Scale adaptive region covariance descriptor for visual tracking

Mohd Fauzi Abu Hassan1,4 Azurahisham Sah Pri1 Zakiah Ahmad2 Tengku Mohd Azahar Tuan Dir3

1Electrical, Electronic and Automation Section, Malaysian Spanish Institute, Universiti Kuala Lumpur, Kulim Hi-Tech Park, 09000, Kulim, Kedah, Malaysia
2Technical Foundation Section, Malaysian Spanish Institute, Universiti Kuala Lumpur, Kulim Hi-Tech Park, 09000, Kulim, Kedah, Malaysia
3Industrial Automation Section, Malaysian France Institute, Universiti Kuala Lumpur, 43650 Bandar Baru Bangi, Selangor. Malaysia

E-mail: mohdfauzzi@unikl.edu.my

Abstract. This paper presents an adaptive approach for scale estimation in a tracking-by-detection framework. The proposed method works by learning covariance descriptor based on multi-layer instance search region. Our results show that the proposed approach significantly improves the performance in term of detection rate compared to region covariance descriptor with using a fixed bounding box (single scale). From this work, it is believed that we have constructed a greater solution in choosing best layer for this descriptor, permitting to move forward to the next issues such as fast motion or motion blur for achieving a robust tracking system.

1. Introduction

Visual tracking is an important topic in computer vision application. The performance of the tracking process depends on some issues such as scale variation, motion blur, in-plane rotation, out of plane rotation, fast motion etc. Scale estimation is a challenging problem in visual tracking. Most existing approaches provide inferior performance when they encountered with scale variations in complex image sequences [1], [2], [3], [4], [5], [6]. Region covariance descriptor has shown some potentials based on the result of previous resources [12], [13]. Visual tracking problem arises when using region covariance descriptor (hereinafter abbreviated as “Cov tracker”) tracking the target moves in scale variation. This is because it does not consider the scale of the target, thus the position of the target cannot be estimated accurately. By using a fixed bounding box (single scale), it is hard to adjust to the changeable object scale. It will lose object information, or it will introduce background information.

To overcome the limitations of a rigid bounding box, we present a novel tracking-by-detection approach which is based on the Cov tracker.

4 To whom any correspondence should be addressed.
In this work, a new approach for estimating the position and scale of a single object is shown in Figure 1. For region appearance representation, we propose an enhanced Cov tracker, called Multi-layer Instance Search region for Online Localization model (hereinafter abbreviated as “MISOL”). The proposed scale estimation method (MISOL) is described in Section 3.

2. Region Covariance Descriptor
For the appearance model, we use the region covariance descriptor [12] as the observation model. The region covariance descriptor can describe the object more accurately by combining different types of spatial and temporal features naturally. Since the dimension of the covariance matrix depends only on the number of the features we used, the problem of high calculation cost in high dimension is avoided. In this work, we select the feature vector as

\[ F(i) = [f_1, f_2, \ldots, f_n] \quad (1) \]

where \( \{f_i\}^n \) are \( d \) dimensional feature column vectors associated with the pixel index \( i \), and \( n \) is the number of total pixels in the image. Feature vectors generated for each pixel. Then the covariance matrix is calculated by

\[ C_r = \frac{1}{MN} \sum_{k=1}^{MN} (f_k - \mu_r)(f_k - \mu_r)^T \quad (2) \]

where \( r \) indicates the rectangle region with the scale: \( MN \), and \( \mu_r \) is the mean of all vectors in the region \( r \). In this way, the region \( r \) is represented as a \( MN \) covariance matrix. Since covariance matrices lie on the Riemannian space rather than the Euclidean space, the Bhattacharyya distance, which is often used for histogram distance metric, cannot be used to measure the distance of covariance matrices. Therefore, the dissimilarity between two covariance matrices is proposed in [12] and calculated by

\[ \rho(C_i, C_j) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k(C_i, C_j)} \quad (3) \]

where \( C_i \) and \( C_j \) are covariance matrices of the target model and the candidate particle, respectively; \( \lambda_k \) is the generalized eigenvalue between \( C_i \) and \( C_j \). Lastly, find target region with minimum distance [12] as the matching region.

\[ \rho_{\text{min}} = \arg \min \rho(C_i, C_j) \quad (4) \]

where \( \rho_{\text{min}} \) is the covariance matrix distance.

3. Proposed multi-layer instance search region for online localization model
Multi-layer approach is used for searching area. After the position estimation, the scale of target is estimated by using multi-layer which is similar to these methods [7], [8], [9] introduced. The basic pipeline of our algorithm of position and scale estimation are shown in Figure 1. The key idea of our method is the same as Cov tracker, but we introduce sampling strategies for feature extraction called MISOL. The target is modelled hierarchically at multi-layer instance (e.g. take 4 layers for example [14]). Following the logic of the model construction and the order of model updating, this section first
introduces the bottom layer model, since the initial features are extracted from this layer. Then, we introduce the bottom layer which is followed by the next top layer. Each layer with minimum distance, \( \rho_{\text{min}} \) [12] as the matching region respectively. After that comparison between two \( \rho_{\text{min}} \) (e.g. layer 1 and layer 2) to find the best matched distance (minimum distance).

**Figure 1.** Pipeline of our algorithm.

### 4. Subjects and method

In this section, it consists of the proposed method, testing image sequences and experiments setting.

#### 4.1. Proposed method

In this section, we first revisit the Cov tracker in Section 2 and the proposed scale estimation method (MISOL) is described in Section 3. The proposed tracking algorithm is tested (four different layers) challenging image sequences from [10]. A bounding box, which includes the coordinates \((x, y)\) of the upper-left corner and the box size (width and height), is used to representing the tracking result in that frame. Following the conventions of [10], our method initialises its target model by using only a given bounding box in the first frame, without any prior information on the target category. Figure 2 shows the flow chart of MISOL algorithm. The proposed method was developed based on the covariance tracking framework.

**Figure 2.** Flowchart of the proposed MISOL algorithm
4.2. Testing image sequences
To evaluate the performance of our proposed MISOL algorithm, we conduct experiments on four challenging image sequences (datasets). The basic information of testing image sequences is the image sequences name, number of frames and the factors that affect the performance of tracking have been listed in Table 1. The total number of all frames is 2711. On these sequences, the object appearances are varied by several factors, including fast motion, out of plane rotation, occlusion, scale variation, background clutters, deformation, low resolution, background clutters, etc. The factors like the scale variation is the main emphasis that we will highlight in this paper.

Table 1. Details and challenges of image sequences in our experiments

| Image sequences | Number of frames | Main challenging factors                           |
|-----------------|------------------|---------------------------------------------------|
| Bolt            | 350              | Occlusion, deformation, in-plane rotation, **out of plane rotation** |
| Walking2        | 500              | **Scale variation**, occlusion, low resolution     |
| Crossing        | 120              | **Scale variation, out of plane rotation**, background clutters, deformation, fast motion |
| Liquor          | 1741             | **Scale variation**, motion blur, **out of plane rotation**, occlusion, background clutters, fast motion, out-of-view |

4.3. Experiments setting
For each image sequences, the location of the target object is manually labelled in the first frame. This means that in the first frame, the target object is initialized manually by a bounding box with prior knowledge. The initial parameters of proposed algorithm are set as: search region, \( d = 20 \), search window, \( e = 5 \) and scaling factor, \( sf = 0.9 \) as shown in Figure 3.

![Figure 3. Initial parameters setting](image-url)
In this work, we select the feature vector as

$$\mathbf{F}(I) = [R \ G \ B \ V \ H \ E]$$  \hspace{1cm} (5)

where $R$, $G$, $B$ represent the red, green and blue values of the pixel, $V$, $H$, $E$ represent the vertical, horizontal and edge values of the grey value. In this way, the region is represented as a 6 x 6 covariance matrix.

5. Results and analysis

Figure 4 till Figure 7 shows the tracking precision for a range of distance thresholds, and the percentages of frames that the tracker is within that distance of the ground truth. In general, our methods (Layer 1, 2, 3 and Layer 4) obtain more precise results on the four challenging sequences compare to Cov tracker. Table 2 summarizes these plots, chose the threshold 20 and report the precision at this point in the curve (e.g. this is the percent of frames for which the tracker was less than 20 pixels from the ground truth). This threshold is used by [11].

The result shows that the proposed method has improved the tracking precision although the results are different. The difference between our methods with Cov tracker can been seen due to the increase of 0.62 (Layer 1), 0.65 (Layer 2), 0.65 (Layer 3) and 0.64 (Layer 4). This is caused by the introduction of multi-scale (e.g. MISOL) as compared to the single scale. In the last row, which is labelled as average, our method (Layer 2 & 3) algorithm obtains the highest average as compared to the others. The second-best result is presented by our method (Layer 4) as compared to the other trackers.

![Figure 4. Precisions plots for Bolt image sequence](image-url)
Figure 5. Precisions plots for Walking2 image sequences

Figure 6. Precisions plots for Crossing image sequences
Figure 7. Precisions plots for Liquor image sequences

Table 2. Tracker precision at a threshold of 20 (percentage of frames where the predicted location is within 20 pixels of the ground truth)

| Image Sequences | Cov  | Layer 1 | Layer 2 | Layer 3 | Layer 4 |
|-----------------|------|---------|---------|---------|---------|
| Bolt            | 0.64 | 0.75    | 0.74    | 0.70    | 0.64    |
| Walking2        | 0.10 | 0.39    | 0.53    | 0.58    | 0.59    |
| Crossing        | 0.83 | 0.98    | 0.98    | 0.99    | 0.96    |
| Liquor          | 0.28 | 0.59    | 0.36    | 0.31    | 0.36    |
| Average         | **0.46** | **0.62** | **0.65** | **0.65** | **0.64** |

6. Discussion

The proposed MISOL algorithm has successfully improved the tracking performance compared to the Cov tracker. The overall results of the proposed algorithm show an improvement in terms of tracking performance details as tracker precision has increased. This shows that by applying the new approach of a search region for online localization, the tracker could be adequately improved especially in addressing scale variation issue. The result shows that the value of tracker precision has increased, MISOL with Layer 2 and Layer 3 performs the best compared to the Cov tracker and others. The result shows that the consequences of the new technique in MISOL algorithm are the value of tracker precision has increased as compared to the Cov tracker. The main problem with visual tracking is drifting was reduced.

7. Conclusion

In this paper, we propose a region covariance descriptor tracking method with MISOL. Our tracker locates the target by adopting the multi-layer instance search region for online localization to realize a more accurate tracking method. It has been confirmed by experiments and evaluations that the proposed MISOL algorithm presents better performance compared with Cov tracker without the scale. However, the MISOL algorithm has a few limitations as identified by the experiment results. Based on the performance results, there are no right layer that can be selected for all image sequences especially
object in fast motion or motion blur. In the next future research, a new process / technique will be proposed to ensure that our tracker performances are improved specially to solve scale variation with the fast motion or motion blur issues.

References

[1] M. Danelljan, G. Häger, and M. Felsberg 2014 Accurate scale estimation for robust visual tracking *Proc. Br. Mach. Vis. Conf.*, p. 2014

[2] E. Ohn-bar and M. M. Trivedi 2016 Multi-scale volumes for deep object detection and localization *Pattern Recognit.*, pp. 1–16

[3] X. Wang, Z. Hou, W. Yu, Z. Jin, Y. Zha, and X. Qin 2017 On Multilayer convolutional features pp. 1–13

[4] C. Zheng, W. Jin, F. Fang, C. Tang, and Y. Ling 2016 Robust visual tracking algorithm based on structural multi-scale features adaptive fusion in co-training *Proc. - 2016 3rd Int. Conf. Inf. Sci. Control Eng. ICISCE 2016*, pp. 588–592

[5] Z. He, Y. Fan, J. Zhuang, Y. Dong, and H. Bai 2018 Correlation filters with weighted convolution responses *Proc. - 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017*, vol. 2018-Janua, pp. 1992–2000

[6] M. Cen and C. Jung 2017 Complex form of local orientation plane for visual object tracking *IEEE Access*, vol. 5, pp. 21597–21604

[7] K. Chaudhary, M. Zhao, F. Shi, X. Chen, K. Okada, and M. Inaba 2017 Robust real-time visual tracking using dual-frame deep comparison network integrated with correlation filters *IEEE Int. Conf. Intell. Robot. Syst.*, vol. 2017-Sept, pp. 6837–6842

[8] K. Chen, W. Tao, and S. Han 2017 Visual object tracking via enhanced structural correlation filter *Inf. Sci. (Ny)*., vol. 0, pp. 1–14

[9] C. Ma, J. Bin Huang, X. Yang, and M. H. Yang 2015 Hierarchical convolutional features for visual tracking *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 3074–3082

[10] Y. Wu, J. Lim, and M.-H. Yang 2013 Object tracking online object tracking: a benchmark 2013 *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2411–2418

[11] B. Babenko, S. Member, M. Yang, and S. Member 2011 Robust object tracking with online multiple instance learning robust object tracking with online multiple instance learning Pami

[12] O. Tuzel, F. Porikli, and P. Meer 2006 Region covariance: A fast descriptor for detection and classification. In *Proc. 9th European Conf. on Computer Vision, volume 2*, pp. 589-600.

[13] P. Cortez-Cargill, C. Undurraga-Rius, D. Mery-Quiroz, and A. Soto 2009 Performance evaluation of the covariance descriptor for target detection. In *International Conference of the Chilean Computer Science Society*, pp 133-141.

[14] Fu Li, Xu Jia, Cheng Xiang and Huchuan Lu 2017 Visual tracking with structured patch-based model *Image Vision Computing* vol. 0, pp. 1–10