Alg.	*MIL*	*OBT*	*LRST*	*VDT*	*CMIL*	*MRAT*
Car	26.9	22.6	16.3	10.7	21.9	**3.8**
Boat	4.7	26.6	60.1	8.1	75.6	**4.2**
Surf	28.4	46.5	44.2	17.2	61.6	**10.1**
Smalltarget	6.2	196.1	7.8	3.5	32.1	**3.1**
Shake	**10.3**	13.2	52.2	13.0	63.3	11.1
Football	11.7	808	100.1	10.5	41.2	**4.4**
Woman	12.2	78.7	86.5	**4.5**	110.1	5.2
Girl	24.9	40.3	24.6	14.2	54.3	**10.8**

Table 1: Quantitative results for the sequences in §4.1

In the *Woman* sequence, the woman undergoes occlusion by the car. In the *Girl* sequence, the girl undergoes large appearance changes by turning around her head and is severely occluded by the man. Some representative frames of the tracking results for these two sequences are shown in Fig.3(g) and Fig.3(h). Both *MRAT* and *VDT* can successfully track the object through all the frames. The *MIL* tracker, it can track the object for most of frames but without accuracy. While the other trackers do not perform well in these two sequences.

4.2 Quantitative Analysis

In this part, we present quantitative evaluations to show the performance of trackers in different video sequences. We calculate the *RMSE* (root mean square error) of the four points in the bounding box between the tracking results and the groundtruth. The groundtruth is marked by hand. The mean of *RMSE* of the seven sequences we test on are given in Table 1. From the quantitative evaluations, *MRAT* provide the lowest *RMSE* in most of the sequences.

4.3 Discussion

The qualitative and quantitative results show that *MRAT* outperforms the other five competing trackers in most of the sequences. *MRAT* is robust in the presence of the background clutter and distraction, large appearance and illumination changes, and occlusion. It improves on the *MIL* through training the tracker under the supervision of the generative model based trackers in a ranking aggregation manner. In the sequences where the *MIL* successfully tracks the object, the proposed aggregation strategy does not worse the performance of *MIL* tracker. However, in the sequences that the *MIL* fails to track the object, *MRAT* improves the performance of the *MIL* tracker and can successfully track the object. As a result, we conclude that *MRAT* provides an effective solution to the problem of *self-training*.

5 Conclusion

In this paper, we propose a novel multi-modality ranking aggregation tracking algorithm to improve performance of the *MIL* under the supervision of the generative model. In our ranking aggregation framework, the discriminative classifier based *MIL* tracker is boosted with the help of the generative

models in ranking aggregation manner. The discriminative model and generative models are fused together seamlessly without the need to compare the performance of each other. The experimental results validate the effectiveness of our algorithm.

Acknowledgments

This work is supported by NSFC (Grant Nos. 61472285, 6141101224, 61473212, 61100147, 61203241 and 61305035), Zhejiang Provincial Natural Science Foundation (Grants Nos. LY12F03016, LY15F030011 and LQ13F030009), Project of science and technology plans of Zhejiang Province (Grants No. 2014C31062).

References

[Avidan, 2007] Shai Avidan. Ensemble tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):261–271, 2007.

[Babenko *et al.*, 2011] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1619–1632, 2011.

[Black and Jepson, 1998] Michael J. Black and Allan D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[Collins *et al.*, 2005] Robert T. Collins, Yanxi Liu, and Marius Leordeanu. Online selection of discriminative tracking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1631–1643, 2005.

[Comaniciu *et al.*, 2003] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[Fligner and Verducci, 1986] Michael A. Fligner and Joseph S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 359–369, 1986.

[Grabner *et al.*, 2006] Helmut Grabner, Michael Grabner, and Horst Bischof. Real-time tracking via on-line boosting. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 46–56, 2006.

[Hager and Belhumeur, 1998] Gregory D. Hager and Peter N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.

[Jepson *et al.*, 2003] Allan D. Jepson, David J. Fleet, and Thomas F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1296–1311, 2003.

[Klementiev *et al.*, 2008] Alexandre Klementiev, Dan Roth, and Kevin Small. Unsupervised rank aggregation with distance-based models. In *Proceedings of International Conference on Machine Learning*, pages 472–479. ACM, 2008.

[Kwon and Lee, 2010] Junseok Kwon and Kyoung Mu Lee. Visual tracking decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1269–1276. IEEE, 2010.

[Lebanon and Lafferty, 2002] Guy Lebanon and John Lafferty. Cranking: Combining rankings using conditional probability models on permutations. In *Proceedings of International Conference on Machine Learning*, volume 2, pages 363–370. Citeseer, 2002.

[Lim *et al.*, 2004] Jongwoo Lim, David A. Ross, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for visual tracking. In *Advances in neural information processing systems*, pages 793–800, 2004.

[Lin *et al.*, 2004] Ruei-Sung Lin, David A. Ross, Jongwoo Lim, and Ming-Hsuan Yang. Adaptive discriminative generative model and its applications. In *Advances in neural information processing systems*, pages 801–808, 2004.

[Lu *et al.*, 2011] Huchuan Lu, Qiuhong Zhou, Dong Wang, and Ruan Xiang. A co-training framework for visual tracking with multiple instance learning. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pages 539–544. IEEE, 2011.

[Mei and Ling, 2011] Xue Mei and Haibin Ling. Robust visual tracking and vehicle classification via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2259–2272, 2011.

[Porikli *et al.*, 2006] Fatih Porikli, Oncel Tuzel, and Peter Meer. Covariance tracking using model update based on lie algebra. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 728–735. IEEE, 2006.

[Saffari *et al.*, 2009] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. On-line random forests. In *Proceedings of International Conference on Computer Vision Workshops*, pages 1393–1400. IEEE, 2009.

[Stauffer and Grimson, 1999] Chris Stauffer and W Eric L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 1999.

[Wang *et al.*, 2007] Hanzi Wang, David Suter, Konrad Schindler, and Chunhua Shen. Adaptive object tracking based on an effective appearance filter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1661–1667, 2007.

[Wu *et al.*, 2013] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2411–2418. IEEE, 2013.

[Yu *et al.*, 2008] Qian Yu, Thang Ba Dinh, and Gérard Medioni. Online tracking and reacquisition using co-trained generative and discriminative trackers. In *Proceedings of European Conference on Computer Vision*, pages 678–691. Springer, 2008.

[Zhang *et al.*, 2010] Xiaoqin Zhang, Weiming Hu, Wei Qu, and Steve J. Maybank. Multiple object tracking via species-based particle swarm optimization. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(11):1590–1602, 2010.

[Zhang *et al.*, 2014] Xiaoqin Zhang, Weiming Hu, Shengyong Chen, and Steve J. Maybank. Graph-embedding-based learning for robust object tracking. *IEEE Transactions on Industrial Electronics*, 61(2):1072–1084, 2014.

[Zhou *et al.*, 2004] Shaohua Kevin Zhou, Rama Chellappa, and Baback Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 13(11):1491–1506, 2004.