Adaptive appearance learning for visual object tracking

ADAPTIVE APPEARANCE LEARNING FOR VISUAL OBJECT TRACKING

Zulfiqar Hasan Khan, Irene Yu-Hua Gu

Dept. of Signals and Systems, Chalmers University of Technology, Sweden
{zulfiqak, irenegu}@chalmers.se

ABSTRACT

This paper addresses online learning of reference object distribution in the context of two hybrid tracking schemes that combine the mean shift with local point feature correspondences, and the mean shift under the Bayesian framework, respectively. The reference object distribution is built up by a kernel-weighted color histogram. The main contributions of the proposed schemes includes: (a) an adaptive learning strategy that seeks to update the reference object distribution when the changes are caused by the intrinsic object dynamic without partial occlusion/intersection; (b) novel dynamic maintenance of object feature points by exploring both foreground and background sets; (c) integration of adaptive appearance and local point features in joint object appearance similarity and local point features correspondences-based tracker to improve [7]; (d) integration of adaptive appearance in joint appearance similarity and particle filter tracker under the Bayesian framework to improve [10]. Experimental results on a range of videos captured by a dynamic/stationary camera demonstrate the effectiveness of the proposed schemes in terms of robustness to partial occlusions, tracking drifts and tightness and accuracy of tracked bounding box. Comparisons are also made with the two hybrid trackers together with 3 existing trackers.

Index Terms— Visual object tracking, dynamic Appearance, SIFT, RANSAC, anisotropic mean shift, particle filters, hybrid trackers.

1. INTRODUCTION

Dynamic appearance modeling in visual tracking has drawn increasing interest in the recent years. The main objective is to model the intrinsic (e.g. pose variation, shape deformation) and extrinsic (e.g. different illumination, camera motion, viewpoint and occlusion) appearance variability. Many techniques have been proposed for adapting appearance changes. [1] extends gradient-based optical flow to handel the appearance variations. [2] handles the fast illumination changes by fast differential EMD tracking. [3] introduces an online subspace learning method with a sample mean update. [4] proposes to label objects and background pixels by ensembles of online learned weak classifiers. Good results have been achieved however challenges and open issues remain in tracking objects in complex scenarios.

To design more robust trackers, some hybrid methods combining MS (mean shift) with local point feature-based tracking or Bayesian framework like PFs (particle filters) are proposed. [5] proposes an expectation-maximization algorithm that integrates SIFT features along with color-based object appearance in MS. [6] uses feature points to handle occlusion and scaling under the MS framework. [7] uses MS and consensus point feature correspondences to improve tracking accuracy via a coarse-to-fine process. [8] proposes to embed the MS in particle filters to track human hands. [9] proposes to weight particles by using the observation model, followed by applying the MS on particles with large weights, called elite particles. [10] combines PFs and anisotropic MS seeking multiple appearance modes by partitioning a rectangular bounding box into sub-regions.

However, the above appearance based tracking methods share a common problem that the appearance models lack dynamics. Motivated by this, we propose an adaptive framework for target object distribution. We model the appearance changes between video frames by exploiting kernel weighted color histogram. A criterion is introduced to find the highest and stable local tracking performance frames that are least affected from the extrinsic appearance variability, especially partial occlusions. This adaptive appearance model makes image matching robust to various dynamics and poses of appearances. The proposed method can be characterized as a generic online learning paradigm which does not require prior training data. The proposed scheme is then integrated with two hybrid methods by combining anisotropic MS with SIFT and PFs. Videos containing partial occlusions, intersections, close color distributed backgrounds or significant deformations are used for testing and measuring the performance. Evaluation and comparison with existing methods shows marked improvement.

2. ONLINE LEARNING OF REFERENCE OBJECT APPEARANCE DISTRIBUTION

The basic idea behind our proposed online learning method is to *only* update the reference model at those frames when they indicate local highest performance of correct tracking without the interference (e.g. occlusion or intersection) from the background or other objects. Further, the frequency of update do not need to be high and can be performed in a fixed frame

interval, since object changes are usually gradual due to the mechanical movement (e.g., a person takes off an overcoat during the walking). To achieve this, we seek the highest and stable local tracking performance in each fixed interval, to decide whether or not the reference object distribution shall be updated in this interval.

The candidate (or the reference) object appearance in a video frame is described by the spatial-kernel weighted color histograms $p = \{p_u\}$ (or $q = \{q_u\}$), as follows:

$$p_u = \frac{c}{|\Sigma|^{\frac{1}{2}}} \sum_{j=1}^{n} k(\tilde{y_j}^T \Sigma^{-1} \tilde{y_j}) \delta[b_u(I(y_j)) - u] \quad (1)$$

where $u = 1, \cdots, m$, m is the total number of bins, $\tilde{y_j} = (y_j - y)$, $\tilde{x_j} = (x_j - x_0)$, Σ is the kernel bandwidth matrix, $b_u(I(y_j))$ is the bin index of color histogram of a candidate object image centered at y_j, c is a constant for the normalization, k is the spatial kernel profile, y is the center of kernel or bounding box, and $I(y_j)$ is the candidate object image within the bounding box.

The appearance similarity between a candidate and the reference object is described by the Bhattacharyya coefficient ρ defined as, $\rho(p, q) = \sum_u \sqrt{p_u(y, \Sigma) q_u}$, where u is the histogram bin.

Let $\rho_t = \sum_u \sqrt{q_u^{j-1} p_u^t}$ be the Bhattacharyya coefficient between the current tracked object from the final tracker and the reference object in the previous $(j-1)$th interval, and $\mathbf{x}_{t,i}$ be the 4 corners in the tracked regions $V_t^{(obj)}$. Noting that q_u^{j-1} implies q_u^t for $t \in [(j-2)S+1, (j-1)S]$, where S is the total frames in the interval (S=25 frames in our tests). Then, the reference model in the jth interval is updated if the following two conditions are both satisfied for all frames within the jth interval:

$$\begin{cases} \text{dist}_t = \sum_{i=1}^{4} \|\mathbf{x}_{t,i} - \mathbf{x}_{t-1,i}\|^2 < T_1, \\ \rho_t > T_2 \end{cases} \quad (2)$$

where $t \in [(j-1)S+1, jS]$, $T_1^{(2)}$ and $T_2^{(2)}$ are empirically selected thresholds ($T_1^{(2)}$=10, $T_2^{(2)}$=0.95 in our tests). $j = 1, 2, \cdots$, then the best stable high performance frame number is chosen from

$$t^* = \text{argmax}_{t \in [(j-1)S+1, jS]} \rho_t \quad (3)$$

and the reference object distribution is updated by

$$q^j = \kappa p^{t^*} + (1 - \kappa) q^{j-1} \quad (4)$$

where q^j (or, q^{j-1}) is the updated reference object pdf in the jth (or $(j-1)$th) interval, κ is the constant controlling the learning rate ($\kappa = 0.1$ in our tests), p^{t^*} is the pdf where t^* is chosen from (3). If (2) is not satisfied, then the reference object distribution is not updated, i.e., $q^j \leftarrow q^{j-1}$.

3. A HYBRID TRACKING SCHEME-1 COMBINING ADAPTIVE APPEARANCE WITH POINT FEATURES

The block diagram of the proposed scheme is shown in Fig.1, where both consensus point feature correspondences and object appearance similarity are used. The final tracker is formulated through maximizing the given criterion based on these two candidate trackers.

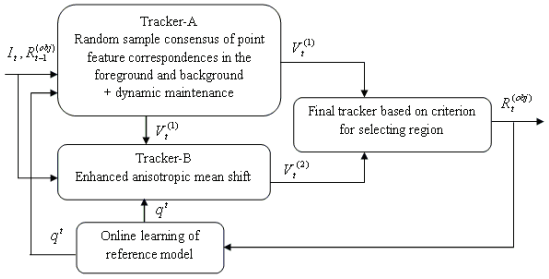

Fig. 1. Block diagram of the proposed *hybrid tracking scheme-1* with online dynamic appearance learning.

For tracker-A, feature point correspondences are estimated through using maximum consensus point correspondences in the foreground and background areas by SIFT [13] and RANSAC [14]. It generates a parameterized candidate region $R_t^{(1)} : V_t^{(1)} = [x^{(1)}, y^{(1)}, w^{(1)}, h^{(1)}, \theta^{(1)}]_t^T$, corresponding to the 2D center, width, height and orientation of the region. Apart from dynamic point maintenance and the use of separate foreground and background areas, a re-initialization process is applied to tracker-A if the similarity between the tracked area and the reference object area becomes small, indicating a potential tracking drift (e.g., due to few correspondence points) which could propagate through frames. For tracker-B, an enhanced anisotropic mean shift is achieved by choosing the center between the candidate region of tracker-A and the previous candidate region of mean shift [11], and by allowing a re-initialization process. The basic idea is to guide the MS to a correct target object location especially when confusing track situations occur (e.g. other objects with similar color distributions, or cluttered), through assigning the tracker to an area that is more agreeable with the local feature correspondences of the target. It generates a parameterized candidate region $R_t^{(2)} : V_t^{(2)} = [x^{(2)}, y^{(2)}, w^{(2)}, h^{(2)}, \theta^{(2)}]_t^T$. A 3rd candidate object region $R_t^{(3)} : V_t^{(3)} = \sum_{i=1}^{2} \tilde{\rho}_t^{(i)} V_t^{(i)}$ where $\tilde{\rho}_t^{(i)} = \frac{\rho_t^{(i)}}{\rho_t^{(1)} + \rho_t^{(2)}}$. The optimal target object region $R_t^{(obj)}$ from the final tracker is then selected by maximizing the following criterion, $\hat{V}_t^{(obj)} = \arg\max_{V_t^{(i)}} (\rho_t^{(1)}, \rho_t^{(2)}, \rho_t^{(3)})$ where $\rho_t^{(i)}$, i=1,2,3, is the Bhattacharyya coefficient measuring the similarity between the reference and the candidate object from the tracked candidate area $R_t^{(i)}$. The online learning of the reference appearance distribution q_u^t is also applied.

Dynamic maintenance of foreground object feature points: The proposed approach for online dynamic maintenance and updating of consensus points is similar in a spirit to the work in [12], however, two major changes are introduced: One is the use of separate point sets for the foreground and the background, another is the utilization of the Bhattacharyya

coefficient. The method is described as follows:

- The Bhattacharyya similarity coefficient of object region $\rho_t^{(1)} = \sum_u \sqrt{p_u^{t,(1)}(y,\Sigma)\, q_u^t}$ is calculated;
- If $\rho_t^{(1)} > T_F$, where T_F is a threshold determined empirically, it indicates that the object area is unlikely to be occluded or newly introduced. The point features in the object area are assigned a score as follows:

$$S_t = \begin{cases} S_{t-1} + 2 & \text{matched consensus point} \\ S_{t-1} - 1 & \text{matched outlier point} \\ median(S_{t-1}) & \text{not matched point} \end{cases} \quad (5)$$

  For inliers, i.e., corresponding points that fit the estimated transformation model, scores are increased. For outliers i.e. point correspondences that do not fit the transformation, scores are decreased. For newly local point features that are only within the candidate object region and do not correspond to any background point features, their score is set to median value of the previous scores.
- If $\rho_t^{(1)} < T_F$, it indicates that the candidate region likely contains newly introduced area e.g. from partial occlusions or object intersections. Hence the dynamic maintenance is temporarily frozen.

Points in the consensus correspondence set are then sorted according to their scores. A pruning process is applied to remove some feature points with small weights in order to maintain a reasonable size for the points set. This is done by keeping the first L_F (L_F=1000 in our tests) highest weight points in the set. New points are added (e.g. due to change pose change or deformation of object) if they fir to the estimated motion model.

Dynamic maintenance of background feature points:
In each frame, all feature points in the background region are added into the background set. Feature points in the set are sorted according to their aging. If the total number of feature points exceeds L_B (L_B=1500 in our tests), then only the newest L_B feature points are kept while the remaining old aging feature points are removed.

4. A HYBRID TRACKING SCHEME-2 COMBINING ADAPTIVE APPEARANCE UNDER BAYESIAN FRAMEWORK

The block diagram of the proposed hybrid scheme-2 is shown in Fig.2. In the PFs (the top block), the state vector $\hat{s}_{t-1}$ describes the shape of a parametric rectangular candidate object box. The initial particles $s_t^{(j)}$ are generated according to the Brownian motion model. Then, the likelihood $p(z_t|\tilde{s}_t^{(j)})$ is computed using the Bhattacharyya distance that is obtained from the the dynamic appearance allocated by the anisotropic MS (the bottom block). In addition, the rectangular bounding box is partitioned into several disjoint concentric areas. This allows the MS to seek multiple modes that give better descriptions of the spatially-dependent distribution of object appearance. Based on the MS estimates, particles are re-distributed to positions related to large weights, and posterior pdf of state vector is then estimated by the PF using the appearance-related likelihood. See [10] for more details. The online learning approach in section 2 is added. Although visual object tracking using joint PFs and MS has been reported [8, 9, 10], non of these joint schemes contain online learning of the dynamic reference object.

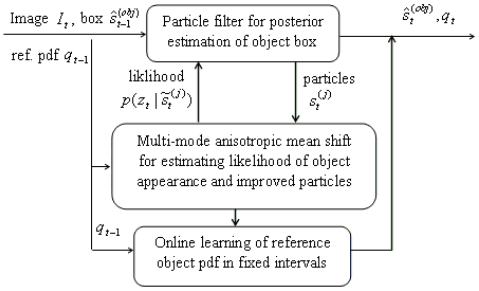

Fig. 2. Block diagram of proposed *hybrid tracking scheme-2* with online dynamic appearance learning.

Adding online learning in this hybrid tracker has led to more robust tracking to long-term partial occlusions and intersections in terms of reducing tracking drift and tracking failure.

5. EXPERIMENTS AND RESULTS

We tested the two hybrid schemes on several videos containing moving objects with a range of complexity (e.g. long-term partial occlusion, object intersection, deformation or, pose changes). Only some tracking results are included in this paper due to space limitations. For each object area (or box) a $32 \times 32 \times 32$ bin histogram of RGB color images is used. The maximum iterations is 10 for the enhanced MS. γ for normalizing the bandwidth matrix of MS was empirically determined as 0.2 for case study $1, 4, 5, 7$, 0.33 for case studies $2, 3, 6$, respectively. $T_1^{(2)}$=10, and $T_2^{(2)}$=0.95 in (2). The 1^{st}proposed scheme in Fig.1, hybrid tracker-1, is compared with two existing methods: anisotropic MS tracker in [11] referred to as *Tracker-1*; A tracker using SIFT [13] followed by RANSAC [14] *referred to as Tracker-2*. Fig.3 shows the tracking results. Also the proposed scheme in Fig.2, hybrid tracker-2, is compared with two existing methods: an anisotropic MS tracker [11] (*Tracker-1*); Combined MS and PFs (MSPF) tracker [8] *referred to as Tracker-3*. Fig.4 shows the tracking results. Further, the two hybrid trackers (in Fig.1 and Fig.2) are compared and results are shown in Fig.5. Fig.6 shows the Euclidean distance between the 4-corners of tracked bounding box between the ground truth box and the tracked box of the two hybrid trackers (with/without online appearance learning). Fig.7 shows the Euclidian distances between the tracked box with ground truth regions for two hybrid trackers, and the existing *tracker-1*, *tracker-2*, *tracker-3*". Observing the results in these figures, it is shown that the two hybrid trackers are very robust to the types of changes (variable head tilting speed, partial occlusion and in intersections, background cluttered). Further online learning of appearance object has significantly enhanced tracking

as compared to their corresponding versions without online learning.

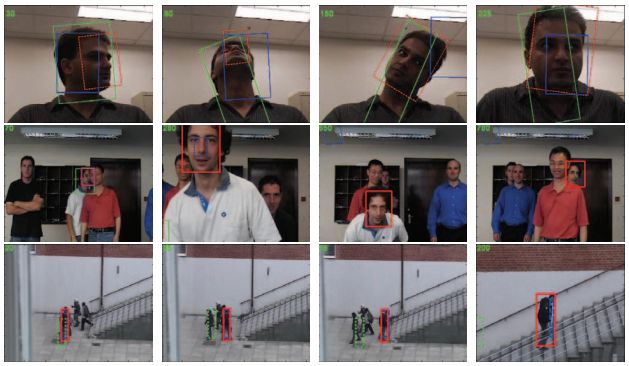

Fig. 3. Tracking results from hybrid tracker-1 (red), *tracker-1* (green), *tracker-2* (blue). Rows 1-3: video "Behzad", "multifaces", "stair walking".

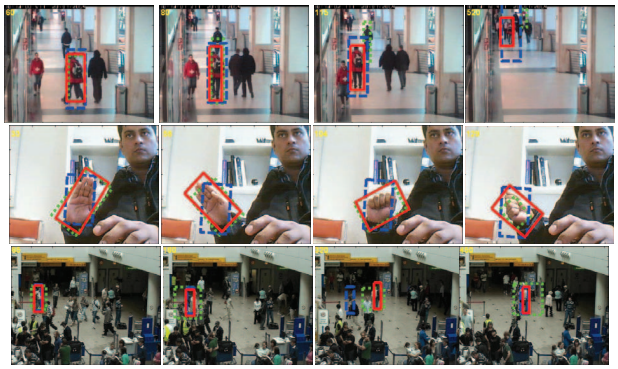

Fig. 4. Tracking results from the proposed hybrid tracker-2 (red), the existing *tracker-1* (green), *tracker-3* (blue). Rows 1-3: for videos "ThreePastShop2Cor", "Hand", "Pets2007_S05_C1" from PETS2007 dataset 5.

Fig. 5. Tracking results from the proposed two hybrid trackers for the video "walking woman": hybrid tracker-1 (red), and hybrid tracker-2 (green).

6. CONCLUSION

A novel online learning method for updating object appearance distribution and dynamic maintaining (i.e. updating, adding and pruning) foreground and background feature points is proposed. The method is then applied to two proposed hybrid tracking schemes (Fig.1 and Fig.2) to further improve the tracking performance. Experimental results have shown that the proposed hybrid trackers are extremely robust, resulting in marked improvement on reducing tracking drift, robustness to long-term partial occlusions or object intersections, as well as complex cluttered background. Comparisons and evaluation with three existing methods have provided further support to the proposed schemes. Comparisons between the two proposed hybrid trackers show that the hybrid

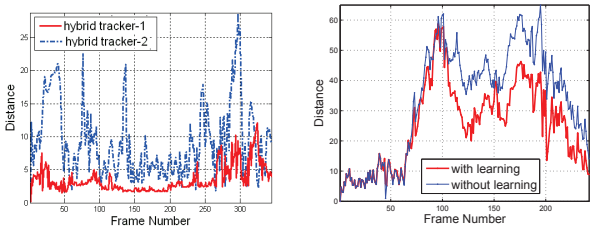

Fig. 6. Comparing Euclidean distances from different trackers. Left: from hybrid tracker-1 (red) and hybrid tracker-2(blue) on the video "woman walk"; Right: from the proposed hybrid tracker-1 with online appearance learning (red) and without learning (blue) for video "stair walking".

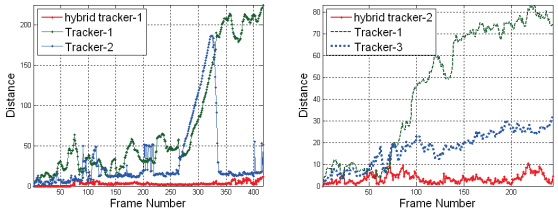

Fig. 7. Euclidean distance. Left: from the proposed hybrid tracker-1 (red), existing *tracker-1* (green), *tracker-2* (blue) on the video "woman walk"; Right: from the proposed hybrid tracker-2 (red), the existing *tracker-1* (green), *tracker-2* (blue) on the video "ThreePastShop2Cor".

tracker-1 is the best compromise in terms of tracking robustness and tracking speed (approx. 10 frames/sec in Matlab code).

7. REFERENCES

[1] G.Hager, P. Belhumeur, "Real-time tracking of image regions with changes in geometry and illumination", *Proc. CVPR*, 1996.

[2] Q.Zhao, S.Brennan, H.Tao, "Differential EMD tracking", In *PProc. ICCV*, 2007.

[3] D.Ross, J.Limy, R.Line, M. Yang, "Incremental learning for visual tracking", *IJCV*, 2007.

[4] S.Avidan, "Ensamble Tracking", *Proc. CVPR*, 2005.

[5] H.Zhou, Y.Yuan, C.Shi, "Kernel-Based method for tracking objects with rotation and translation", *IJCV*, 2008.

[6] A.Chen, M.Zhu, Y.Wang, C.Xue, "Mean shift tracking combining SIFT", in *Proc. ICALIP*, 2008.

[7] Z.Khan, I.Y.H. Gu, T.Wang, A.Backhouse, "Joint Anisotropic Mean Shift and Consensus Point Feature Correspondences for Object Tracking in Video", in *Proc. ICME*, 2009.

[8] C.Shan, Y.Wei, T.Tan, F.Ojardias, "Real time hand tracking by combining particle filtering and mean shift", *Proc. ICAFGR*, 2004.

[9] S. Zhong and F. Hao, "Hand Tracking by Particle Filtering with Elite Particles Mean Shift", in *Proc. IWFCST*, 2008.

[10] Z. Khan, I.Y.H. Gu, A.Backhouse, "Joint particle filters and multi-mode anisotropic mean shift for robust tracking of video objects with partitioned areas", *proc. ICIP*, 2009.

[11] Q.Sumin, H.Xianwu, "Hand tracking and gesture gecogniton by anisotropic kernel mean shift", in *Proc. ICNNSP*, 2008.

[12] S.Haner, I.Y.H.Gu, "Combining Foreground / Background Feature Points and Anisotropic Mean Shift For Enhanced Visual Object Tracking", *proc. of ICPR 2010.*

[13] D.G.Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *Int. Journal of Computer Vision*, vol. 60, pp. 91-110, 2004.

[14] M.A.Fischler, R.C.Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartographys", *Communications of the ACM*, 1981.