Visual tracking using multi-layer appearance approach

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

¹Electrical, Electronic and Automation Section, Malaysian Spanish Institute, Universiti Kuala Lumpur, Kulim Hi-Tech Park, 09000, Kulim, Kedah, Malaysia
²Engineering 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: mohdfauzi@unikl.edu.my

Abstract. This paper investigated single target tracking of arbitrary objects. Tracking is a difficult problem due to a variety of challenges such as scale variation, motion, background clutter, illumination etc. To achieve better tracking performance under these severe conditions, this paper proposed covariance descriptor based on multi-layer instance search region. Our results show that the proposed approach significantly improves the performance in term of centre location error (in pixels) compared to covariance descriptor with using a fixed bounding box. From this work, it is believed that we have constructed a great solution in choosing best layer for this descriptor. This will be addressed in the next future work such as consider target motion during tracking.

1. Introduction

Visual tracking has been one of the most important components in computer vision with many applications such as surveillance, robotics, interactive visual media editing, intelligent monitoring, human–computer interaction, augmented reality, autonomous and so on. Scale variation is a one of challenging problem in visual tracking. Recently, covariance descriptor, that has been widely used in recognition [1] and detection [2], had fascinated boundless devotion in object tracking thanks to their promising precision and high speed capability. However, these methods still suffer from problem of scale variation due to moving object are not same in size in every frame [3-7]. It is difficult to adjust to the changeable object scale by using a fixed bounding box during tracking. Figure 1 shows a new approach to overcome this problem. For appearance representation, we propose an improved covariance descriptor explained in next section.

2. MISOL Tracker

The object scale is estimated by using multi-layer approach which is similar to these methods [8-10] introduced. The object is modelled as multi-layer (e.g., four different size layers) [11]. For region appearance representation, propose an improved Cov tracker, called Multi-layer Instance Search region for Online Localization (MISOL) approach. The overview of our proposed flowchart is illustrated in Figure 1.
There are four (4) layers used in the proposed model. It means that the size of the bounding box is increase 10 pixels on each layer upwards. Table 1 shows parameter values used in multi-layer representations. The target is modelled hierarchically at a multi-layer instance (e.g., take 4 layers for example). This model presents the bottom layer model, and the initial features are selected from this layer. Then, it presents the bottom layer which is continued by the next top layer. Each layer comes with a minimum distance, \( \rho_{\text{min}} \) as the matching region, respectively. After that, a comparison is drawn between two \( \rho_{\text{mins}} \) (e.g., layer 1 and layer 2) to find the best matched distance (minimum distance).

**Table 1. Parameter values used in multi-layer representations**

| Layer | Scaling Factor | Pixels |
|-------|----------------|--------|
| 1     | 0.9            | -10    |
| 2     | 1.0            | 0      |
| 3     | 1.1            | 10     |
| 4     | 1.2            | 20     |

![Flow chart of MISOL](image-url)

**Figure 1.** Flow chart of MISOL
3. Subjects and Method

3.1. Proposed method
The proposed model (e.g., take 4 layers) is tested with covariance tracker (Cov) without the scale and MISOL on challenging video sequences from [12]. Given one ground truth object annotation in the first frame, the develop algorithms to track the target object in the remaining frames. As there is no prior knowledge of the target object except for the first frame annotation. Then run Cov and the proposed tracker [12]. Figure 2 shows the overview of the proposed block diagram tracking method.

![Block diagram of proposed method](image)

**Figure 2.** Block diagram of proposed method

3.2. Testing video sequences
There are four (4) challenging video sequences are used in this experiment to evaluate the tracking performance. All these datasets are explained in Table 2.

| Video sequences | No. of frames | Challenging factors |
|-----------------|--------------|---------------------|
| Bolt            | 350          | In-plane rotation, deformation, occlusion, out of plane rotation |
| Walking2        | 500          | Occlusion, scale variation, low resolution |
| Crossing        | 120          | Background clutters, scale variation, fast motion etc |
| Liquor          | 1741         | Out-of-view, fast motion, occlusion, motion blur background clutters, scale variation |

3.3. Experiments setting
The inclusive framework of the proposed tracker is: (a) Object initialization; (b) Appearance (MISOL) model; (c) Motion estimation model; (d) Object localization. Firstly, use experiment setting. The parameters of the proposed model are set as: search window, \( e = 5 \) and search region, \( d = 20 \). Hence, the significant of the proposed tracker to use layer value of 3 with the scaling factor, \( sf = 0.9 \). Selected feature vector \( F(i) = [ R \ G \ B \ V \ H \ E ] \). \( R, G \) and \( B \) are red, green, and blue values of the
pixel. \( V \) and \( H \) are image intensity first derivative in vertical and horizontal. \( E \) is absolute value vertical and horizontal. All parameters involved in the proposed model are chosen empirically [11].

4. Results and Analysis

This section, we test the performance of Cov tracker on sequences from CVPR2013 tracking benchmark datasets [12]. These datasets are currently used as the state-of-the-art benchmarks in the 2D visual object tracking community. Moreover, Figure 4 shows the centre location error (in pixels) [13] versus the number of frames for each sequence. It is as a displacement of target centre from ground truth for four (4) video sequences (Bolt, Walking2, Crossing, Liquor) for Cov tracker and our methods (e.g., Layer 1, 2, 3 and Layer 4). In the four (4) video sequences, our methods (tracker) can achieve more better and stable results than the Cov tracker. Table 3 shows the mean centre location error (in pixels) for four (4) video sequences by Cov tracker and our methods. Based on the result of the average, it shows that the method has successfully reduce centre location error to 24.43% (Layer 1), 24.50% (Layer 2), 24.60% (Layer 3) and 24.40% (Layer 4). The main reason is caused by multi-layer (scale) compares to the single scale.

In Bolt sequence, the Cov tracker almost loses the target during tracking. However, our proposed methods have better tracking performance with lower centre location error values. In Walking2 sequence, the Cov tracker gradually generates larger centre location error. Our methods have better tracking performance with lower centre location error values except for Layer 1 where it loses the target for several frames during tracking. In Crossing sequence, the Cov and Layer 1, 2, 3 and Layer 4 have better tracking performance with lower centre location error values. In Liquor sequence, long term tracking capability is needed. Hence, in this sequence, Cov tracker gradually generates larger centre location error, but our trackers show better tracking results. However, our method (Layer 4) almost loses the target in several frames during tracking.
Table 3. Centre location error (in pixels) for four (4) video sequences for Cov tracker and proposed method

| Video Sequences | Cov tracker | MISOL tracker |
|-----------------|-------------|---------------|
|                 | Layer 1     | Layer 2       | Layer 3       | Layer 4       |
| Bolt            | 20.75       | 18.94         | 19.49         | 20.75         |
| Walking2        | 82.01       | 34.17         | 31.96         | 31.38         |
| Crossing        | 17.65       | 12.71         | 12.55         | 12.65         |
| Liquor          | 82.73       | 39.35         | 40.77         | 40.77         |
| Average         | 50.79       | 26.29         | 26.19         | 26.39         |

Figure 4. Centre location error plot for Cov and MISOL tracker

5. Discussion
The MISOL tracker has achieved more better and stable results in all four (4) video sequences compared to Cov tracker. In all four (4) video sequences, the proposed tracker uses multi-layer instance search region approach to tackle problem of moving object are not identical in size in every frame. The comparison of pixels values of the mean centre location error for the video sequences is shown in Table 3. In the last row, which is the Average, our method (Layer 3) obtains the lowest average centre location error among the other trackers. In general, our method (Layer 3) obtains more stable and better results in the four (4) challenging sequences. Based on the average, the result shows that it has successfully reduce center location error compared with Cov tracker. This is caused by introducing MISOL component. In general, MISOL model has overcome scale issue.
6. Conclusions
The proposed MISOL method succeeded in improving tracking performance as compared to the Cov tracker. MISOL tracker performs better than Cov tracker. Although the qualitative evaluation shows a noteworthy improvement of the proposed model, a few limitations that have been identified from the experiment results especially in the issue of dynamic scale variation. In the next work, a new improvement will be proposed to make the proposed tracker can improve scale variation due to motion during tracking.

References
[1] Hariri W, Tabia H, Farah N, Benouareth A and Declercq D 2016 3D face recognition using covariance based descriptors Pattern Recognit. Lett. vol. 78 pp. 1–3
[2] Porikli F, Tuzel O and Meer P 2006 Covariance tracking using model update based on means on Riemannian manifolds Proc. IEEE Conf. Comput. Vis. Pattern Recognit. vol. 1 pp. 728–735
[3] Danelljan M, Häger G and Felsberg M 2014 Accurate scale estimation for robust visual tracking Proc. Br. Mach. Vis. Conf.
[4] Wang X, Hou Z, Yu W, Jin Z, Zha Y and Qin X 2017 On Multilayer convolutional features pp. 1–13
[5] Zheng C, Jin W, Fang F, Tang C and Ling Y 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. pp. 588–592
[6] He Z, Fan Y, Zhuang J, Dong Y and Bai H 2018 Correlation filters with weighted convolution responses Proc. 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017 pp. 1992–2000
[7] Cen M and Jung C 2017 Complex form of local orientation plane for visual object tracking IEEE Access vol. 5 pp. 21597–21604
[8] Chaudhary K, Zhao M, Shi F, Chen X, Okada K and Inaba M 2017 Robust real-time visual tracking using dual-frame deep comparison network integrated with correlation filters IEEE Int. Conf. Intell. Robot. Syst. pp. 6837–6842
[9] Chen K, Tao W and Han S 2017 Visual object tracking via enhanced structural correlation filter Inf. Sci. (Ny). vol. 0 pp. 1–14
[10] Ma C, Bin Huang J, Yang X and Yang M. H. 2015 Hierarchical convolutional features for visual tracking Proc. IEEE Int. Conf. Comput. Vis. pp. 3074–3082
[11] Abu Hassan MF, Sah Pri A, Ahmad Z and Tuan Dir TMA 2020 Scale adaptive region covariance descriptor for visual tracking IOP Conference Series Materials Science and Engineering 932:012090
[12] Wu Y, Lim J and Yang MH 2013 Object tracking online object tracking: a benchmark IEEE Conf. Comput. Vis. Pattern Recognit. pp. 2411–2418
[13] Babenko B, Member S, Yang M and Member S 2011 Robust object tracking with online multiple instance learning robust object tracking with online multiple instance learning PAMI