One-Stream Framework for Tracking with Template Update Algorithm

MUHAMMED ZEYN [1]

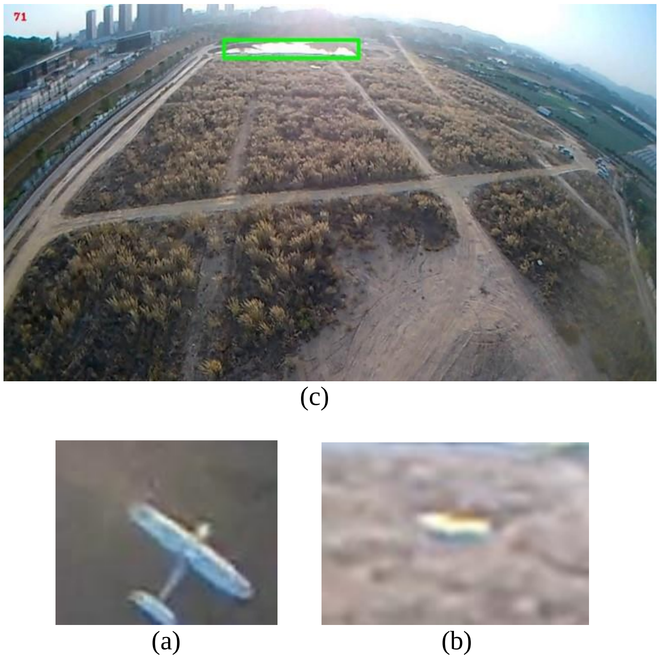
[1] Department of Mechatronics Engineering, Yildiz Technical University, Besiktas, Istanbul, Turkey
muhammed.zeyn@std.yildiz.edu.tr

Abstract

The combination of feature learning and relation model [1] in visual object tracking showed tremendous impact on the tracker's performance. Since then, most of the state-of-the-art trackers adapted this method to enhance the performance. Despite the high performance of the trackers, it still has poor accuracy on long term tracking, the idea behind that is the model fails to track the object after deformation on the appearance of the object. In this study we propose a new template update mechanism that feeds the model with the deformation of the object during long term tracking.

I. Introduction

The main goal of this work is to enhance the performance of the proposed tracker in long term tracking. Figure 1. shows the change of the appearance during tracking process on video from UAV2UAV dataset.



(c)

(a)

(b)

Figure 1. *(a) shows the first template used to feed the model to extract features at initializing the tracker. (b) the template to be tracked at the 474-th frame. (c) shows the frame of the 473-th frame of the video, the green rectangle shows the tracker output.*

The output of the model clearly shows the insufficiency of running the tracker de-pending on the initializing template; therefore, we aim to design a template update mechanism depends on the first template as static template and a dynamic template that changes repeatedly during training process.

II. RELATED WORKS

Most trackers rely on the backbone output to predict the object location without using online tracking methods. Although some recent trackers used basic template update mechanisms such as (yan et al. 2021) where the template is updated depending on MLP score prediction module that generates confidence score and update the template depending on it. (Zhang et al. 2023) updates the template depending on time spent from initializing. Where the first template has the highest score, and the last template has the lowest score.

III. DATA

The Data used to train and test the model was obtained from LaSOT [5] Large-scale Single Object Tracking benchmark. LaSOT benchmark consists of 1,400 sequences with more than 3.5M frames in total. Each frame in these sequences is carefully and manually annotated with a bounding box, making LaSOT the largest, to the best of our knowledge, densely annotated tracking benchmark. The average video length of LaSOT is more than 2,500 frames, and each sequence comprises various challenges deriving from the wild where target objects may disappear and re-appear again in the view.

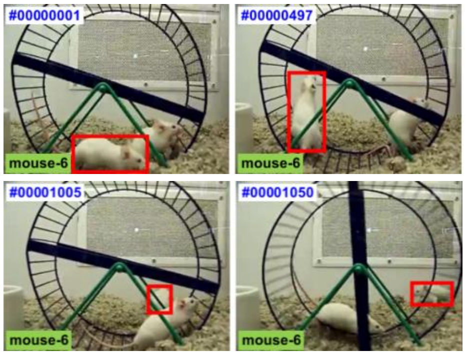


Figure 2. *Mouse-6: "white mouse moving on the ground around another white mouse".*

IV. METHOD

The proposed method contains two steps, predicting the score of the current template by calculating the similarity between the first template and current search region. The second step is feeding the model with the dynamic template. The model architecture is shown in Figure 3.

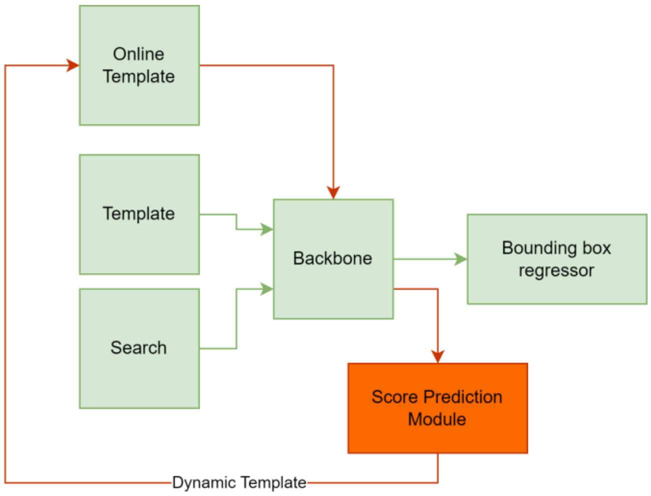


Figure 3. *The Architecture of the Proposed Template Update Mechanism. (showed in red)*

a) Similarity Prediction Model

The similarity prediction model is a framework that contains a self-attention layer with an MLP network to produce a similarity score. The input of the model is score token which is a learnable parameter that serves the model as a query. The Template and Search features that are produced from the vision transformer backbone are concatenated together and fed to the attention layer as key and value parameters. The MLP network takes the output of the attention layer and produces a score between 0-1 where 1 is the same object to be followed and 0 is a different object. The model architecture is shown in figure 4.

b) Training and Inference

The model contains of three main parts: backbone, bounding box regression head and similarity prediction model. The backbone and Bounding box regression head were trained in the OSTrack [1] paper and shared publicly. In this paper we trained the similarity prediction model by extracting the features of the template and search region and fitting the model with them. The model training was conducted using NVIDIA RTX 3060 ti with 8 gigabyte VRAM. Cross entropy was used as a loss function. AdamW was used as an optimizer. Charts 1 and 2 show the training result on 50 epochs.

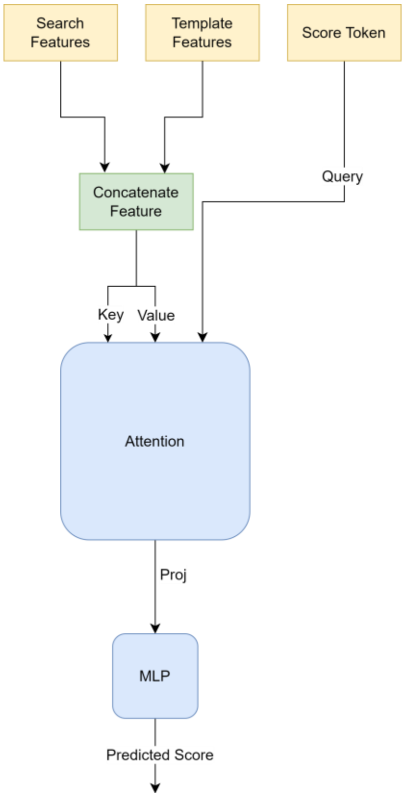


Figure 4. *The Architecture of The Proposed Similarity Prediction Model.*

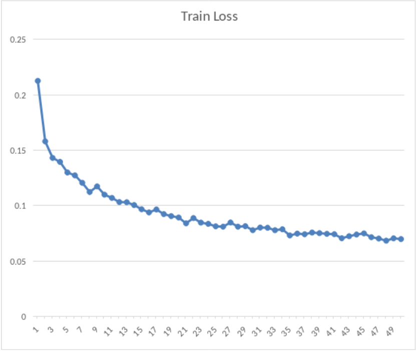


Chart 1. *The Train Loss of Similarity Model*

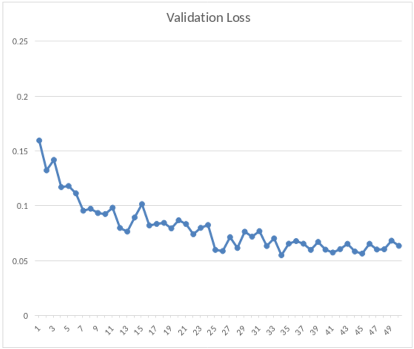


Chart 2. *The Validation Loss of Similarity Model*

V. EXPERIMENTS

The dataset was divided by 80/20 ratio for train and test. Each category contains 20 sequences so 4 sequences from each category were used to test the model. The evaluation metrics that were used to evaluate the model listed below:

- **AUC (Area Under the Curve):** AUC measures the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is a graphical representation of the trade-off between true positive rate and false positive rate at various thresholds.
- **Precision: Definition:** Precision is a measure of the accuracy of the positive predictions made by the model. It is calculated as the ratio of true positives to the sum of true positives and false positives.
- **Normalized Precision:** Normalized precision is a variation of precision that considers the imbalance in class distribution by normalizing the precision with respect to the proportion of positive instances in the dataset.

Tables 1-2 show comparison between OSTrack and OSTrack with template update mechanism. Updating the template made noticeable difference in some attributes while it still needs improvements in others.

LaSOT	AUC	OP50	OP75	Precision	Norm Precision
OSTrack	64.20	76.07	62.55	75.19	76.82
OSTrack with Template	**67.97**	**82,98**	**64.95**	**78.83**	**82,23**

Table 1. *Evaluation on Fast Motion Attribute on LaSOT Dataset*

LaSOT	AUC	OP50	OP75	Precision	Norm Precision
OSTrack	**68.21**	**79.85**	**66.11**	**74.38**	**78.01**
OSTrack with Template	67.38	79.41	64.81	73.07	77.59

Table 2. *Evaluation on Fully Occlusion Attribute on LaSOT Dataset*

VI. CONCLUSION

This paper proposes a new algorithm for template update mechanism. Using the attention in the similarity prediction model and merging the inputs to inference at one-stream made the inference time faster considering previous works. For future works, the update frequency might be adjusted to satisfy the length of the video. Template update mechanism might be able to decide where to search for object while the similarity score low.

REFERENCES

[1] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan and Xilin Chen. “Joint feature learning and relation modeling for tracking: A one-stream framework” in *European Conference on Computer Vision* (ECCV), 2022.

[2] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. “Learning spatio-temporal transformer for visual tracking” in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (ICCV), 2021

[3] Kaijie He, Canlong Zhang, Sheng Xie, Zhixin Li and Zhiwen Wang “Target-Aware Tracking with Long-term Context Attention” in *Association for the Advancement of Artificial Intelligence* (AAAI), 2023.

[4] Y. Wang, Z. Huang, R. Laganière, H. Zhang, and L. Ding, ‘‘A uav to uav tracking benchmark,’’ KnowledgeBased Systems, vol. 261, p. 110197, 2023. [Online]. Available: www.sciencedirect.com/science/article/pii/S095070512201293X

[5] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling. “LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking,” in *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019.