CONTEXTUAL OBJECT TRACKER WITH STRUCTURE ENCODING

Tanushri Chakravorty[⋆] *Guillaume-Alexandre Bilodeau* [⋆] *Eric Granger* [†]

[⋆] LITIV Lab., Polytechnique Montréal
[†] LIVIA, École de technologie supérieure, Montréal

ABSTRACT

Motivated by the problem of object tracking in video sequences, this paper presents a new Contextual Object Tracker with Structural Encoding (CTSE). The novelty in our tracking approach lies in the application of contextual and structural information (that is specific to a target object) into a *model-free* tracker. This is first achieved by including features from a *complementary* region having correlated motion with the target object. Second, a *local structure* that represents a spatial constraint between features within the target object are included. SIFT keypoints are used as features to encode both these information. The tracking is done in three steps. Firstly, keypoints are detected and described to encode object structure. Secondly, they are matched in every frame. Finally, each matched keypoint votes for the target object location locally in a *voting matrix* by using the encoded object structure. The voting method gives more priority to the keypoints that have been matched more often and are closest to the target's center than the rest. The proposed tracker is competitive with state-of-the art trackers while being significantly faster. It ranks as first or second most accurate tracker in experiments with standard datasets.

Index Terms— *Object tracking, Model-free tracking, Context, Appearance model, Object structure, Keypoints*

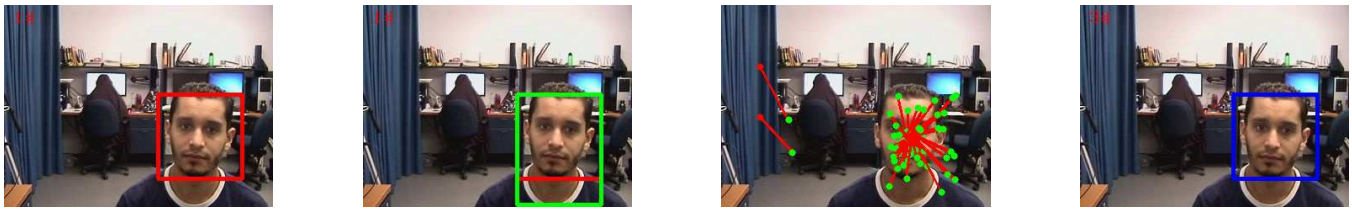
1. INTRODUCTION

Even after decades of research, object tracking in a real-world unconstrained environment remains an arduous task. The core problem in any tracking algorithm occurs due to abrupt and frequent appearance changes of the target object because of illumination, occlusion, scale and presence of objects having similar appearance to the target object (distractors) in the environment. Several approaches have been proposed to design strong appearance model, in order to discriminate an object from the background and match that appearance model in every frame, so as to have strong similarity measure for accurate tracking. Still, it is difficult to simultaneously address the problem of appearance changes caused by occlusion and illumination, and the problem of distractors. This paper presents a tracking method that can jointly address these problems.

The previous research gives an appropriate direction to solve these problems by focusing the target's appearance model – not only on the object's region description but also on the visual cues around the target object. The addition of contextual information besides the target object's region has been shown successful in the domain of object recognition [1] and semantic segmentation [2]. Approaches like [3], [4] explore the use of context in tracking. Both methods create a structure (topology) between correlated regions (having similar motion) and the main target object, and exploit this structure for object tracking. In recent work, the idea of using contextual information is slightly different. In [5], they use *Supporters* which are keypoint features spread over all the image and not necessarily around the target object that bear a correlated motion with the target. In [6], instead of using a set of image features from the whole frame, authors use features from the target itself and create an internal structure for all such features. The first common aspect is the use of contextual information. It is the data available from or around the target object having a correlated motion with the target object. The second common aspect is to use the structural information between the target object and the correlated features for efficient object tracking, thus dealing with distractors.

Another inspiration for our tracker comes from part-based trackers [7], [8], [9], where the target object is described by decomposing the objects region description into parts or patches. In [7], they use generative representation that belong to the target object only with patches pre-defined in a grid. These patches vote for the target object position in a competitive approach. However, their method becomes inappropriate for tracking non-rigid objects as the grid is unable to adjust to changes that occur due to deformations. In [8], they sample a set of overlapped patches and track object using visible patches during partial occlusion. In [9], they propose to use a histogram based model to encode the object structure. Part-based trackers like [10] and model-free trackers like [11], [12], [13], use discriminative representation and learning approaches to distinguish the target object from the background. From this, the common aspect is that decomposing the object's region is robust to partial occlusions.

This work was supported by fonds de recherche du Québec - Nature et technologies Team research project grant #167442



(a) Target Region (b) ROI (Region of Interest) (c) Voting by keypoints (d) Final Output

Fig. 1: Proposed Tracking Method (a) Target Region (b) ROI (Target + Complementary Region (below the red line)) (c) Voting by keypoints (green dots) for the target location (red dot) (d) Final Output by tracker

This paper presents a new model-based tracker entitled the Contextual Object Tracker with Structural Encoding (CTSE). It takes as input all the information about the target object and its context from the first frame and then tries to locate and update this pattern of input correctly for the rest of the video sequences. The CTSE is illustrated in Figure 1 and follows a three step process [1]. First, SIFT keypoints are extracted and described for the ROI (target + complementary). They provide invariance to illumination and robustness against distractors. The local structure for the target region is also computed. This provides the robustness against occlusion because the location of the target is described uniquely with respect to each individual keypoints. Thus, each keypoint behaves as a part of the object's region description. Second, these keypoints are matched, and each matching keypoint votes individually for the target position in a voting matrix. Third, in locations where multiple votes form a cluster, the global maximum of the obtained votes is selected as the final target location. Finally, the model is updated. Some key contributions to the state-of-the-art are as follows :

1. Keypoints having a *structure* spatial constraint and a *motion* correlation with the target center, are shown to be robust features for object tracking. Hence, principles from both context and structure may be combined into object tracking.
2. To achieve greater tracking accuracy, the inherent noise of the tracking method is modulated by utilizing a technique called *voting by keypoints*. In this, the voting for the target location is done using the structural configuration of each feature (keypoint).
3. The *quality* for every keypoint feature is estimated by maintaining a *structural configuration* for each keypoint. This helps in achieving a finer global prediction for the target location in every frame. The structural configuration is updated on the fly so as to accommodate the appearance changes.

Section 2 describes our appearance model and the tracking method. Section 3 describes the update steps for the structural configuration for a keypoint. Section 4 includes experimental results, and Section 5 draws conclusions. Results of the proposed tracker are compared to reference trackers using the video sequences in [14], and [6] respectively.

[1] https://bitbucket.org/tanushri/ctse

2. TRACKING METHOD

Generally, a tracking algorithm includes two main components: (1) appearance model that represents the characteristics of the target object, and (2) a search strategy to estimate the target's position in every frame.

2.1. Appearance model (target and complementary region)

Considering the underlying concept of model-free tracking, our tracker is initialized in the first frame by annotating a bounding box and a complementary region for a target face as shown in Figure 1b. Using this as our region of interest, SIFT keypoints are first detected all over the frame and then the keypoints contained inside the ROI are stored. Consider a target with a set of keypoints in the appearance model stored in a vector K. Let each keypoint be denoted by k_i, such that $k_i \in K$. We used SIFT keypoints as literature has shown that they are invariant to scale, translation, illumination and can handle small rotation variation, which makes it a very suitable feature for object tracking [15].

2.1.1. Encoding structure

As soon as the keypoints are detected and their descriptors are computed, a *structural* configuration for each keypoint is initialized. The structural configuration is represented as S_{k_i} = $[d_{k_i}, X_{k_i}, C_{k_i}, p_{k_i}]$ and consists of the following:

1. d_{k_i} = descriptor of keypoint
2. $X_{k_i}[\Delta x, \Delta y]$ (Spatial Constraint Vector) = Describes the keypoints location with respect to the target center.
3. C_{k_i} (Correlation Factor) = Indicates the keypoint's motion correlation with respect to the target center.
4. p_{k_i} (Proximity Factor) = Describes the importance of the keypoint's proximity to the target center. A keypoint located nearby to the target center will have higher proximity value as compared to others. It has a

direct effect on the C_{k_i} parameter of keypoint configuration as we will see in later subsections.

The encoded structure with spatial constraints helps in predicting the target's position in the next frame as the structure will remain mostly unchanged for the future frames of the video sequences. Therefore, when the target moves in the next frame, the points that have a correlated motion with the target center will also move by the same spatial translation in the next frame, but the relative distance (spatial constraint vector, X_{k_i}) of these points from the center will be constant. Hence, by re-detecting and matching the same keypoints as present in the appearance model for a target in the next frame t+1, we can estimate the new position of the target. The advantage of using such a structure aids in tracking during occlusion because even if a single keypoint is matched during occlusion (as rest of the keypoints will be hidden), the target object can still be tracked.

2.2. Search strategy

First SIFT keypoints are detected and described in the whole frame. These are matched with those present in the appearance model by comparing the Euclidean distance similarity between their descriptors. The advantage of detecting the keypoints in the whole frame helps in matching keypoints with the appearance model even if the target undergoes large or abrupt motion. The keypoint matching outputs a region with keypoints that co-occur with those present in the appearance model. We use a similar criteria as used [16] for removing erroneous matches and keep only those matches that have a distance ratio less than 0.8. The matching output gives a region, which is a coarse estimation of the target. Therefore, for finer prediction of the target location we have to use a different strategy called *voting by keypoints*.

2.2.1. Voting by keypoints using encoding structure

During tracking, there is inherent noise of the system, which will influence the target's center prediction by each keypoint. Therefore, we consider this inherent noise while estimating the final target location. We assume that all the pixels in the frame are affected by the same inherent noise and associate a single Gaussian pdf (probability density function) to all k_i. We want to vote in such a manner that a pixel on a patch around k_i, will have the highest vote with its closeness to the patch's center indicated by k_i (similar to a Gaussian function). Thus, k_i votes for the target's center by using its X_{k_i} parameter of S_{k_i}. Lets say the current position of k_i is x in the frame t and its corresponding structural spatial constraint is denoted by X_{k_i}. Hence, the Gaussian pdf with which k_i will cast its vote can be written as:

$$P(x|k_i) \propto \frac{1}{\sqrt{2\pi|\Sigma|}} exp(-0.5(x-X)^T\Sigma^{-1}(x-X)) \quad (1)$$

Here, Σ is a covariance matrix. Therefore, the local prediction by given k_i for the target's new center location is given by:

$$x_{LocPred_{k_i}} = P(x|k_i)C_{k_i}\mathbb{1}_{(k_i \in K)} \quad (2)$$

Hence, each keypoint *votes* for the target's center location with a Gaussian pdf and its Correlation factor C_{k_i} and $\mathbb{1}_{(k^{(i)} \in K)}$ is an indicator function, which is set to one for keypoints that are matched in current frame. All such individual votes are summarized in a vote matrix. In order to select the most probable location of the target center, we find the location inside the vote matrix (VM) where the sum of individual votes is highest, resulting in a cluster of votes. This shows that a cluster of keypoints have voted for the same center location for a target object. Hence, the final target center is given by Equation 3, and is represented as follows :

$$x_{targetCenter} = \arg\max_{x \in VM}(\sum_{i=0}^{K}(P(x|k_i)C_{k_i}\mathbb{1}_{(k_i \in K)})) \quad (3)$$

3. DETERMINING KEYPOINT QUALITY

3.1. Adaptive correlation and proximity factor

As seen from Equation 2, the correlation factor C_{k_i} plays a major contribution in determining the global prediction for the target center. Initially all the keypoints in K are assigned with an initial value for C_{k_i}. With every new frame processed t, the C_{k_i} parameter value in the structural configuration of keypoint k_i updates with *learning factor*, α using Equation 4 as follows:

$$C_{k_i}^{t+1} = (1-\alpha)C_{k_i}^t + \alpha p_{k_i}^t \mathbb{1}_{(k_i \in K)} \quad (4)$$

Here the term $p_{k_i}^t$ represents the proximity factor for a particular keypoint, k_i at frame t. The p_{k_i} for a particular varies non-linearly with its closeness to the target's center and is evaluated by using a function, given by the following Equation 5:

$$p_{k_i}^t = max((1-|\lambda(x_{TargetCenter} - x_{LocPred_{k_i}})|), 0.0) \quad (5)$$

Here λ is a constant. Hence, a keypoint that is closer to the predicted target center ($x_{TargetCenter}$), will have more importance in contributing its vote for in the Gaussian pdf (Refer Equation 2) in the next frame $t+1$, than those which are far from the target center. By doing this, we achieve higher accuracy for target center location because even if certain keypoints that are erroneously matched, they will have a very less contribution in vote matrix. For the rest of the keypoints that have not been matched, their C_{k_i} reduces (Refer Equation 4).

4. EXPERIMENTAL RESULTS

For comparison, we use state-of-the-art evaluation criteria namely, bounding box *Overlap Ratio* (**OR**) and average *Center Location Error* (**CLE**). OR is the average percentage of

Table 1: Comparison of CLE and OR of proposed tracker with respect to state-of-the-art part-based trackers. **Bold** red indicates the best results and blue italics indicates the second best.

Videos	**SPT**[10]		**SCMT**[9]		**AST**[8]		**SAT**[6]		**CTSE**(proposed)		**CTSE**(no context)	
	CLE	OR	CLE	OR	CLE	OR	CLE	OR	CLE	OR	CLE	OR
FaceOcc1	116.84	0.05	5.07	**1.00**	85.43	0.25	14.26	*0.99*	**3.77**	**1.00**	*3.89*	**1.00**
Girl	**8.97**	**0.84**	201.27	0.19	53.42	0.17	*10.01*	**0.84**	10.52	*0.78*	10.61	*0.78*
David	36.09	*0.62*	33.81	0.60	68.57	0.37	**10.48**	**1.00**	*26.38*	0.60	*26.38*	0.60
Cliff bar	*22.11*	0.51	77.31	0.24	35.35	**0.69**	25.33	*0.60*	26.13	0.51	**20.68**	0.59
jp1	35.21	0.18	17.74	0.78	16.66	0.84	*7.03*	*0.89*	**5.95**	**0.99**	**5.95**	**0.99**
jp2	30.58	0.39	69.44	0.55	45.15	0.55	*7.25*	*0.93*	**3.91**	**0.99**	**3.91**	**0.99**
wdesk	79.92	0.13	34.17	0.57	80.97	0.32	**11.12**	**0.90**	*11.23*	*0.85*	*11.23*	*0.85*
wbook	11.27	0.98	**5.09**	**1.00**	8.68	*0.99*	11.87	*0.99*	*6.92*	*0.99*	*6.92*	*0.99*

frames where the overlap of BB's (bounding boxes) of tracker and ground truth is at least 50%. CLE is the Euclidean distance between the center's of BB's of tracking output and ground truth. The videos for validation have the following attributes: partial and long term occlusion (*FaceOcc1*, [7], *wbook and wdesk*, [6]), illumination, large camera motion and background change (*David*, [11], *Girl*, [17]), Background Clutter (*Cliffbar*[11]), and moderately crowded scene (*jp1, jp2* [6]). As seen from Table 1, the performance of our method is very good for scenes with distractors, *jp1*, *jp2*. Our method delivers a precision of 0.99 with the least error as compared to rest of the trackers, because the encoded structure and keypoints prevents the tracker from switching to distractors. The voting by keypoints using the structure helps in greatly reducing the error. The encoded structure with the complementary region helps in prediction of target during long-term partial occlusion with a precision of 1.00 in *FaceOcc1*, 0.85 in *wdesk*, and 0.99 in *wbook*, as the subset of features help in target prediction even when a significant part of the target object remains hidden for a long time. For *David* and *Girl*, our method gives a competitive performance with the ability to track objects during large camera motion and illumination change. When experimented our tracker without a complementary region (no context), the error slightly increases for *FaceOcc1*, and *Girl*, as the target region remains hidden for sometime. This results in lesser matching, and lesser impact of correlation factor in voting. Whereas for *Cliffbar*, the results significantly improve without the context, as the complementary region takes into account the background region with no motion correlation with the target, which is otherwise observed in videos having torso and head (*FaceOcc1,Girl*). Note, if both (head+torso) are occluded at the same time, the context is less advantageous, and thus being scenario dependent. Hence, context is useful for videos where other objects have correlated motion with the target. For a typical 320x240 resolution video sequence, our tracker runs with 10 frames per second on Intel Core i7, 3.40 GHz machine. Figure 2 shows the qualitative results of our tracking method.

Fig. 2: Qualitative Results. From left to right row-wise and top to bottom : *Girl*, *jp1*, *jp2* , *Cliffbar* video sequences.

5. CONCLUSION

In this paper, a new tracker has been proposed that combines the concept of context and structure for object tracking. Experimental results have shown that using keypoint features that have a correlated motion with the target center and that are organized in a structure having spatial constraints with respect to the target center, are robust features for object tracking in video sequences. Our results emphasize that by adapting the structural configuration parameters of the keypoints, improves tracking for challenges such as partial and long-term occlusion, illumination, and distractors, etc. However, the robustness of our tracker depends on the keypoint detector. The future research will seek to increase the precision of our tracker by combining a detector with our method.

6. REFERENCES

[1] A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin, "Context-based vision system for place and object recognition," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, Oct 2003, pp. 273–280 vol.1.

[2] R. Mottaghi, S. Fidler, Jian Yao, R. Urtasun, and D. Parikh, "Analyzing semantic segmentation using hybrid human-machine crfs," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 3143–3150.

[3] Lu Zhang and L. Van Der Maaten, "Structure preserving object tracking," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 1838–1845.

[4] Ming Yang, Ying Wu, and Gang Hua, "Context-aware visual tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 7, pp. 1195–1209, July 2009.

[5] H. Grabner, J. Matas, L. Van Gool, and P. Cattin, "Tracking the invisible: Learning where the object might be," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 1285–1292.

[6] W. Bouachir and G.-A. Bilodeau, "Structure-aware keypoint tracking for partial occlusion handling," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, March 2014, pp. 877–884.

[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 2006, vol. 1, pp. 798–805.

[8] Xu Jia, Huchuan Lu, and Ming-Hsuan Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 1822–1829.

[9] Wei Zhong, Huchuan Lu, and Ming-Hsuan Yang, "Robust object tracking via sparsity-based collaborative model," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 1838–1845.

[10] Shu Wang, Huchuan Lu, Fan Yang, and Ming-Hsuan Yang, "Superpixel tracking," Los Alamitos, CA, USA, 2011, vol. 0, pp. 1323–1330, IEEE Computer Society.

[11] 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, pp. 1619–1632, Aug 2011.

[12] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-n learning: Bootstrapping binary classifiers by structural constraints," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 49–56.

[13] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proceedings of the British Machine Vision Conference*. 2006, pp. 6.1–6.10, BMVA Press, doi:10.5244/C.20.6.

[14] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Online object tracking: A benchmark," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 2411–2418, 2013.

[15] Jared Heinly, Enrique Dunn, and Jan-Michael Frahm, "Comparative evaluation of binary features," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part II*, Berlin, Heidelberg, 2012, ECCV'12, pp. 759–773, Springer-Verlag.

[16] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[17] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, Jun 1998, pp. 232–237.