Online Robust Visual Tracking via Superpixel-based Collaborative Appearance Model

Abstract

1 Introduction

Recent years have witnessed significant progress in visual tracking with the development the efficient algorithms. Most of recent algorithms for visual tracking in terms of three modules: target representation, search scheme and model update. Sparse representation have been widely applied in visual tracking [1][2][3][4][5], Mei and Ling[1] apply sparse representation in visual tracking, the target candidate is sparsely represented as a linear combination of the atoms of a dictionary which consists of dynamic target templates and trivial templates. And occlusion, corruption and other challenging issues can be deal with by trivial templates. In [3], a tracking algorithm based on histogram of local sparse representation is proposed, and the target object is located by a local representation based voting map and reconstruction error regularized mean-shift. In [4], the structural local sparse appearance model exploits both partial information and spatial information of the target based on a alignment-pooling method. In [5], a collaborative model which exploits both holistic templates and local representation is proposed.

Mid level visual cues also have been widely used in visual tracking, due to its effective representation with sufficient structure and great flexibility when compared with high level appearance models and low level features. Specifically, superpixels [9] have been one of the most effective and efficient image features, which have proved to be success in applications such as image segmentation [8], body model estimation and object recognition. In [7], a tracking method based by repeated figure/ground segmentation, which uses a superpixel_based CRF to match region. Wang et al. [6] propose a discriminative appearance model with structural information captured in superpixel. However, in[7], it processed the entire frame with CRF for region match which leads to high computational complexity. In[6]，as its feature dictionary is redundant, its also has high cost of computation. Further, its confidence map does not have consideration of distance of spatial. To overcome these problem, we proposed a tracking method with feature selection and clustering method with spatial information.

2 Superpixel-based Collaborative Appearance Model

In this section, we present the proposed algorithm in details. We first introduce the theory background of our algorithm, and then describe the clustering method with spatial information and superpixel-based Collaborative Appearance Model .The update scheme with feature selection of our appearance model is then present.

2.1 Bayesian Tracking Framework with Superpixel

Our algorithm is carried out with the Bayesian inference framework. Given the observation set of the target $z_{1:t} = \{z_1, \ldots, z_t\}$ up to the t-th frame, the target state variable x_t can be computed by the maximum a posterior estimation,

$$\widehat{x}_t = \arg\max_{x_t^i} p(x_t^i \left| z_{1:t} \right.) \qquad (1)$$

where x_t^i denotes the state of the i-th sample. The posterior probability $p(x \left| z_{1:t} \right.)$ is inferred by the Bayesian theorem recursively,

$$p(x_t \left| z_{1:t} \right.) \propto p(z_{1:t} \left| x_t \right.) \int p(x_t \left| x_{t-1} \right.) p(x_{t-1} \left| z_{1:t-1} \right.) dx_{t-1} \text{ , (2)}$$

where $p(x_t \left| x_{t-1} \right.)$ denotes the dynamic motion model and $p(z_t \left| x_t \right.)$ denotes the observation model .

The dynamic motion model $p(x_t \left| x_{t-1} \right.)$ describes the temporal correlation of the states between two consecutive frames. In the tracking framework, we apply the affine transformation with six parameters $x_t = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, t_x, t_y)$ to model the target between two consecutive frames, where $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ are the deformation parameters and (t_x, t_y) are the 2D translation parameters. The states transition distribution is formulated by a Gaussian distribution. We also assume the six parameters of the affine transformation are independent.

The observation model $p(z_t \left| x_t \right.)$ denotes the likelihood of the observation z_t at state x_t .It plays a key role in the object tracking. The observation model in our method is constructed by:

$$p(z_t \left| x_t \right.) \propto \widehat{C}(x_t) \qquad (3)$$

where $\widehat{C}(x_t)$ denotes the confidence of the observation z_t at state x_t . With the feature dictionary updated incrementally, the observation model is able to adapt to the appearance change of the target.

2.2 Efficient Superpixel Clustering Scheme

Given the image set of the target templates $T = [T_1, T_2, \ldots, T_n]$. For each template T_i , we first segment it into N_i superpixels. Each superpixel $sp(i,r)(i = 1, \ldots, n,\ r = 1, \ldots, N_i)$ is represented by a feature vector f_i^r , then we obtain a feature dictionary $D = \{ f_i^r \left| i = 1, \ldots, n,\ r = 1, \ldots, N_i \right. \}$. Next we apply the kmeans clustering algorithm on the feature dictionary D, and obtain m clusters, which denote Dic_cluster. Each Dic_cluster

$Dic\_clst(j)(j=1,\ldots,m)$ is represented by its cluster center $f_{Dic\_c}(j)$, its cluster radius $r_{Dic\_c}(j)$, and its own cluster members $\{f_i^r \mid f_i^r \in Dic\_clst(j)\}$.

Every Dic_cluster $Dic\_clst(j)$ has a score to evaluate how probable its superpixel members belong to the target or background, the higher the score, the higher the probability its members belonging to the target, the lower the score, the higher the probability its members belonging to the background. The score of every Dic_cluster contains two part: $S^+(i)$ and $S^-(i)$. The $S^+(i)$ denotes the number of pixels belonging to foreground, the $S^-(i)$ denotes the number of pixels belonging to background. Intuitively, the larger the $S^+(i)$, the more likely the superpixel members of $Dic\_clst(j)$ appear in the target area. So we evaluate the score of each Dic_cluster $Dic\_clst(j)$ by a confidence measure between 1 and -1 to indicate the probability its superpixel members belong to the target or background,

$$Dic\_C_j^c = \frac{S^+(i)-S^-(i)}{S^+(i)+S^-(i)} \qquad (4)$$

When a new frame arrives, we first exact a surrounding region of the target and segment it into N_t superpixels. If we direct using cluster algorithm to cluster N_t superpixels into m clusters according to Dictionary, it ignores the spatial information between superpixels. So we propose a bilateral kmeans clustering algorithm(BKC) to cluster the superpixels by taking both spatial information and intensity information into account. The BKC is constructed by two part: pre_cluster and post_cluster. In the pre_cluster, we first randomly select K superpixels as centers, then we calculate the spatial distance d_s and intensity distance d_r between each superpixel in the current frame and each center, if $d_r < \theta_r and\ d_s < \theta_s$, then this superpixel belongs to the center, we call it BKC_center. Each BKC_cluster $SKC(k)(k=1,\ldots K)$ is represent by its cluster center $f_{SKC\_c}(k)$,its cluster radius $r_{SKC\_c}(k)$ and its own cluster members $\{f_t^r \mid f_t^r \in SKC(k)\}$. In the post cluster, we then cluster $f_{SKC\_c}(k)$ into the Dic_cluster $Dic\_clst(j)$ with the smallest distance. By using our BKC algorithm, the superpixels with small spatial distance and intensity distance would be in the same cluster. However, the superpixels with small intensity distance but large spatial distance would be in the different cluster.

2.3 Superpixel-based Collaborative Appearance Model

When a new frame arrives, we have extracted a surrounding region of the target and segment it into N_t superpixels and cluster them into different clusters. A certain superpixel has its own members, which cluster it belongs to and the distance between this superpixel and its corresponding cluster center in the dictionary. Intuitively, the closer the superpixel lies to its corresponding cluster center, the more likely this superpixel belongs to. The confidence measure of each superpixel is computed as follows:

$$dis(r,c(i)) = \exp(-\sigma \times \frac{\left\| f_t^r - f_{Dic\_c}(i) \right\|}{r_{Dic\_c}(i)} \quad (5)$$

$$C_r^s = dis(r,c(i)) \times C_i^c \quad \forall r = 1,\ldots,N_t, i = 1,\ldots,n \quad (6)$$

where $dis(r,c(i))$ denotes the weighting term based on the distance between this superpixel f_t^r and its corresponding cluster center $f_{Dic\_c}(i)$ in the dictionary. $r_{Dic\_c}(i)$ denotes the cluster radius of $Dic\_clst(i)$ in the feature dictionary.

2.4 Robust Template Update

Since the appearance of the object often changes during the tracking process, the update scheme is essential and important. However, if we update the dictionary too often, the small errors are introduced each time. The errors are likely to accumulate and the tracker drifts from the target. If we do not update the dictionary, the tracker is prone to fail in dynamic scene due to the dictionary without capturing the appearance variation, because of the illumination and pose change. So we solve this problem by dynamically updating the template.

In many tracking methods, the earlier tracking results are more accurate so they should be stored longer than newly acquired results in the template stack. To balance the old and new templates, our templates consist of two parts: static templates and dynamic templates. The static templates are constructed by our training process, the first n tracking results (4 in our experiments), and the static templates remain the same in the entire tracking process. In order to capture the new appearance and adapt to the dynamic scene, we create the dynamic templates by adding the follow m tracking frames (6 in our experiment). For every W frames, we put the new frame into the dynamic templates and delete the oldest one.

Intuitively, the occlusion to pollute our template may not be desirable. So we present a simple but efficient method to handle occlusion problem. For the current state, the candidate with largest confidence value Cmax is what we need, and Cmean is the average of the confidence in the templates, Csum is the sum confidence in the template,so we set a threshold to detect the occlusion condition:

Occlusion= (Cmean – Cmax)/Csum>th

If occlusion is larger than th, it indicates that the confidence value of the current frame is much less than the average of the confidence in the templates, and it has high probability to the background, so we don’t update the dynamic template. By this template update scheme, our tracker can handle the occlusion much better, even more, when heavy occlusion happened, we can recover from it and robust results can be obtained.

4 Experimental Results

4.1 Quantitative Comparison

4.2 Qualitative Comparison

5 Conclusion

References

[1] X. Mei. H. b. Ling, Robust visual tracking using L1 minimization. In International Conference On Computer Vision, 2009.

[2] B. Chenglong , W. Yi , L. Haibin and J. Hui. Real time robust L1 tracker using accelerated proximal gradient approach.Providence. In International Conference On Computer Vision and Pattern Recognition, 1830-1837, 2012

[3] L. Baiyang , H. Junzhou , Y. Lin and C. Kulikowsk. Robust tracking using local sparse appearance model and K-selection.Providence. In International Conference On Computer Vision and Pattern Recognition, 1313-1320, 2011

[4] J. Xu , L. Huchuan and Y. Ming-Hsuan. Visual tracking via adaptive structural local sparse appearance model.In International Conference On Computer Vision and Pattern Recognition, pages 1822-1829, 2012

[5] Z. Wei , L. Huchuan and Y. Ming-Hsuan. Robust object tracking via sparsity-based collaborative model. In International Conference On Computer Vision and Pattern Recognition, 1838-1845, 2012

[6]W. Shu , L. Huchuan , Y. Fan and Y. Ming-Hsuan. Superpixel tracking.In International Conference on Computer Vision, pages 1323-1330, 2011.

[7] R. Xiaofeng and J. Malik. Tracking as Repeated Figure/Ground Segmentation. In International Conference On Computer Vision and Pattern Recognition, pages 1-8, 2007

[8] B. Fulkerson , A. Vedaldi and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In International Conference on Computer Vision, pages 670-677, 2009.

[9] R. Achanta , A. Shaji , K. Smith , A. Lucchi , P. Fua and X. Su, et al. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11): 2274-2282,2012