# Soft-Assigned Bag of Features Tracking

Zhongyan Qiu[1,3], Tong Yu[2], Tongwei Ren[1,2,*], Yan Liu[4], Jia Bei[1,2]

[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]Software Institute, Nanjing University, Nanjing, China
[3]School of Software, Nanchang University, Nanchang, China
[4]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

## ABSTRACT

Bag of features (BoF) provides an effective and efficient representation for object tracking in video sequences. However, hard assignment used in BoF generation inevitably brings in quantization errors, which may lead to inaccuracy even failure in tracking. In this paper, we propose a novel soft-assigned bag of features tracking approach (SABoF), which can significantly reduce the influence of quantization errors and obtain more accurate and stable tracking results. We initialize tracking by specifying the tracked object and constructing the codebook. Then, we represent each candidate target with soft-assigned BoF and measure its similarity to the tracked object. The most similar candidate target in each frame is selected as the tracked result. To improve tracking performance, we also refine the tracking results by combining incremental PCA tracking. The proposed approach is evaluated on the challenging video sequences from CAVIAR dataset. Experiments show our approach outperforms current dominant methods in complex conditions.

## Categories and Subject Descriptors

I.4.8 [**Scene Analysis**]: Tracking

## General Terms

Algorithms

## Keywords

Soft assignment, bag of features, object tracking

## 1. INTRODUCTION

Object tracking in video sequences has drawn much attention in research due to its wide applications, such as event surveillance, video indexing and human-computer interaction [6]. One of the essential problems in object tracking is what kind of feature should be used in object representation [5]. A good object representation can reduce the computation in feature extraction and similarity measurement and make accurate and stable tracking results.

Global features are utilized in the early tracking methods [8]. They require low computational cost, but usually fail in presence of partial or complete occlusion [15]. Recently, numerous local feature based tracking methods are proposed [7, 12]. They focus on local information of the tracked objects, which are useful in handling object occlusion, but require high computational cost in similarity measurement.

Bag of features (BoF) provides an effective and efficient approach to represent the tracked object with local features [3]. It assigns local features to several codewords, which are usually clustered from the sampled features, and represents the object as a collection of codewords. In this way, BoF not only depicts the local visual characters of the tracked object but also represents the object with a singular feature to simplify similarity measurement. Hence, BoF is widely used in object categorization, image retrieval, and it is first applied in object tracking by Yang *et al.* [15].

BoF tracking obtains good performance in complex conditions, such as object occlusion. Nevertheless, it hard assigns each local feature to a singular codeword, which inevitably brings in quantization errors. For the codewords are closer in tracked object representation than other applications, small differences caused by object occlusion or illumination variation may lead to two corresponding local features assigned to different codewords. Meanwhile, a local feature extracted from image noise is also assigned to its nearest codeword and further matched to other local features, even it is not similar to the codeword. For local features are treated completely same or totally different if assigned to the same codeword or not, hard assignment will cause the inaccuracy in similarity measurement, and even result in failures in BoF tracking.

To solve the above problem, we propose a novel soft-assigned bag of features tracking approach (SABoF). We utilize soft assignment strategy [10, 13], which assigns each local feature to several nearest codewords with different weights, and determines the assigned weights according to the distances from local features to the codewords. We generate the codewords and the candidate targets as original BoF tracking, and apply soft assignment in feature extraction for candidate target selection. It shows that the application of soft assignment can significantly reduce the influence of inaccuracy in similarity measurement, and obtain more stable tracking results.

[*]Corresponding author (Email: rentw@nju.edu.cn)

*ICIMCS'13,* Aug. 17–19, 2013, Huangshan, Anhui, China.
Copyright 2013 ACM 978-1-4503-2252-2/13/08 ...$15.00.

## 2. SOFT-ASSIGNED BOF TRACKING

We first initialize tracking with tracked object specification and codebook construction. Then object position in current frame is located by selecting the most similar candidate target to the tracked result in the previous frame using soft-assigned BoF. To improve tracking performance, codebook is continuously updated and tracking result is refined by combining incremental PCA tracking.

### 2.1 Initialization

Similar to original BoF tracking [15], we initialize tracking by manually specifying the tracked object in the first frame, and apply incremental PCA tracking in the first 5 frames [11] to collect sufficient training data. Within the manually specified object or each tracked result, we randomly select $N_P$ patches and extract a feature $f_i$ for each patch $p_i$. Though amounts of local features are effective for tracking, we use RGB descriptor and local binary pattern descriptor in our experiments [16], for comparing performance with original BoF tracking on the same features. Based on these features, the collected $5 \times N_P$ patches are quantized into $N_C$ clusters by approximate $k$-means algorithm [9]. The cluster centers are treated as the codewords to compose the initial codebook, and each codeword $c_j$ has a feature as above. With the codebook, the tracked object is represented as a bag of features, which is a histogram of the occurrence frequencies of its containing codewords.

### 2.2 Candidate target selection

When tracking the object in a new frame $F^t$, we select $N_T$ candidate targets around the position of tracked result in the previous frame $F^{t-1}$ using the predefined affine parameters. Within each candidate target, we randomly select $N_P$ patches and extract their features.

To represent the candidate targets as bags of codewords, a simple solution is hard assigning each patch $p_i$ to its nearest codeword $\hat{c}_i$:

$$\hat{c}_i = \arg\min_{c_j} \|\boldsymbol{f_{p_i}} - \boldsymbol{f_{c_j}}\|, j \in \{1, ..., N_C\}. \qquad (1)$$

where $\boldsymbol{f_{p_i}}$ and $\boldsymbol{f_{c_j}}$ are the features of patch $p_i$ and codeword $c_j$, respectively.

However, hard assignment inevitably brings in quantization errors and influences accuracy in tracking. For object occlusion, illumination variation and distortion are always included in surveillance video and other tracking sceneries, they may lead to small changes in feature values of two corresponding patches and assignment of these patches to different codewords. Furthermore, the patch containing noise will be assigned to its nearest codeword, no matter how dissimilar they are, and treated completely same to other patches similar to the codeword. Obviously, hard assignment may result in inaccuracy in candidate target selection and even failures in tracking.

Fig. 1 shows an example of hard assignment drawbacks. In the previous frame $F^{t-1}$, three patches in the tracked result (red) are assigned to codeword A, B and C respectively. After the patches are selected from the candidate targets in current frame $F^t$, they are quantized by codebook. Though the patches in the ground truth in frame $F^t$ (orange) are similar to the patches in the corresponding positions within the tracked result in frame $F^{t-1}$, they are quantized to codeword C and D for quantization errors. Meanwhile, the patches selected from another candidate target in frame $F^t$ (green) are quantized to codeword A, B and C, though they have larger distances to the patches within the tracked result in frame $F^{t-1}$. Especially, the green patch in lower left corner is far away from all the codewords, but it is still assigned to codeword B. In this way, the green candidate target will be closer to the tracked result in BoF similarity measurement than the ground truth, even it is less similar in appearance.

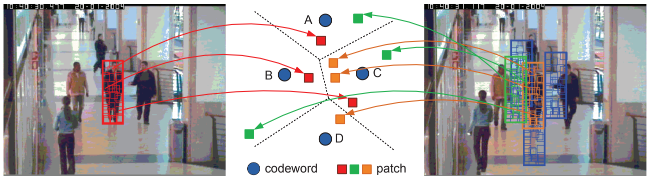

**Figure 1: Example of hard assignment drawbacks. (left) Previous frame $F^{t-1}$. (middle) Hard assignment for patch quantization. (right) Current frame $F^t$.**

To overcome the drawbacks, we utilize soft assignment [10] in our approach. Different to hard assignment, soft assignment allows a patch assigned to several nearest codewords with different weights. As is usual in soft assignment, we use an exponential function of the distance between patch and codeword to calculate the assigned weight:

$$w(p_i, c_j) = \exp\left(-\frac{\|\boldsymbol{f_{p_i}} - \boldsymbol{f_{c_j}}\|^2}{\sigma^2}\right). \qquad (2)$$

where $\sigma$ is a parameter to adjust the weight values, which is usually influenced by the distances between codewords.

Each patch is assigned to the nearest $r$ codewords, and the occurrence frequency of each codeword in a candidate target is calculated as the sum of assigned weights from all the patches in the candidate target. In this way, a candidate target is represented as a histogram of the occurrence frequencies of codewords:

$$\boldsymbol{h_{T_k^t}} = \left[\sum\nolimits_i w(p_i^{T_k^t}, c_1), ..., \sum\nolimits_i w(p_i^{T_k^t}, c_{N_C})\right]. \qquad (3)$$

where $T_k^t$ is the $k$th candidate target in frame $F^t$; $p_i^{T_k^t}$ is the $i$th patch in candidate target $T_k^t$.

We measure similarities between the candidate targets in frame $F^t$ and the tracked result in frame $F^{t-1}$ based on BoF distances, and select the most similar candidate target as the tracked result in frame $F^t$:

$$\hat{T}^t = \arg\min_{T_k^t} \|\boldsymbol{h_{\hat{T}^{t-1}}} - \boldsymbol{h_{T_k^t}}\|, k \in \{1, ..., N_T\}. \qquad (4)$$

where $\hat{T}^{t-1}$ and $\hat{T}^t$ are the tracked results in frame $F^{t-1}$ and $F^t$; $\boldsymbol{h_{\hat{T}^{t-1}}}$ and $\boldsymbol{h_{T_k^t}}$ are BoF features of traced result $\hat{T}^{t-1}$ and candidate target $T_k^t$.

### 2.3 Codebook update and result refinement

For the appearance of tracked object continuously changes in video sequence, we update the codebook per $N_F$ frames processed. The new codewords are generated using approximate $k$-means clustering on both the new coming $N_F \times N_T$ patches and the previous codewords. After the codebook is

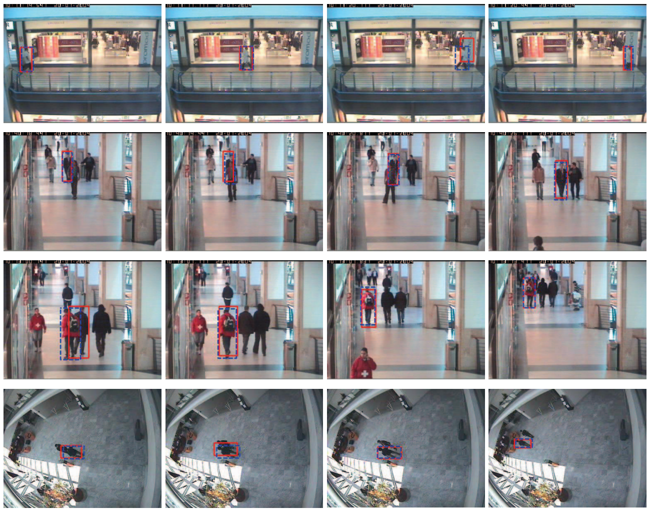

**Figure 2: Tracking results of our approach. (first row) Tracking Result in the frames 15, 94, 169 and 175 of sequence *OneStopEnter1front*. (second row) Tracking result in the frames 170, 196, 216 and 347 of sequence *ShopAssistant2cor*. (third row) Tracking result in the frames 69, 78, 245 and 397 of sequence *ThreePastShop2cor*. (fourth row) Tracking result in the frames 79, 106, 110 and 196 of sequence *Meet_WalkSplit*.**

updated, BoF representation of the tracked object should be regenerated according to the new codebook.

We adopt incremental PCA tracking to refine tracking results [2] as original BoF tracking. If the distance between tracked object and the most similar candidate target is larger than a predefined threshold, the affine parameters for candidate target selection, including central points, size and rotation angle, are adjusted by combining the parameters of incremental PCA tracking with a weight $\alpha$.

## 3. EXPERIMENTS

We first discuss parameter selection in soft-assigned BoF tracking. Then, we show the qualitative tracking results of our approach on four challenging video sequences from CAVIAR dataset. To demonstrate the advantages of our approach, we compare our approach with five dominant tracking methods with quantitative evaluation.

### 3.1 Parameter selection

Several key parameters are used in soft-assigned BoF tracking, which influence the tracking performance. To demonstrate the advantage of soft assignment, we use the same parameter setting to original BoF tracking [15]. In our experiments, we select $N_T = 300$ candidate targets in each frame, and randomly select $N_P = 50$ patches with the size of 12×12 pixels in each candidate target. The number of codewords is constantly 20 in tracking procedure. The codebook is updated when processing per $N_F = 5$ frames, and the combination weight $\alpha$ in result refinement equals 0.7. In soft assignment, each patch is assigned to the nearest $r = 3$ codewords, and $\sigma$ in assinged weight calculation is set 1/9, for the distances between codewords are small in tracked object representation.

**Table 1: Comparison in average position errors on five video sequences.**

| | Frag | IVT | NBS | TLD | BoF | SABoF |
|---|---|---|---|---|---|---|
| *OneStopEnter1front* | 179.63 | 180.81 | 179.11 | 175.63 | 13.22 | **12.65** |
| *ShopAssistant2cor* | 7.87 | 7.02 | 10.47 | 16.20 | 4.76 | **3.43** |
| *ThreePastShop2cor* | 26.81 | 44.05 | 32.92 | 57.31 | 5.68 | **4.15** |
| *Meet_WalkSplit* | 10.35 | 18.45 | 35.12 | 20.00 | 5.64 | **5.25** |

### 3.2 Tracking results

We evaluate the proposed approach on four challenging video sequences from CAVIAR dataset, which includes video clips in the different scenarios of interest with hand-labeled tracking ground truths [2]. Four selected video sequences are *OneStopEnter1front*, *ShopAssistant2cor*, *ThreePastShop2cor* and *Meet_WalkSplit*. They have significant object occlusion and obvious illuminance variation.

Fig. 2 illustrates the tracking results of our approach on these four video sequences. The blue boxes indicate hand-labeled ground truths, and the red boxes are the tracked results of our approach.

In sequence *OneStopEnter1front*, the tracked person passes several regions with various illumination and a signboard with similar color to his cloths. It shows that our approach can handle obvious illuminance variation (frame 15 and 94), and quickly adjust the tracked result to accurate position (frame 175) after the disturbance of signboard (frame 169).

In sequence *ShopAssistant2cor*, the tracked person is seriously occluded by another person in walking. It shows that our approach can stably track the person though he is occluded (frame 170, 196 and 216), and handle illumination variation well (frame 347).

In sequence *ThreePastShop2cor*, the tracked person walks with two person, and passes a person with similar color cloths. It shows that our approach can rectify the position error in person interference (frame 69 and 78), prevent the influence of the person with similar color cloths, and provide accurate tracking results when illumination varies (frame 245) and object scale changes (frame 397).

In sequence *Meet_WalkSplit*, the tracked person meets with another person with similar color cloths, and splits from the person after walking together for a while. It shows that our approach can keep stable tracking when the two persons meet (frame 79, 106 and 110), and track the person accurately when his scale changes (frame 196).

To quantitatively evaluate the performance, we compare our approach with five dominant tracking methods: robust fragments-based tracking (Frag) [1], incremental PCA tracking (IVT)[11], discriminative nonorthogonal binary subspace tracking (NBS) [5], tracking-learning-detection (TLD) [4] and original BoF tracking (BoF) [15]. The position error is measured as the centroid distance from the tracked result to the ground truth in pixels. Fig. 3 illustrates a comparison of our approach with other five tracking methods in position error on each frame. It shows that our approach obtains accurate tracking results on all the video sequences and it is more stable than original BoF tracking. Table. 1 presents the average position errors of these five methods on each video sequence. It shows that our approach can achieve more accurate tracking results

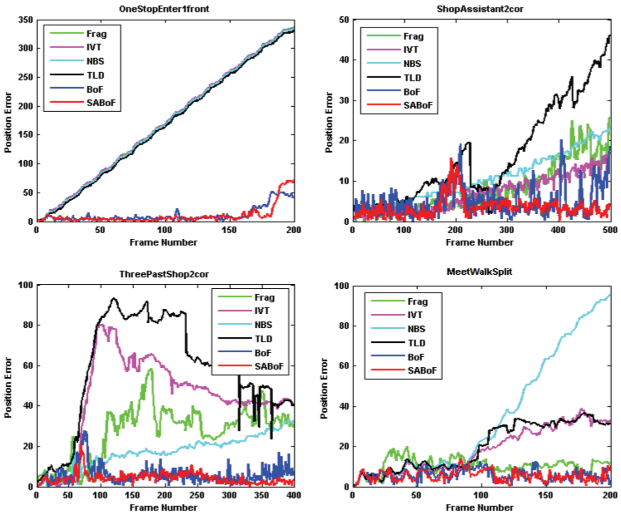

**Figure 3: Comparison in pixel error for each frame. The horizontal axis is frame number, and the vertical axis is position error. (top left) Position errors on *OneStopEnter1front*. (top right) Position errors on *ShopAssistant2cor*. (lower left) Position errors on *ThreePastShop2cor*. (lower right) Position errors on *Meet_WalkSplit*.**

under object occlusion, illumination variation and other challenging conditions.

### 3.3 Discussion

There has been lots of prior work on object tracking or bag of features. Due to the page limitation, we only survey the work related to our approach here.

One kind of related work is the previous object tracking methods, especially the methods use local features in object representation. In experiment section, we have compared the proposed approach with five dominant object tracking methods on the challenging video sequences from CAVIAR dataset. It showed that our approach could obtain more accurate and stable tracking results in complex conditions. Among these compared methods, BoF tracking proposed by Yang et al. [15] is the most similar method to our approach, which first applied BoF in object tracking and also obtained acceptable results. But BoF tracking used hard assignment in similarity measurement, which inevitably brought in quantization errors and led to inaccuracy in tracking. More details can be found in Fig. 3 and Table 1.

Another kind of related work is other BoF applications using soft assignment, such as object categorization [14] and image retrieval [10]. But different applications require different soft assignment strategies. For example, Wang et al. realized soft measurement with Gaussian Mixture model [14], but Philbin et al. [10] and we utilized the strategy of assigning each local feature to multiple codewords. Furthermore, our approach also adopted the completely different features, codebook scale and parameters to image retrieval [10] for the intrinsical requirements of object tracking.

## 4. CONCLUSION

We proposed soft-assigned bag of features tracking approach to reduce the influence of quantization errors in BoF tracking. The proposed approach is evaluated on the challenging video sequences with object occlusion, illumination variation, distortion and other complex conditions, and compared with current dominant tracking methods. It shows that out approach obtains more accurate and stable tracking results.

In the future, we would like to adopt soft assignment in different tracking methods, and apply soft-assigned BoF tracking in video indexing and other applications.

## 5. ACKNOWLEDGMENTS

This paper is supported by Natural Science Foundation of China (61202320), and Natural Science Foundation of Jiangsu Province (BK2012304).

## 6. REFERENCES

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *CVPR*, pages 798–805, 2006.

[2] CAVIAR. http://homepages.inf.ed.ac.uk/rbf/caviar/.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop SLCV*, pages 1–22, 2004.

[4] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking learning detection. *TPAMI*, 34(7):1409–1422, 2012.

[5] A. Li, F. Tang, Y. Guo, and H. Tao. Discriminative nonorthogonal binary subspace tracking. In *ECCV*, pages 258–271, 2010.

[6] A. Maalouf and M. Larabi. Robust foveal wavelet-based object tracking. In *ICASSP*, pages 1489–1492, 2012.

[7] X. Mei and H. Ling. Robust visual tracking using $l_1$ minimization. In *ICCV*, pages 1436–1443, 2009.

[8] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV*, pages 661–675, 2002.

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8, 2007.

[10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, pages 1–8, 2008.

[11] D. Ross, J. Lim, R. Lin, and M. Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1-3):125–141, 2008.

[12] D. Serby, E. Meier, and L. Gool. Probabilistic object tracking using multiple features. In *ICPR*, volume 2, pages 184–187, 2004.

[13] J. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders. Kernel codebooks for scene categorization. In *ECCV*, pages 696–709, 2008.

[14] Y. Wang, X. Liu, and Y. Jia. Soft measure of visual token occurrences for object categorization. *CAIP*, 5702:774–782, 2009.

[15] F. Yang, H. Lu, and Y. Chen. Bag of features tracking. In *ICPR*, pages 153–156, 2010.

[16] F. Yang, H. Lu, W. Zhang, and G. Yang. Visual tracking via bag of features. *IET Image Processing*, 6(2):115–128, 2012.