IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Issue-11 E-ISSN: 2347-2693

Object Tracking Based on Sparse Discriminative and Generative Model

Renuka Devi SM

Department of Electronic and Communication, GNITS, Hyderabad, India

e-mail: renuka.devi.sm@gmail.com

Available online at: www.ijcseonline.org

Received: Oct/26/2016 Revised: Nov/02/2016 Accepted: Nov/24/2016 Published: Nov/30/2016

***Abstract*— Real time object tracking is a challenging task in computer vision. Many algorithms exist in literature like mean shift method, kernel method , pixel based, Silhouette based and sparity based, method. Of these methods robust appearance model that exploits both holistic templates and local representations is the sparsity-based discriminative classifier (SDC) and a sparsity-based generative model (SGM). SDC module, is effective method to compute the confidence value that assigns more weights to the foreground than the background in the SGM module. Further the histogram-based method is also discussed that takes the spatial information of each patch into consideration with an occlusion handing scheme. Furthermore, the update scheme considers both the latest observations and the original template, thereby enabling the tracker to deal with appearance change effectively. Experimental results show that the above method gives good performance and accuracy even in the presence of occlusion.**

Keywords- Object tracking, Target feature modelling, sparsity-based generative model, sparsity-based discriminative classifier

I. INTRODUCTION

Object tracking is defined as the process of segmenting an object of interest from a video scene and keeping track of its motion, orientation even in presence of occlusion and noise. The first step of tracking involves extraction detection of the moving objects in video streams. Next steps are the tracking of such detected objects from frame to frame and analyse the object tracks, to recognize their behavior. The goal of object tracking is to estimate the states of a target object in an image sequence.

Much of the literature focuses on meanshift algorithm, kalman filtering, pixel based, Silhouette based and sparsity based and various other methods of object tracking in a video sequence[1]. Extensive literature work is discussed in [2][3].An image, usually from a video sequence, is divided into two complimentary sets of pixels. The first set contains the pixels which correspond to foreground objects while the second complimentary set contains the background pixels. This result is often represented as a binary image or as a mask. It is difficult to specify an absolute standard with respect to what should be identified as foreground and what should be marked as background because this definition is somewhat application specific [4][5]. Generally, foreground objects are moving objects like people, boats and cars and everything else is background. Many a times shadow is represented as foreground object which gives improper output.. The basic steps for tracking an object are Object Detection, Object Representation, Object Tracking

In this paper the appearance model[6][7] is focused since it is usually the most crucial component of any tracking algorithm. This plays a critical role in numerous vision applications.

The paper is organized as follows: Section 2 deals with the related work in object tracking, Section 3 presents in detail the object tracking by sparse collaborative model. Section 4 gives the experiment results for three cases of video. Finally Section 5 presents the conclusion

II. RELATED WORK

The section gives the overall steps in object tracking algorithm. As seen basically tracking involves[2]:

A. Object Detection Methods:

Every tracking method requires an object detection mechanism either in every frame the various methods of object detection are

a. Temporal differencing method
b. Frame Differencing
a. Optical Flow
b. .Background Subtraction

The comparison of each of the methods are given in [2]

B. Object Representation Methods:

In a tracking scenario, Objects can be represented by their shapes and appearances. The extracted moving object may be different objects such as humans, vehicles, birds, floating clouds, swaying tree and other moving objects. Hence shape

features are usually used to represent motion regions. As per literature survey, approaches to represent the objects are as follows:

a. Shape-based Representation
b. Motion-based Representation
c. Color-based Representation
d. Texture-based Representation

C. Object Tracking Methods:

Tracking can be defined as the problem of approximating the path of an object in the image plane as it moves in a scene. Object tracking can be classified as point based tracking, kernel based tracking and silhouette based tracking.

III. OBJECT TRACKING BY SPARSE COLLABORATIVE MODEL[6]

The goal of visual tracking is to determine a posteriori probability, $p(x_t|z_{1:t})$, of the target state where x_t is the object state, z_t is the observation at time t.

$$x_t=[l_x, l_y, \theta, s, \alpha, \emptyset]^T$$

Where $l_x, l_y, \theta, s, \alpha, \emptyset$ denote x, y translations, rotation angle, scale, aspect ratio, and skew respectively. These affine parameters are independent and modeled by six scalar Gaussian distributions. The motion model $p(x_t|x_{t-1})$ predicts the state at t based on the immediate previous state, and the observation model $p(z_t|x_t)$ describes the likelihood of observing z_t at state x_t. The particle filter is an effective realization of Bayesian filtering, which predicts the state regardless of the underlying distribution[13].[14]

Most tracking methods use rectangular image regions to represent targets, and thus background pixels are inevitably included as part of the foreground objects. Consequently, classifiers based on local representations may be significantly affected when background patches are considered as positive ones for update. On the other hand, the holistic appearance encoded by a target template is more distinct than the local appearance of local patches[15]. Thus, holistic templates are more effective for discriminative models to separate foreground objects from the background. In addition, local representations are more amenable for generative models because of flexibility. A collaborative observation model integrates a discriminative classifier based on holistic templates and a generative model using local representations[16].

A. *SPARSE DISCRIMINATIVE CLASSIFIER (SDC)[6]*

Motivated by the demonstrated success of sparse representation for vision tasks, a sparse discriminative classifier[17] for object tracking is given. In the following, the vector x is used to represent intensity values of a raster scanned image.

Extraction of Positive and Negative Templates: Training image set is built consisting of N_p positive templates and N_n negative templates [9]. Firstly, N_p positive sample images have to be taken around the target location. These images selected are to be down samples to a standard size. In our experiments we use 32x32 images which are then standard bilinear interpolated to obtain efficiency. The down sampled image is stacked together as shown in Fig 1 to form the set of positive templates .Similarly, the negative training set is composed of images further away from the target location, example within an annular region some pixels away from the target object as shown in Fig 1. Also the negative training set consists of both the background and images with parts of the target object are also taken as shown in figure 1. This helps to obtain better object localization as samples containing only partial appearance of the target are treated as the negative samples and the corresponding confidence values are likely to be small.

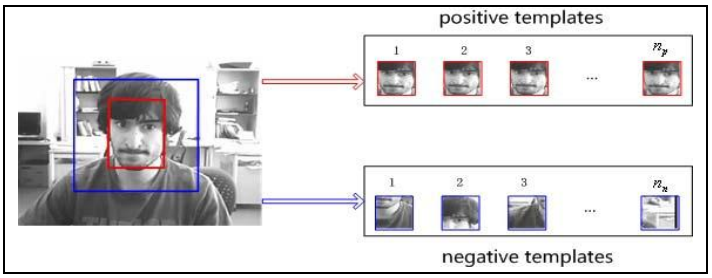


Fig 1: Positive and Negative templates[7]

3.2 Feature Selection[7]

The above-mentioned obtained patches of positive images and negative images are considered to be in gray-scale feature space and these are rich but are redundant in the sense that the dimensionality will be huge. Hence determinative features are best selected to distinguish the foreground object(positive images) from the background(negative images). This can be extracted by learning a classifier,

$$\min_s \|A^T s - p\|_2^2 + \lambda_2 \|S\|$$

Where A is composed of N_p positive templates A+ and N_n negative templates A−, K is the dimension of the features, and λ_2 is a weight parameter. Each element of the vector p represents the property of each template in the training set A, i.e., +1 for positive templates and −1 for negative templates. The solution to this equation is the sparse vector s, whose nonzero elements correspond to discriminative features selected from the K-dimensional feature space. This feature selection scheme adaptively chooses suitable number of discriminative features in dynamic environments via the l1 constraints. This is dynamic because as the moves the set of samples in A changes are hence the discriminative features. The features are projected to a subspace via a projection matrix S which is formed by removing all-zero

rows from a diagonal matrix S'and the elements are determined by

$$S'_{ii} = \begin{cases} 0, & s_i = 0 \\ 1, & \text{otherwise} \end{cases}$$

Both the training template set and the candidates sampled by a particle filter are projected to the discriminative feature space. Thus, the training template set and candidates in the projected space are A' = SA and x' = Sx.

B. *CONFIDENCE MEASURE*

The proposed SDC method is developed based on the assumption that a target image region can be better represented by the sparse combination of positive templates while a background patch can be better represented by the span of negative templates. Given a candidate region x, it is represented by the training template set with the coefficients α computed by

$$\min_{s} \|x' - A'\alpha\|_2^2 + \lambda_1 \|\alpha\|$$

where x' is the projected vector of x and λ3 is a weight parameter. A candidate region with smaller reconstruction error using the foreground template set indicates it is more likely to be a target object, and vice versa.

Thus, we formulate the confidence value H_c of the candidate x by

$$H_c = \frac{1}{1 + e^{-(\varepsilon_b - \varepsilon_f)/\sigma}}$$

Where $\varepsilon_f = \|x' - A'_+ \alpha'_+\|_2^2$ is the reconstruction error of the candidate x with the foreground template set A+, and α+ is the corresponding sparse coefficient vector. Similarly, $\varepsilon_b = \|x' - A'_- \alpha'_-\|_2^2$ is the reconstruction error of the candidate x using the background template set A−, and α− is the corresponding sparse coefficient vector. The variable σ is fixed to be a small constant that balances the weight of the discriminative classifier and the generative model. The reconstruction error is computed based on the target (positive) templates, which is less effective for tracking since both the negative and indistinguishable samples (e.g., patches covering some part of a target object) have large reconstruction errors when computed with the target (positive) template set. Thus, it introduces ambiguities in differentiating whether such patches are from the foreground or background. In contrast, our confidence measure exploits the distinct properties of the foreground and the background in computing the reconstruction errors to better distinguish patches from the positive and negative classes.

Recent advances of sparse coding for image classification as well as object tracking motivated to use this concept. Generative model is built for object representation that takes local appearance information of patches and occlusions into consideration.

Histogram Generation[6]: For simplicity, grayscale features are used to represent the local appearance information of a target object where each image is normalized to 32 × 32 pixels. We use overlapped sliding windows on the normalized images to obtain M patches and each patch is converted to a vector yi, where G denotes the size of the patch. The sparse coefficient vector β of each patch is computed by

$$\min_{\beta_i} \|y_i - D\beta_i\|_2^2 + \lambda_4 \|\beta_i\|$$

where the dictionary D is generated from J cluster centers using the k-means algorithm [3] on the M patches from the first frame (which consists of the most representative patterns of the target object) as shown in Fig. 4.2, and λ4 is a weight parameter.

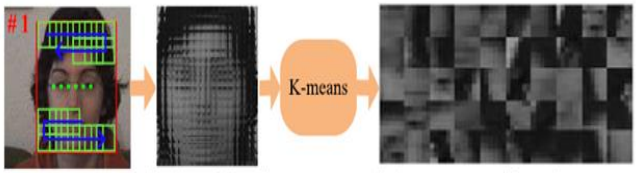


Fig 2(a) shows the First frame, Collection of all patches, (Dictionary generated from cluster centers[7].

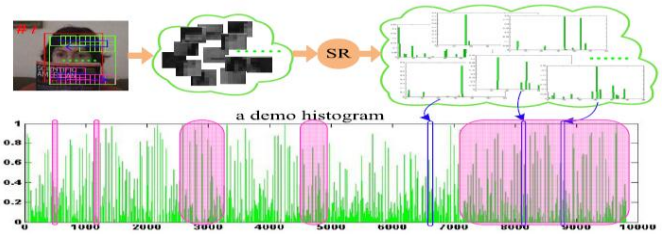


Fig 2(b): Histogram generation[7]

The first frame is scanned with overlapped sliding windows. Then the dictionary is generated with cluster centers of all the collected patches. The sparse coefficient vector βi of each patch is normalized and concatenated to form a histogram [7] by

$$\rho = [\beta_1^T, \beta_2^T, \ldots\ldots \beta_M^T]^T$$

where ρ is the proposed histogram for one candidate region, as shown by Fig.2(b).

A candidate region in the t-th frame with overlapped sliding windows is scanned. The sparse coefficient vectors of all the patches are concatenated to form the histogram of this candidate region. The histogram segments in magenta are coefficient vectors of the occluded patches. These segments and their counterparts in the template histogram are not taken into account when computing the similarity of this histogram and the template histogram.

Occlusion Handling[6]: In order to deal with occlusions [9], the constructed histogram is modified to exclude the occluded patches when describing the target object. A patch with large reconstruction error is regarded as occluding part and the corresponding sparse coefficient vector is set to be zero. Thus, a weighted histogram is generated by

$$\varphi = \rho \odot O$$

where $\odot$ denotes the element-wise multiplication. Each element of o is an indicator of occlusion at the corresponding patch and is obtained by

$$o_i = \begin{cases} 1, & \varepsilon_i < \varepsilon_0 \\ 0, & \text{else} \end{cases}$$

Where $\varepsilon_i = \|y_i - D\beta_i\|_2^2$ the reconstruction error of patch yi, and ε0 is a predefined threshold which determines whether the patch is occluded or not. Thus there is a sparse histogram for each candidate region. This representation scheme takes spatial information of local patches and occlusion into account, thereby making it more effective and robust.

Similarity Function

The histogram intersection function is used to compute the similarity of histograms between the candidate and the template due to its effectiveness by

$$L_c = \sum_{j=1}^{J\times M} \min(\varphi_c^j, \Psi^j)$$

Where φ_c^j and ψ are the histograms for the c-th candidate and the template. The histogram of the template ψ is generated and the patches y are all from the first frame and the template histogram is computed only once for each image sequence. It is updated every several frames. The vector o reflects the occlusion condition of the corresponding candidate. The comparison between the candidate and the template should be carried out under the same occlusion condition, so the template and the c-th candidate share the same vector oc. For example, when the template is compared with the c-th candidate, the vector o of the template is set to o_c.

C. *COLLABORATATION OF SDC and SGM*

A collaborative model [7] using the SDC and the SGM modules within the particle filter framework is considered. In this tracking algorithm, both the confidence value based on the holistic templates and the similarity measure based on the local patches contribute to an effective and robust description of probabilistic appearance model. The likelihood function of the c-th candidate region is computed by

$$p(Z_t|X_t^c) = H_c L_c = \left(\sum_{j=1}^{J\times M} \min(\varphi_c^j, \Psi^j)/1 + e^{-(\varepsilon_b - \varepsilon_f)/\sigma}\right)$$

and each tracking result is the candidate with the maximum a posteriori estimation. The multiplicative formula is more effective in our tracking scheme compared with the alternative of additive operation. The confidence value H_c gives higher weights to the candidates considered as positive samples (i.e., ε_f smaller than ε_b) and penalizes the others. As a result, it can be considered as the weight of the local similarity measure L_c.

D. *UPDATE SCHEME*

Since the appearance of an object often changes significantly during the tracking process, the update scheme is important and necessary. An update scheme is developed in which the SDC and SGM modules are updated independently. For the SDC module, we update the negative templates every several frames from image regions away (e.g., more than 8 pixels) the current tracking result. The positive templates remain the same in the tracking process. As the SDC module aims to distinguish the foreground from the background, it is important to ensure that the positive and negative templates are all correct and distinct.

For the SGM module, the dictionary D is fixed during the tracking process. Therefore, the dictionary is not incorrectly updated due to tracking failures or occlusions. In order to capture the appearance changes and recover the object from occlusions, the new template histogram ψn is computed by

$$\Psi_n = \mu\Psi_f + (1-\mu)\Psi_l \qquad \text{if } O_n < O_o$$

Where ψf is the histogram representing the manually set tracking result in the first frame and it is generated with the way shown in Fig. 4.3). The notion ψ_l is the histogram last stored before update, and μ is the weight. The variable On denotes the occlusion measure of the tracking result in the new frame. It is computed by the corresponding occlusion indication vector on using

$$O_n = \frac{1}{J\times M}\sum_{i=1}^{J\times M}(1 - o_n^i)$$

The appearance model is updated as long as the occlusion measure O_n in this frame is smaller than a predefined constant O_0. This update scheme preserves the first template ψ_f and takes the most recent tracking result into account.

IV. RESULTS

The images used for experiment analysis are used from the site[18][19].First the performance measures used in the evaluation of the algorithm are discussed and later the experimental results of tracking.

Performance measures:

The metrics that are used for a quantitative assessment of the various trackers are described. These metrics [5] are generally calculated on a frame-by-frame basis, but can also be calculated from the metrics over an entire sequence that summarizes a tracker's performance more succinctly. The measures employ centroid distance and overlap are commonly used for assessing the performance of tracking techniques. By choosing popular metrics, it is easier to compare the results of our evaluations to those of others.

Below is the image describing a person, their associated bounding box in blue and the position of a tracker denoted by red ellipse. The centroid distance is represented by the green line.

Fig.3: Centroid distance and overlap metrics

Centroid Distance

For a tracker centered at (x_t, y_t) and a ground truth (set of measurements that is known to be more accurate as compared to the measurements from the testing system) bounding box with centre (x_b, y_b) we define the centroid distance as dist centroid,

$$\text{dist}_{\text{centroid}} \triangleq \sqrt{(x_t - x_b)^2 + (y_t - y_b)^2}$$

In order to have a distance measure between the tracker and the ground truth that is comparable across objects of different sizes, we define the normalized centroid distance in terms of the width w_b and the height h_b of the bounding box normalized dist centroid,

$$normalised\_distcentroid \triangleq \sqrt{(\frac{x_t - x_b}{w_b})^2 + (\frac{y_t - y}{h_b})^2}$$

Overlap:

The proportion of the ground truth bounding box that is occupied by the tracker in a given frame is another useful measure of the tracker's accuracy. This metric is referred to as the overlap:

$$\text{overlap} \triangleq \frac{\text{area}_{\text{common}}}{\text{area}_{\text{bounding\_box}}}$$

Note that the tracker is treated as rectangular (as opposed to elliptical) for the purposes of the calculation.

Case I: TRACKING A BALL

Below are the results of tracking a ball in a video with moving cam, moving target, rotation and fast direction changes. Figure 4 shows the first frame of ball. Figure 5 shows the selection of the target of interest i.e. ball . Figure 6 and 7 shows the negative and the positive samples considered respectively. Figure 8 Patches after clustering and the cluster centroids respectively. Finally figure 9 gives the tracking ball in two of the frames in shown.

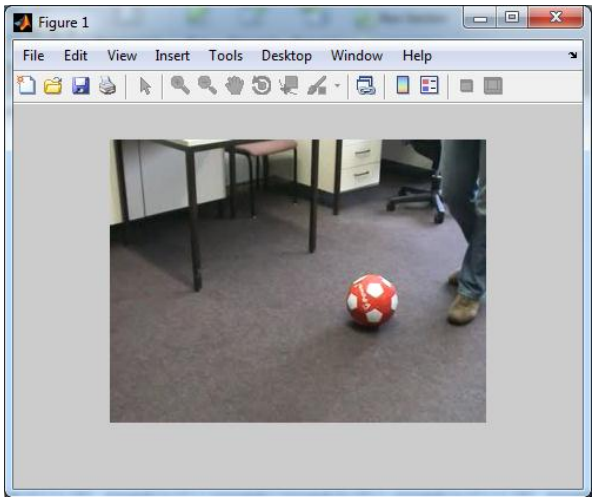


Fig. 4 First frame of ball tracking

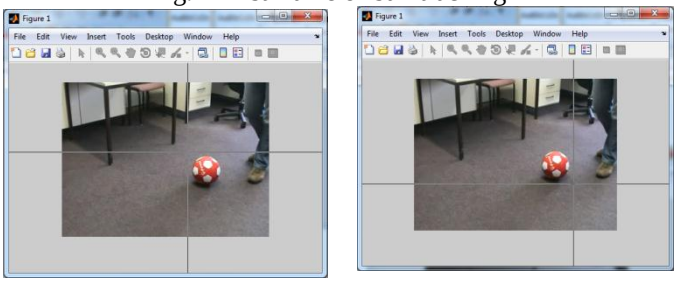


Fig 5. Selection of the ball

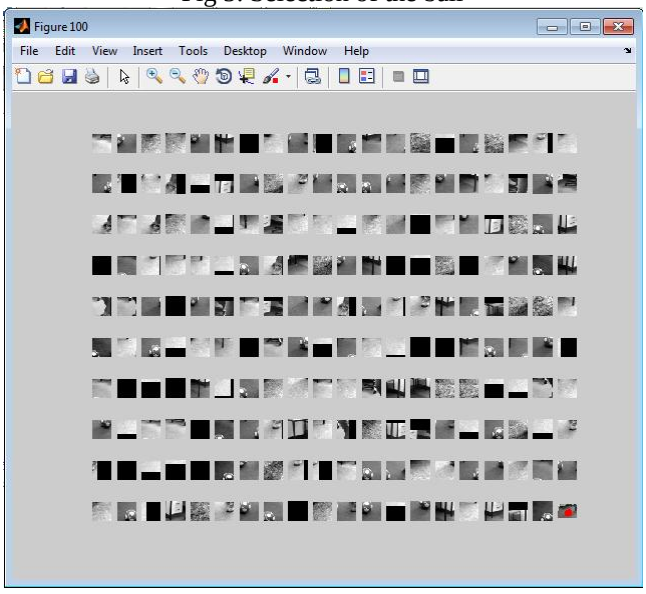


Fig 6: Negative samples

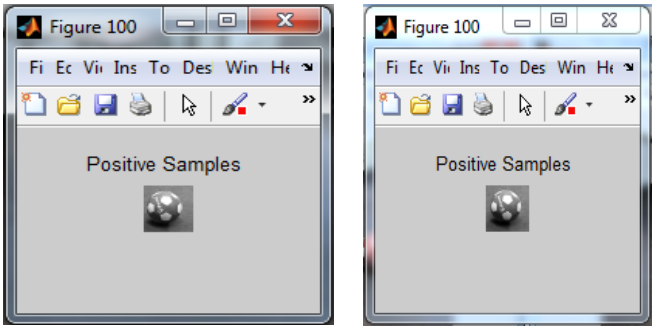


Fig 7 Positive samples

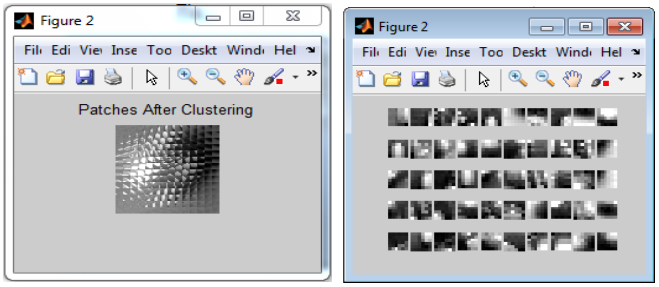


Fig 8: Cluster centres and patch centroids

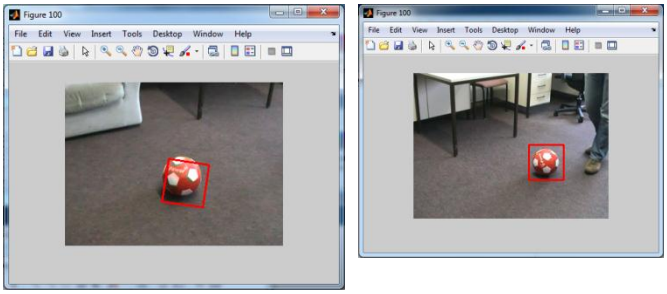


Fig 9 Tracking of the ball

Case 2:TRACKING A PERSON IN AN OCCLUDED VIDEO

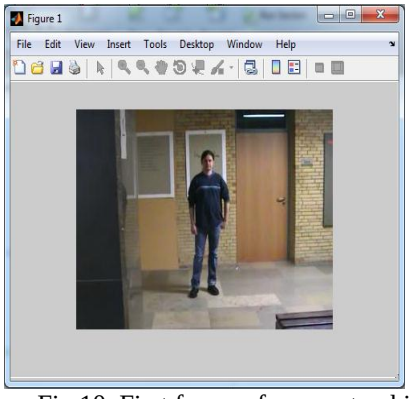


Fig 10. First frame of person tracking

Below are the results of tracking a person from a video with moving cam. This example shows that even with partial occlusion object can be tracked. Figure10-15 shows the similar results as in the case of tracking the ball.

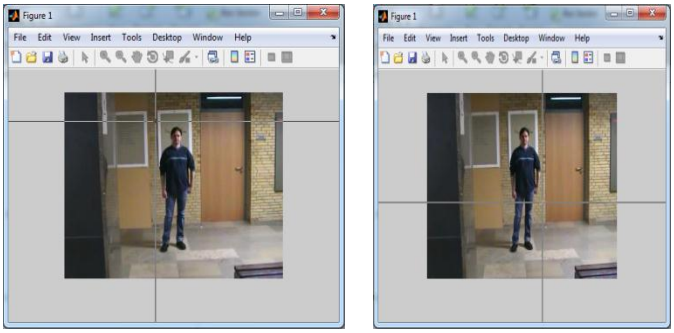


Fig 11 Selection of the target person

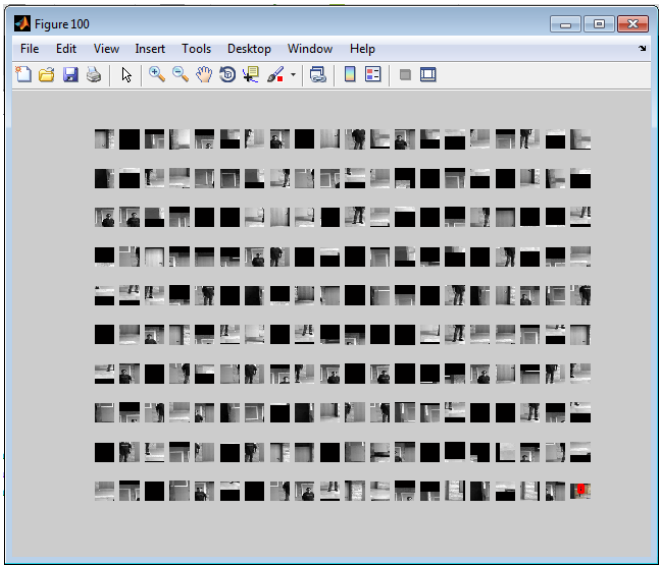


Fig 12 Negative samples

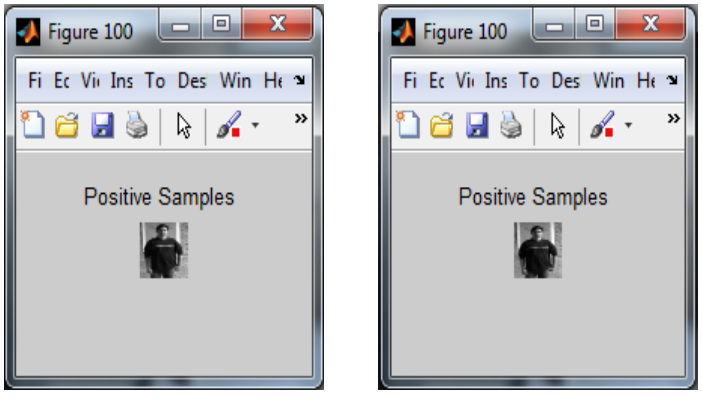


Fig 13 Positive samples

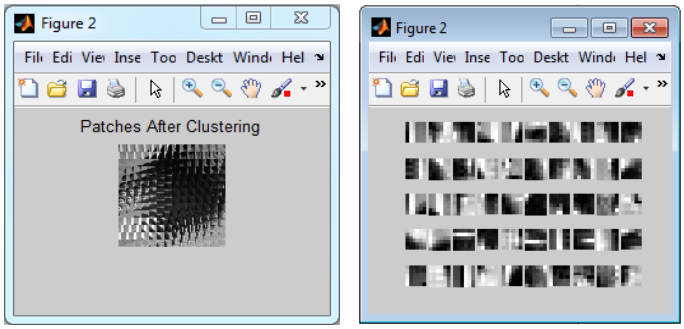


Fig 14 Patches after clustering and Cluster centres

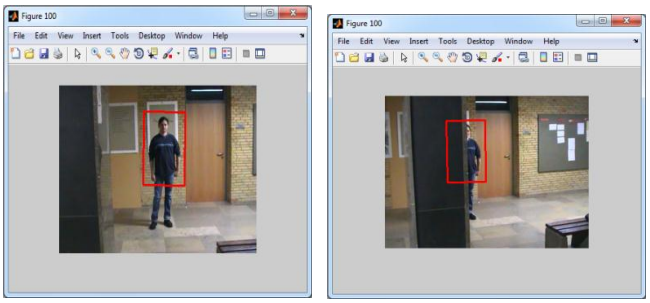


Figure 15 Tracking of the person with occlusion

Case 3: REAL TIME TRACKING VIDEO : Similar results for a video shot taken practically in tracking the teddy is shown in figure 16-21.

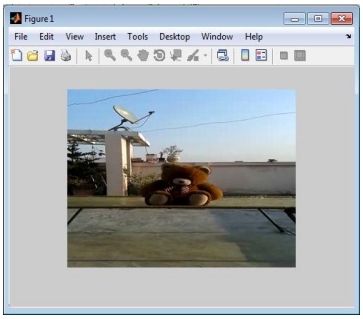


Fig16: First frame of teddy tracking video

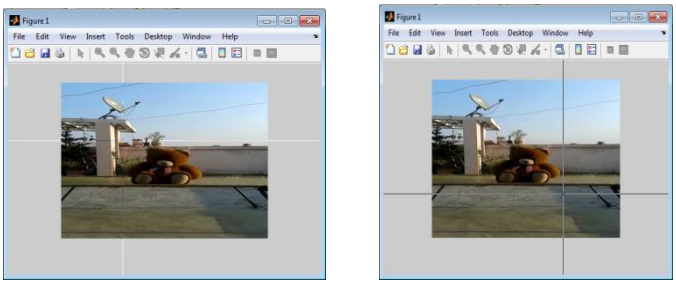
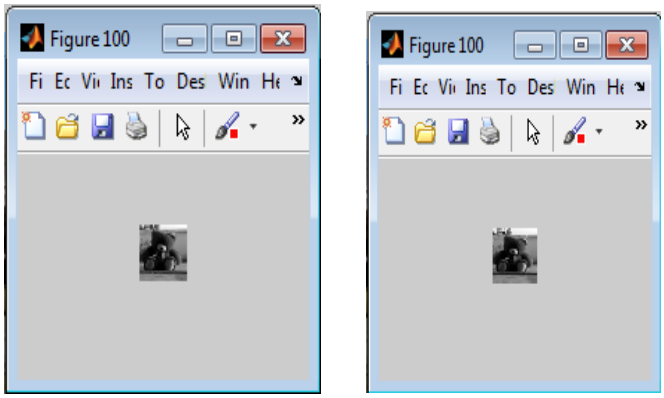


Fig 17: Selecting the teddy to be tracked



18: Fig Positive samples

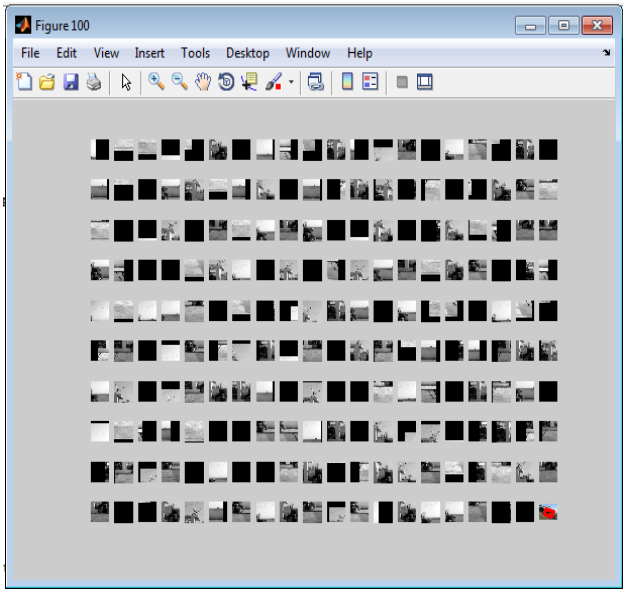


Fig 19: Negative samples

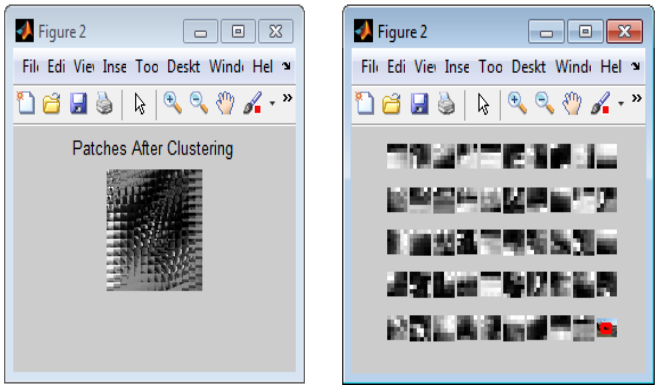


Fig 20: Patches after clustering and cluster centres

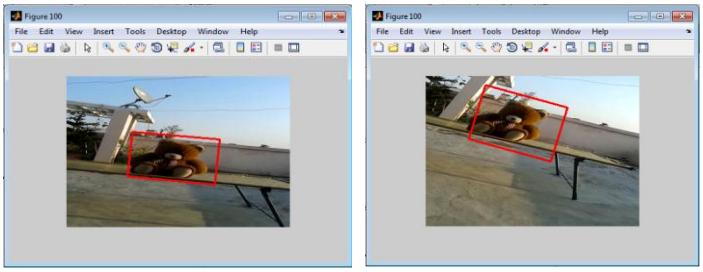


Fig21 Tracking teddy in real time video

V. CONCLUSIONS

In this paper, an effective and robust tracking method based on the collaboration of generative and discriminative modules is demonstrated. In this tracking algorithm, holistic templates are incorporated to construct a discriminative classifier that can effectively deal with cluttered and complex background. Local representations are adopted to form a robust histogram that considers the spatial information among local patches with an occlusion handling module, which enables tracker to better handle heavy occlusions. The holistic discriminative and local generative modules are integrated in a unified manner. Furthermore, the online update scheme enhances this method to adaptively account for appearance changes in dynamic scenes.

REFERENCES

[1] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2006, pp. 798–805.

[2] Gandham. Sindhuja, Renuka Devi S.M.: "A Survey on detection and tracking of objects in a Video sequence", International Journal of Engineering Research and General Science Volume 3, Issue 2, Part 2, March-April, 2015, pp.418-426.

[3] Allan D. Jepson, "Robust online appearance models for visual tracking", IEEE Conference on computer vision and pattern recognition, Kauai, 2001, vol. 1, pp. 415-422.

[4] . J. Black and A. D. Jepson, "Eigen tracking": Robust matching and tracking of articulated objects using a view based representation, Tech. Report T95-00515, Xerox PARC, Dec 1995.

[5] Gandham Sindhuja, Renuka Devi S.M. Comparative analysis of mean shift in object tracking. IEEE Conference on Power, Control, Communication and Computational Technologies for Sustainable Growth (PCCCTSG),2015, 283-287.

[6] Zhong, Wei, Huchuan Lu, and Ming-Hsuan Yang. "Robust object tracking via sparsity-based collaborative model." Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[7] Zhong, Wei, Huchuan Lu, and Ming-Hsuan Yang. "Robust object tracking via sparse collaborative appearance model." IEEE Transactions on Image Processing 23.5 (2014): 2356-2368.

[8] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust tracking using local sparse appearance model and k-selection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2011, pp. 1313–1320.

[9] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. Kulikowski, "Robust and fast collaborative tracking with two stage sparse optimization," in Proc. 11th Eur. Conf. Comput. Vis., 2010, pp. 624–637.

[10] I. Matthews, T. Ishikawa, and S. Baker, "The template update problem," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 6, pp. 810–815, Jun. 2004.

[11] Q. Wang, F. Chen, W. Xu, and M.-H. Yang, "Online discriminative object tracking with local sparse representation," in Proc. IEEE Workshop Appl. Comput. Vis., Jan. 2012, pp. 425–432.

[12] S. Avidan, "Support vector tracking," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 8, pp. 1064–1072, Aug. 2004.

[13] T. B. Dinh and G. G. Medioni, "Co-training framework of generative and discriminative trackers with partial occlusion handling," in Proc. IEEE Workshop Appl. Comput. Vis., Jan. 2011, pp. 642–649.

[14] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsitybased collaborative model," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 1838–1845.

[15] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 1822–1829.

[16] X. Mei and H. Ling, "Robust visual tracking using _1 minimization," in Proc. IEEE 12th Int. Conf. Comput. Vis., Oct. 2009, pp. 1436–1443.

[17] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 10, pp. 1728–1740, Oct. 2008.

[18] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[19] http://faculty.ucmerced.edu/mhyang/pubs.html, and http://ice.dlut.edu.cn/lu/publications.html,

