ROBUST VISUAL TRACKING VIA PART-BASED SPARSITY MODEL

Pingyang Dai, Yanlong Luo, Weisheng Liu, Cuihua Li, Yi Xie

Xiamen University
Computer Science Department
Xiamen, China

ABSTRACT

The sparse representation has been widely used in many areas including visual tracking. The part-based representation performs outstandingly by using non-holistic templates to against occlusion. This paper combined them and proposed a robust object tracking method using part-based sparsity model for tracking an object in a video sequence. In the proposed model, one object is represented by image patches. The candidates of these patches are sparsely represented in the space which is spanned by the patch templates and trivial templates. The part-based method takes the spatial information of each patch into consideration, where the vote maps of multiple patches are used. Furthermore, the update scheme keeps the representative templates of each part dynamically. Therefore, trackers can effectively deal with the changes of appearances and heavy occlusion. On various public benchmark videos, the abundant results of experiments demonstrate that the proposed tracking method outperforms many existing state-of-the-arts algorithms.

Index Terms— Visual tracking, part-based, sparsity model

1. INTRODUCTION

Visual tracking is one of the cardinal problems of computer vision. It has been widely used in many applications such as surveillance, driving assistant systems, interactive games, and augmented reality to human-computer interaction. But most of state-of-the-art tracking algorithms could not achieve satisfied requirements, let alone compare with human's performance. The challenges in designing a robust visual tracking algorithm are presented by noise, occlusion, varying viewpoints, clutter background and illumination changes. A variety of tracking algorithms have been proposed to solve these problems.

This research was supported by the National Defense Basic Scientific Research Program of China, National Defense Science and Technology Key laboratory Fund, Doctoral Program of Higher Specialized Research Fund (No. 20110121110020) and Fundamental Research Funds for the Central Universities of the People's Republic of China (No. 2010121066), Shenzhen City Special Fund for Strategic Emerging Industries (JCYJ2012).

The ensemble tracker [1] formulates tracking as a binary classification problem, where an ensemble of weak classifiers is trained online to distinguish objects from background. Grabner et al. [2] proposed an online boosting method to update discriminative features and a semi-supervised online boosting algorithm was proposed to handle the drift problem [3]. The method IVT [4] utilizes an incremental subspace model to adapt the changes of appearance. This method performs well when target objects encounter illumination changes and pose variations. But its holistic appearance model cannot effectively handle heavy occlusion. Babenko et al. [5] learned a classifier as the appearance model via multiple instance boosting. The classifier is online updated by means of a forgetting factor. Kalal et al. [6] took a different approach based on the P-N learning algorithm, in which the effective classifiers are learnt by exploiting the underlying structure of positive and negative samples for object tracking. Adam el al. [7] proposed a fragment-based method to handle occlusion. The target is located by a vote map which formed by comparing the histograms of candidate patches with the ones of corresponding template patches. However, the used template is not updated and the method is sensitive to the large variation of appearance. Recently, sparse representation has been used in the L1-tracker [8] where an object is modeled by a sparse linear combination of target templates and trivial templates. The template set is dynamically updated according to the similarity between the tracking result and the template set. However, occlusion is still one of the most challenging problems in object tracking.

In this paper, we develop an effective method which combines the part-based model and sparse representation for object appearance. This method handles partial occlusion and other challenging factors. Compared with the part-based model, our method maintains the local part appearance information and a vote map of multiple patches, which provides a compact representation of a tracked target. In addition, our tracker can be adaptively updated such that it keeps the representative templates of each part throughout the tracking process.

The rest of this paper is organized as follows. The detail of the part-based sparsity model is described in Section 2. Experimental results on public available challenging se-

quence are analyzed in Section 4. Finally we summarize our work in Section 4.

2. PROPOSED METHOD

In this section, we present the proposed method in details. We first discuss the motivation of this work. Next, we describe how the part-based method and sparse representation are combined, which forms a part-based sparsity model. The update scheme of our appearance method is then presented.

2.1. Problem formulation

Sparse representation has been extensively studied and successfully applied in pattern recognition and computer vision. With the constrain of sparsity, one signal can be represented in the form of linear combination of a few basis vector. Mei et al. [8] proposed an algorithm by casting the tracking problem as finding the most likely patch with sparse representation and handling partial occlusion with trivial templates respectively.

$$\mathbf{y} = \mathbf{T}\mathbf{a} + \epsilon = \begin{bmatrix}\mathbf{T}, & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{a}\\ \mathbf{e}\end{bmatrix} = \mathbf{B}\mathbf{c}. \tag{1}$$

In Equation (1), $\mathbf{y} \in \mathbf{R}^d$ is an image patch in the current frame, which is normalized to keep the same size as the template. The vector of trivial templates $\mathbf{I} = [i_1, i_2, \ldots, i_d] \in \mathbf{R}^{d\times d}$ is an identity matrix and $\mathbf{e} = [e_1, e_2, \ldots, e_d]^T \in \mathbf{R}^d$ is a trivial coefficient vector. Then each candidate of image patches is sparsely represented by a set of target templates and trivial templates. Equation (1) can be solved as a $\ell 1$-regularized least squares problem, which is known as a typically yield sparse solution.

$$min\|\mathbf{B}\mathbf{c} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{c}\|_1. \tag{2}$$

In Equation (2), $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the norms of $\ell 1$ and $\ell 2$ respectively. Therefore, the tracking with sparse representation is formulated as a problem that searchs for samples with minimal reconstruction errors from the space of learned templates. The underlying assumptions of this approach can be used to handle partial occlusion and noise by modeling the error *e* as arbitrary sparse noise. However, several problems still exist. For example, the background pixels in the target templates do not lie on the linear template subspace. The reconstruction error from background pixels might be smaller than the one from target pixels under the conditions of occlusion or noise, which would affect the accuracy of sparse representation. Moreover, the spatial relation among adjacent patches and the temporal correlation among target templates have not been considered.

The usage of parts or components is well known in object recognition. Deformable part models known as pictorial structures provide an elegant framework for object detection. In tracking, the part-based model often works well due to its simple models such as rigid templates. Adam et al. [7] proposed an approach which combines fragment-based representation and voting. The tracking algorithm is robust to partial occlusion and pose variation. But it does not include any update scheme and cannot handle the large pose variations and clutter background.

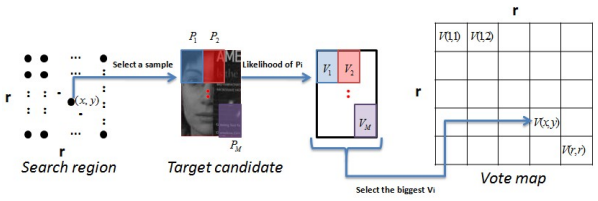

Fig. 1. Main components of the proposed method.

Therefore, we develop a part-based sparsity model that integrates part-based model and sparse representation with a template update scheme. The part-based model with update scheme can handle partial occlusions and pose changes. At the same time, reconstruction error in sparse representation can be used to measure the similarity with according patch. Therefore, the proposed method can handle various challenges. Fig.1 illustrates how the vote values can be calculated from the patches in target candidate and be combined into a vote map. More details describes in Section 2.2.

The contributions of this paper are summarized as follows. Firstly, a unified part-based sparsity framework which combines part-based model and $\ell 1$ regularization is proposed. Secondly, the update template scheme combines with the part-based model and sparse representation are employed to handle pose variation, appearance change and heavy occlusion. Experiments on several challenging benchmark image sequences demonstrate that our proposed tracking method outperforms many existing state-of-the-arts algorithms.

2.2. Part-based sparsity model

Given the initial frame I, we use vertical and horizontal patches to construct an appearance model for a target, shown in Fig.1. Each patch $p_i, i = 1, \ldots, M$ is non-overlapped and represents one fixed part of the target object. Therefore, local patches can represent the complete structure of the target. We generate template set $\mathbf{T}^i = [T_1^i, T_2^i, \ldots, T_n^i]$ for each patch p_i by using zero-mean-unit normalization and perturbing one pixel in four directions of the original pixels, which is similar to [8]. Each patch has a dictionary $\mathbf{B}_i = [\mathbf{T}^i\ \mathbf{I}]$. For a target candidate, we accordingly extract the local patches from target candidates which can be represented as linear combination of only a few basis elements of the dictionary. These local patches can be solved by a $\ell 1$-regularized least squares problem as below.

$$min\|\mathbf{y}_i - \mathbf{B}\mathbf{c}_i\|_2^2 + \lambda\|\mathbf{c}_i\|,\ s.t.\ \mathbf{c}_i \geq 0. \tag{3}$$

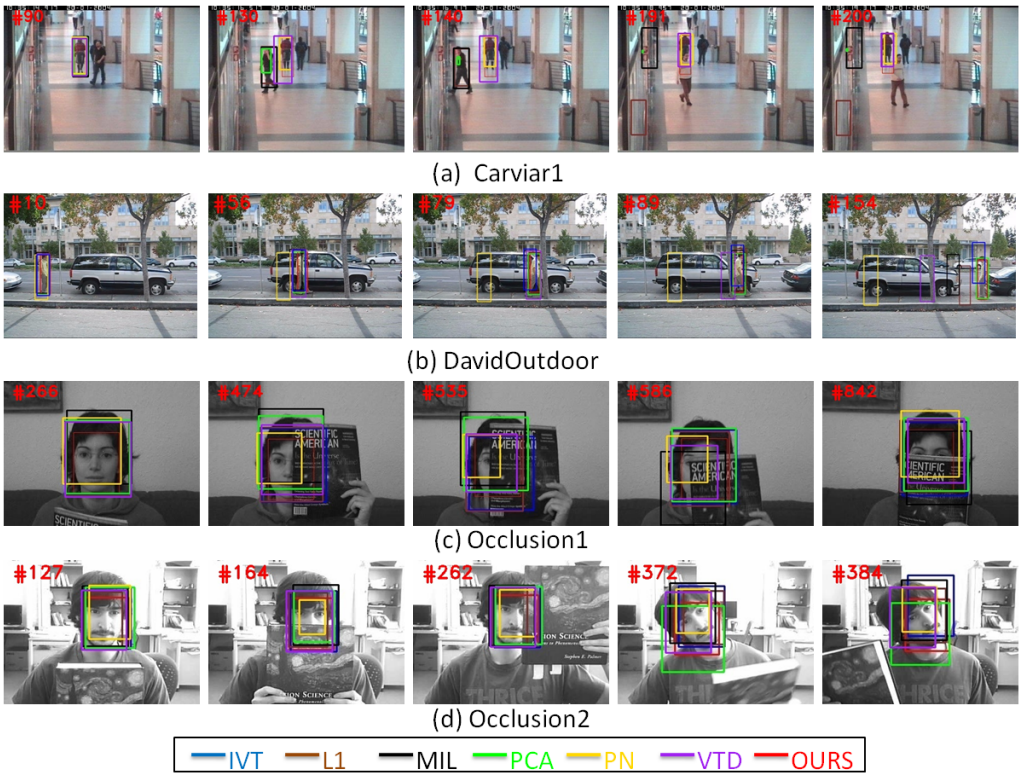

Fig. 2. Qualitative comparison results.

We can calculate the reconstruction error for each patch p_i by $\mathbf{e}_i = \|\mathbf{y} - \mathbf{T}\mathbf{c}_i\|$ and calculate the vote value v_i for the i-th patch. After obtaining v_i, we obtain a vote map $V(\cdot,\cdot)$ for every possible position (x, y) of the target in current frame. Then we choose the maximal value in the vote map $V(\cdot,\cdot)$ by $v(x, y) = max\{v_1, v_2, \ldots, v_M\}$ and obtain the location of the target in the current frame.

This appearance model is able to deal with partial occlusion. When occlusion occurs, the change of appearance increases the reconstruction error of the occluded local patches. However, the local patches which are not occluded still have lower reconstruction errors and then have high vote values in the vote map. Therefore, this model can accurately track targets and handle noise, heavy occlusion and appearance change.

2.3. Template update

It is essential to update the observation model for handling the appearance change of a target object in visual tracking. Since the appearance of an object often changes significantly due to partial occlusion, illumination and pose variation during the tracking process. Tracking with fixed object templates is prone to fail in such dynamic scenes. However, if we update the templates too frequently according to new observations, the errors are likely accumulated and the tracker is susceptible to drift. We handle this problem by dynamically updating the target templates set $\mathbf{T}^i$ for each patch. To dynamically update the target templates set $\mathbf{T}$, we use the weights of importance for the templates which are similar to the ones in [8]. One important feature of $\ell 1$ minimization is that it favors templates with large norms because of the regularization part $\|\mathbf{c}\|_1$. The larger the norm of T_j^i according to patch p_i is, the smaller coefficient c_j^i is needed in the approximation $\|\mathbf{y} - \mathbf{T}\mathbf{c}_i\|_2$. We can introduce a weight $\omega_i = \exp(-\|\mathbf{y} - \mathbf{T}c_j^i\|_2^2)$. But in order to make the best of c_j^i, we compute the reconstruction error of each observed image patch by Equation (4).

$$\omega_i = \frac{max(c_i)}{\|y_i - B_i c_j^i\|}. \tag{4}$$

We create each patch template at the first frame. And the templates set $\mathbf{T}_i$ is updated with respect to the coefficients of tracking result respectively.

3. EXPERIMENTS

In order to evaluate the performance of the proposed tracking method, we compare our tracking algorithm with 6 state-of-the-arts methods on 4 challenging sequences which are

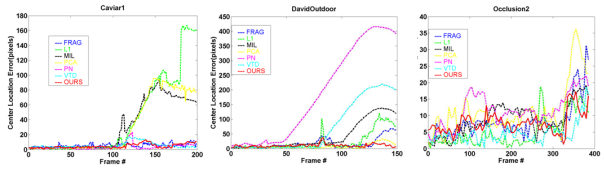

Fig. 3. Qualitative comparison results.

publicly available from prior work [5] and Caviar dataset (http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/). The tracking methods for comparision include incremental visual tracking (PCA) method [4], fragment-based (FragTracker) tracking method [7], L1 tracker [8], multiple instance learning (MIL) tracker [5], visual tracking decomposition (VTD) method [9] and P-N learning(PN) tracker [6]. To be fair, we use the source codes provided by the authors with the tuned parameters for their best performance. Our proposed algorithm is implemented in C++. We extract non-overlapped 3×3 patchers from the target image and resize each patch to 20×20 pixels. For each sequence, the location of the target object is manually labeled in the first frame.

The first experiment uses *caviar*1 sequence. The walking person is subjected to partial occlusion. Some tracking result frames are given in Fig.2. The frame indexes are 90, 130, 140, 191 and 200. It can be observed that other trackers start tracking the man when the woman is partially occluded at frame 130 and 191 except PN, VTD and our method. And our method locates the target well. Compared with other trackers, our tracker is more robust to the partial occlusion.

The outdoor environment is very challenging for visual tracking since tracking object may suffer from occlusion, pose changes and clutter background. The second experiment uses *Davidoutdoor* Sequence. The indexes of five representative frames are 10, 56, 79, 89 and154. All trackers except PCA and our method lose the target. Our tracker tracks the target very well throughout the whole sequence.

In the third and fourth experiments, the sequence occlusion1 and occlusion2 involve severe occlusion which is very challenging for tracking. Their indexes of representative frames are 266, 474, 535, 586, 842 and 127, 164, 262, 372, 384 respectively. As shown in Fig.2, our tracker achieves better tracking results than other trackers.

The results of quantitative comparisons also verify that our tracker is superior to the above-mentioned algorithms. The relative center position error (in pixels) between the ground truth center and the tracking results is presented in Fig.3. Our tracker consistently produces a smaller distance error than other trackers in most of cases.

4. CONCLUSION

In this paper, we have proposed an effective and robust tracking method based on the part-based sparsity model. In our trackers, the part-based model and sparse representations are combined for object appearance model, where $\ell 1$ regularization is introduced into the part-based model. The update scheme reduces drifts and enhances the capability against appearance occlusion. We have provided quantitative and qualitative comparisons among our proposed method and six state-of-the-arts algorithms. On several challenging image sequences, our tracker demonstrates better performance in terms of robustness.

5. REFERENCES

[1] S. Avidan, "Ensemble tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 2, Feb.

[2] H. Grabner and H. Bischof, "On-line boosting and vision," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, June, vol. 1.

[3] Helmut Grabner, Christian Leistner, and Horst Bischof, "Semi-supervised on-line boosting for robust tracking," in *ECCV (1)*, 2008.

[4] David A. Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, 2008.

[5] B. Babenko, Ming-Hsuan Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, Aug.

[6] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, July.

[7] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, June, vol. 1.

[8] Xue Mei and Haibin Ling, "Robust visual tracking and vehicle classification via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 11, Nov.

[9] Junseok Kwon and Kyoung-Mu Lee, "Visual tracking decomposition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June.