Part-Based Tracking with Appearance Learning and Structural Constrains

Wei Xiang and Yue Zhou

Institute of Image Processing and Pattern Recognition
Shanghai Jiao Tong University
Shanghai, China
{xwsjtu,yue_zhou}@126.com

Abstract. Adaptive tracking-by-detection methods are widely used in computer vision for tracking objects. Despite these methods achieve promising results, deformable targets and partial occlusions continue to represent key problem in visual tracking. In this paper, we propose a part-based visual tracking method. First, we take advantage of the existing online learning appearance model to learning the appearance of each part. Second, we propose a novel part initialization method and an affine invariant structural constrain between these parts. Third, a tracking model based on the appearance of each part and the spatial relationship between the parts is proposed. We make use of an optimization algorithm to find the best parts during tracking, update the appearance model and the structural constraints between parts simultaneously. In this paper we show our method has many advantages over the pure appearance learning based tracking model. Our method can effective solve the partial occlusion problem, and relieve the drift problems. What's more, our method achieves great result while tracking the target of which geometric appearance changes drastically over time.

Keywords: tracking, appearance model, structural constrains, part-based.

1 Introduction

Object tracking has many practical applications and has long been studied in computer vision. An approach to tracking which has become particularly popular recently is tracking-by-detection [1, 2, 3, 4, 7, 11]. It has been shown that an adaptive appearance model, which updating the model during the tracking process, is the key to good performance.

These methods train a model to separate the object from the background via a discriminative classifier can often achieve superior results. These methods often don't consider the motion model and have been termed "tracking by detection". For example, Avidan[4] used a Support Vector Machine as an off-line binary classifier to distinguish target from background. Helmut Grabner[3] used an on-line boosting method to choose discriminating features and classify the target and background. Babenko et al.[5]employ an online Multiple Instance Learning based appearance

C.K. Loo et al. (Eds.): ICONIP 2014, Part I, LNCS 8834, pp. 594–601, 2014.

model to resolve the sample ambiguity problem. All of these method delineate the tracked object by a single regular bounding box(e.g., a rectangular or circular window), which renders them sensitive to partial occlusions and shape deformations.

Deformable part-based models have been successfully applied to object detection and object recognition on numerous occasions. Part-based appearance models [6, 8, 10] have been shown to have favourable properties such as robustness to partial occlusions and articulation. In those cases, highly variable objects are represented using mixtures of deformable part-based models. But, some difficulties in extending part-based appearance models to visual tracking are still remain to solve. Firstly, it is hard to decide which parts are the "good" parts, which means these parts can well distinguish the object from background. Secondly, because the appearance model needs to be updated online, we have to adjust parts dynamically, reducing or adding parts according to the part deformation. Thirdly, when encountering partial occlusions, these parts need to be updating dynamically.

In this paper, we proposed a novel tracking method, which is based on learning the appearance of parts of the object and the spatial relationship between these parts. Our method adopts an effective appearance model to learning the appearance of each part. At the same time, we propose a novel structure description method to show spatial relationships between the parts. Our final tracking model can well handle the partial occasion and strong deformation problem.

In summary, our main contributions are: (1) we propose a novel structure description method to show spatial relationships between the parts (2) we propose a tracking model which makes full use of the appearance of each part and the spatial relationships between these parts. The appearance of each part and the structural constrains update online. We discuss the appearance model in section 2.Section 3 introduce our new tracking model, section 4 presents our experimental results, and section 5 concludes the paper.

2 Appearance Model

In this paper, we choose the compressive based features extracted appearance model [1] as our part learning model. The appearance model is generative as the object can be well represented based on the features extracted in the compressive domain. It is also discriminative because it uses these features to separate the target from the surrounding background via a naive Bayes classifier. In this appearance model, features are selected by an information-preserving and non-adaptive dimensionality reduction from the multi-scale image feature space based on compressive sensing theories.

This method assumes that the tracking window in the first frame has been determined. At each frame, sample some positive samples near the current target location and negative samples far away from the object center to update the classifier. To predict the object location in the next frame, drawing some samples around the current target location and determine the one with the maximal classification score. The fig1 illustrates the appearance model.

The appearance model is efficient and robust, but its tracking box is kept unchanged. We use the method as our part appearance model and propose the structural constrain among these parts, which make our method is robust to deformable targets.

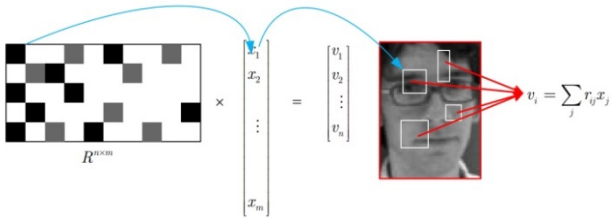

Fig. 1. The illustration of the compressive based features extracted appearance model .The blue arrows illustrate that one of nonzero entries of one row of R sensing an element in x is equivalent to a rectangle filter convolving the intensity at a fixed position of an input image.

3 Parts Based Tracking

3.1 Part Initialization

Our algorithm is based on part learning. The initial position of parts has to be chosen so as to be good for image representation. It is obvious that the part which contains affluent target character information should be chosen. We propose a novel method to select the effective parts. Firstly, we use the FAST algorithm to find key points. FAST(features from accelerated segment test) is an efficient key point detecting algorithm. We assume that the area where these key points aggregating contain abundant information and should be regard as one part. Therefore, we use the K-means clustering algorithm to classify these key points into several categories. For each category, we exclude the outlier points and use a rectangle to encompass these points. These rectangles are the parts we chosen. Fig.2 show the illustration of initialization of parts. In experiment, we found that 3 or 4 parts usually is a good option.

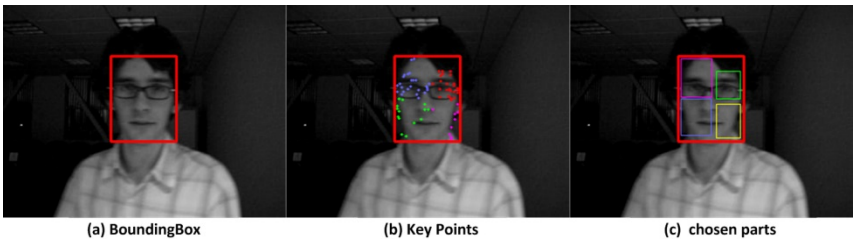

Fig. 2. Example of initialization of parts in David seq. (b) displays key points detected by the FAST algorithm. (c) shows the four initialized parts.

3.2 Structure Description

Most patch-based algorithms use the distance between patch center and target center as structure description. The description is rotation-invariant. However, it cannot resist scale variance. In this paper, we propose a new method to descript the structure constrain as shown in Fig3. The position of each part is defined by $X_t = (X_t^c, X_t^1, ..., X_t^m, D_t^a, D_t^1, ..., D_t^m)$ where X_t^c denotes the center position of an object, X_t^i indicates the center position of the ith local part, we define the spatial relationship between these parts.

$$D_i = R_i / (\sum_{j=1}^{n} R_j / n), i = 1, 2, ..., n \tag{1}$$

Where R_i is the distance between the center of ith part and target center. The description is affine invariant, Eqn.(2)proves it can resist scale variance.

$$aR_i / (\sum_{j=1}^{n} aR_j / n) = aR_i / (a\sum_{j=1}^{n} R_j / n) = R_i / (\sum_{j=1}^{n} R_j / n) = R_i / R_{avr} \tag{2}$$

These Distances describe the structural constraints between these parts. When the target suffers rotating, scaling, translation and changes in view angle, these distances should stay the same.

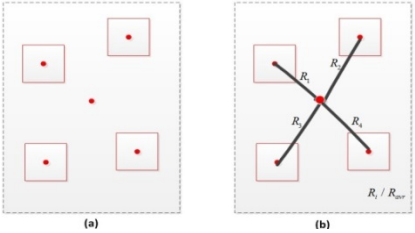

Fig. 3. Description of the structure of the target.(a) the initial part (b) illustrates the structure description of a target.

3.3 Tracking

Our tracking algorithm is based on the appearance model of each part and the structural constraints between these parts. For each part, we train the appearance model to online learning the appearance of each part. Assume four parts, P_1, P_2, P_3, P_4, we construct four trackers, T_1, T_2, T_3, T_4 based on our appearance model. During tracking, each tracker can get some candidate rectangles of each part, $L_i^1, ..., L_i^n$ with scores $S_i^1, ..., S_i^n$. These scores indicate the possibility of the rectangle to be the corresponding part. The higher score a rectangle gets, the higher possibility the rectangle can be the corresponding target. Since each contains partial information of the target, it may

drift easily. We use the structural constrains between these parts to solve that problem, which lead to our final tracking model. This can be expressed by the following optimization problem.

$$\arg\max \sum_{i=1}^{n} w_i + \lambda \sum_{i=1}^{n} f(p_i) \tag{3}$$

While w_i represents the appearance similarity of each parts and $f(p_i)$ represents the structure term among each part, λ control the relative weight between the appearance term and structure term. To solve the optimization problem, we can get the appearance score of each part through our appearance model, but the structure model is hard to determine. We don't know the center of our tracking object, which should be determined by the parts. At the same time, the location of each part is still unknown. Since the problem is unsolvable, we have to find the approximate solution. We consider the fact that the variance between two serial frames in video is relative small. Thus the target is impossible to change drastically. We use the template matching method to roughly locate the object and get the center of the object. Then we can solve the Eq3 to precise positioning the location of each part. Finally, we use the minimum enclosing rectangle method to get the final tracking result. Update the appearance model of each part and the structural constrains between these parts as well.

In all, tracking procedure mainly includes three parts: (a) Using the appearance model to locate every candidate parts and get the similarity. (b) Template matching method to get the center. (c) Optimizing the Eq.2 to precisely locate each part. And finally, we use these parts to get the position of the target. The main steps of our algorithm are summarized in Algorithm1.

Algorithm 1. Part-based learning and structural constrains tracking

Input: ith video frame, the structural constrain of object in previous frame

1. Use the each appearance model of part (P_1, P_2, P_3, P_4) to sample rectangles, and Calculate the likelihood score of these rectangles ($W_1^{1,\ldots,n}, W_2^{1,\ldots,n}, W_3^{1,\ldots,n}, W_4^{1,\ldots,n}$).
2. Maximize the equation 3 to find the best parts (P_1, P_2, P_3, P_4) .
3. Update the appearance model of each part and the spatial relationship between these parts.
4. Get the final tracking result through the location of each part.

Output: Tracking location L_t and the structural constrain.

3.4 Discussion

Our tracking algorithm mainly has two advantages over some other algorithms. Firstly, it can well handle the situation that the target suffers dramatic deformation. For many tracking methods based on appearance learning, when the target or the

background changes greatly, the tracking result may contain background information which may lead to the model drifting gradual. For our algorithm, due to the fact that each part contains the partial information of target object, and the structural term express the global information of the target. When the appearance of the target changes dramatically, each part can find the corresponding part. At the same time, because the structure constrain, these parts are unlikely to drift. As shown in Fig4, our tracking algorithm perform well on the ice skater video.

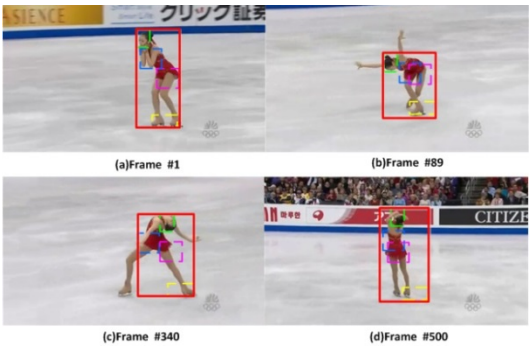

Fig. 4. Tracking result of ice skater

Another big advantage of our algorithm is that it can well handle the partial Occlusion problem. Each part is learned by our appearance model, and we can get some candidate rectangles with corresponding scores. When a part is occluded while other parts not, we can know this happen due to the fact that the average score of this part is much less than the other parts get. So we can dynamically reduce the learning rate of occluded part.Fig4.

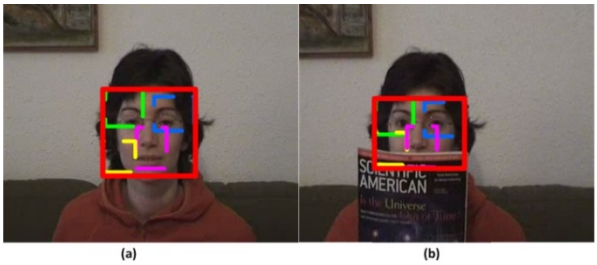

Fig. 5. (a)The target (b) The face is partial occupied by a book. In our algorithm, we can detect the yellow rectangle and the purple rectangle is occupied by comparing the average scores they get.

4 Experiment

In experiments, we will verify the performance of our algorithm. Development environment of our algorithm is Visual Studio 2010 and Intel OpenCV library on Intel

Core(TM)2 Duo CPU E7500 and 3.00GB RAM. We evaluate our tracking algorithm with the online AdaBoost method and the compressive tracking algorithm on 4 challenging sequences.

For the David indoor sequence shown in Fig 6(1), the illumination and pose of the object both change gradually. Our method performs well but the OAB method failed and the CT approach can't locate accurately. For the Singer sequence shown in Fig 6(2) the illumination and the scale changes drastically, the proposed tracker is robust to scale and illumination change. Our method has the structural constrain term, so it can well handle this situation. For Fig 6(3), the pose of the object changes drastically, our method performs well while the other methods fail to track. For Fig 6(4), the background and the object both change drastically, our method achieved great result while the other method failed. Fig5 has demonstrated our method can well handle the partial occlusion well. Our method is very efficient, which runs at 40 frames per second (FPS) on our development environment.

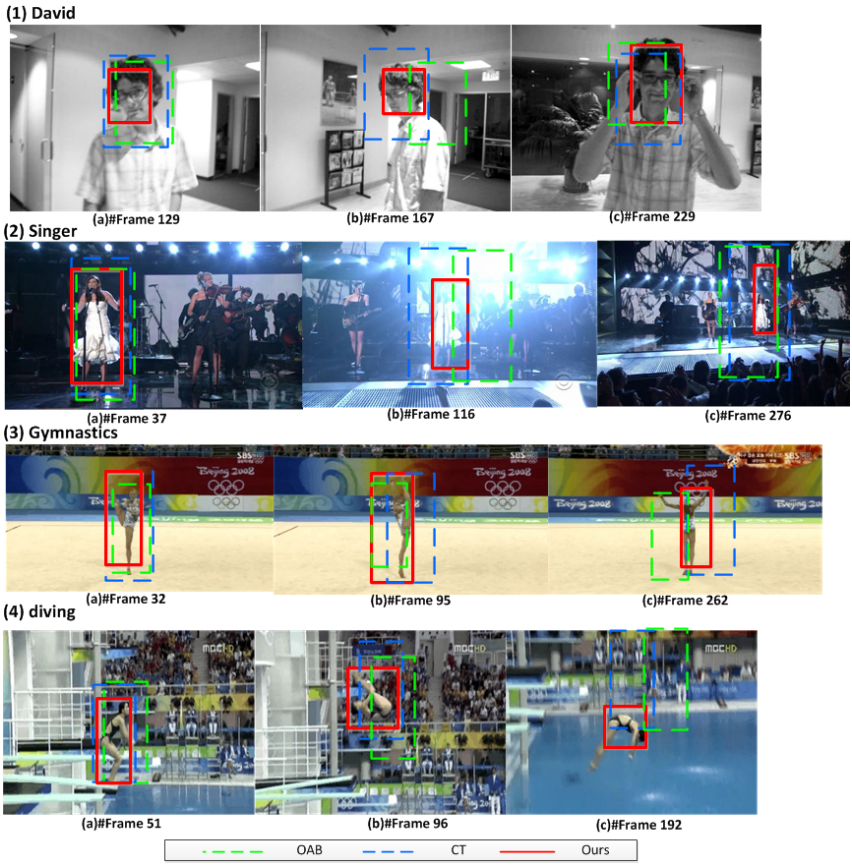

Fig. 6. Screenshots of some sampled tracking results

5 Conclusions

In this paper, we presented a novel approach to tracking based on part learning and structural constrains between these parts. Our formulation generalizes previous methods by combining the appearance learning model and the spatial relationships among parts. The appearance of parts and their spatial relationship is updating online. In contrast to other based on appearance learning model, our algorithm can well handle large variation in pose. Our approach models an object as a set of parts, each parts is trained and updated individually using the compressive based features extracted appearance model. Our algorithm can well solve the partial occlusion problem via the parts learning method. The Structural constraints between parts are rotation-invariant. So our approach can well handle the rotation and scale variation problem. When the object is very small or the resolution of target is very low, each part contains little information and may drift easily. So our approach does a relatively poor job on that situation.

References

1. Zhang, K., Zhang, L., Yang, M.-H.: Real-time compressive tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 864–877. Springer, Heidelberg (2012)
2. Yao, R., et al.: Part-based visual tracking with online latent structural learning. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2013)
3. Grabner, H., Grabner, M., Bischof, H.: Real-Time Tracking via On-line Boosting. BMVC 1(5) (2006)
4. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: 2011 IEEE International Conference on Computer Vision (ICCV). IEEE (2011)
5. Babenko, B., Yang, M.-H., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009. IEEE (2009)
6. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008. IEEE (2008)
7. Li, H., Shen, C., Shi, Q.: Real-time visual tracking using compressive sensing. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2011)
8. Hare, S., Saffari, A., Torr, P.H.S.: Efficient online structured output learning for keypoint-based object tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2012)
9. Wu, Y., Lim, J., Yang, M.-H.: Online object tracking: A benchmark. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2013)
10. Kwon, J., Lee, K.M.: Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009. IEEE (2009)
11. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(7), 1409–1422 (2012)