Co-segmentation for visual object tracking

Luka Čehovin Zajc

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
E-pošta: luka.cehovin@fri.uni-lj.si

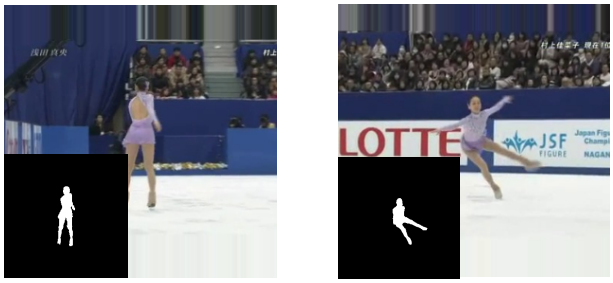
Abstract

Visual trackers have already achieved remarkable robustness due to discriminative template matching paradigm and deep descriptive features that are employed for this purpose. Yet, they are limited in their accuracy due to bounding box constraint and low spatial resolution of deep features. In recent works it has been shown that improving accuracy of discriminative deep trackers with one-shot segmentation module may not only improve the accuracy, but also boost their robustness. In this work we are exploring co-segmentation as another option for determining accurate position and size through the task of segmenting an object. We present our preliminary results on a challenging VOT2020 dataset where the proposed method achieves competitive performance to the state-of-the-art.

1 Introduction

Visual tracking is an important building block in many real-world computer vision applications. Using the most widely accepted definition of the problem, a computer vision algorithm is required to predict the state of the object in a sequence of images using only its initial state to support the prediction. Support data scarcity and the need for robustness of predictions have led most tracking algorithms to only address prediction of low-dimensional state, i.e. target position and scale. According to the recent benchmark [1], the dominant paradigm in the field is correlation bounding box tracking using deep features [2, 3] where the target represented by a multi-channel template is localized by cross-correlation between the template and a search region. These trackers are very robust, yet they operate on fixed-size rectangular patches and cannot deal with deformations that cause aspect changes.

Recently, a proposed combination of a discriminative template matcher and a deep generative segmentation model was proposed in [4]. The template matcher is responsible for coarsely localizing the target, addressing any distractors in the surrounding area. The segmentation network then finely localizes the target and infers its size and aspect ratio by determining its per-pixel mask. This combination has been shown to improve not only the accuracy of the tracker, but also its robustness. In this paper we are investigating an alternative approach to object segmentation using limited support data during the tracking session. Our basis is co-segmentation, where an objective is to jointly segment semantically similar objects in two or more images. We present our adaptations of a co-segmentation method (see Figure 1) to the visual object tracking domain and show some preliminary results that demonstrate the feasibility of the approach.



Anchor image and mask — Query image and mask prediction

Figure 1: Co-segmentation in visual tracking. Our method uses an anchor image and a mask of an object of interest to infer its mask in another frame.

2 Related work

For more than a decade, the most robust approaches to visual tracking have viewed the problem as a machine-learning task of discrimination between the object and the background [5]. In this time the features used to describe an image and the methods that enabled fast learning of a discriminative model have changed dramatically. Most recent methods use deep features, either pre-trained for discriminative localization of an object [2] or train a multi-channel discriminative correlation filter on-line [3, 6].

Different than localization, size estimation was commonly handled using a multi-scale pyramid, which is a greedy but inaccurate approach [7] or not taken into account at all. With the success of deep learning and acceptance of more power-hungry methods, some researchers begun experimenting with region proposal methods to estimate target's size and aspect ratio [8, 9], others employed box refinement methods [10, 6]. While segmentation output was considered in [9] it was not used to improve the overall tracking process. In [4] a segmentation was used in a mutually beneficial interaction with a dis-

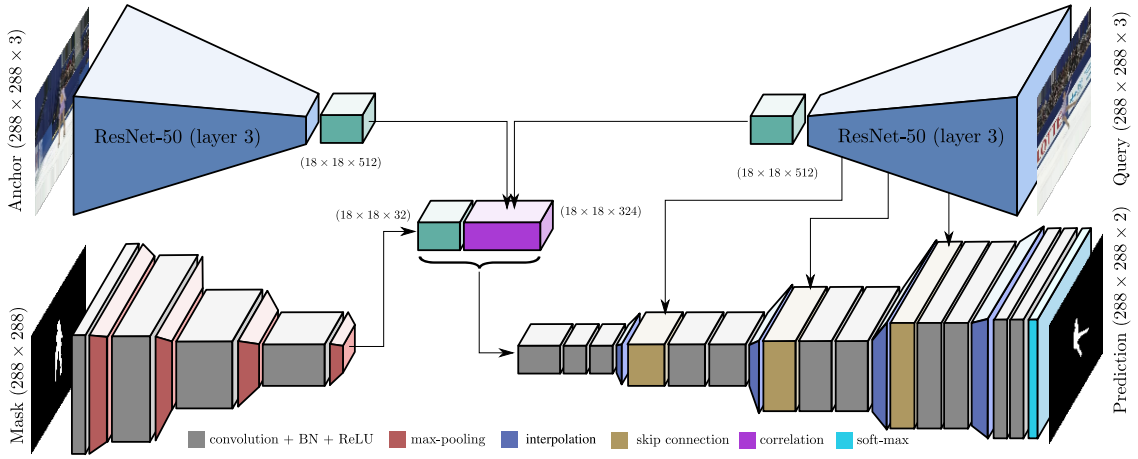


Figure 2: Schematic representation of our co-segmentation model for visual tracking.

criminative tracker for the first time. The discriminative tracker provided initial localization that is then refined using object segmentation. Segmentation is also used to adjust sampling size. All this contributes to increased accuracy and robustness. But the method is still partially handcrafted. In this work we investigate an approach based on co-segmentation method that is fully differentiable and can be trained end-to-end.

The field of co-segmentation is a special branch of segmentation where an objective is to segment one or more objects in more than one image at the same time [11]. There are many types of co-segmentation with assumptions about number of salient objects [11, 12], background constancy [13]. While co-segmentation seems like a good match for visual tracking, the objects of interest in co-segmentation are very loosely defined by salience and the emphasis is on generality. Resulting segmentation lacks specificity to segment only a single object which is the goal in visual tracking which also makes no assumptions about the appearance of the object itself as well as its surrounding area. In the following section we present a modification one of the deep co-segmentation methods [14] that makes it usable in visual tracking scenario.

3 The co-segmentation model

In this work we are evaluating a simple single-shot multi-object co-segmentation approach, inspired by [14]. The main concept behind the original work is a feature correlation layer that originates from FlowNet [15] for deep optical flow estimation. Instead of comparing deep features locally, as it is common in optical flow estimation, [14] compares all features from a reference image with all features from the query image, building an affinity tensor. This is then followed by an up-sampling decoder that is trained to reconstruct a mask of semantically meaningful objects found in both images.

In our co-segmentation network, displayed in Figure 2, the VGG16 [16] backbone is replaced with a pre-trained ResNet50 [17] that we have found to be better suited for segmentation task. We use features from the third layer of the feature extractor to calculate an affinity tensor between the anchor image and the query image. In contrast to the co-segmentation task we have introduced a separate encoder to encode the input mask from the anchor frame to its high-dimensional representation that we use during the prediction mask inference process together with the computed affinity tensor. In contrast to the original work, we also introduce skip connections from the feature encoder for the query image into the decoder to improve the details of the predicted mask.

3.1 Training

The model was trained on YouTube-VOS [18] video segmentation dataset. Each training sample contained two patches selected from the same sequence with corresponding segmentation masks of the target object. Samples are drawn randomly from the dataset and augmented using geometrical transformations (translation, rotation, scale) and pixel alteration (brightness, contrast, noise, blur). Examples of training samples are shown in Figure 3.

The ResNet feature extractor backbone was not updated during training, we have only trained mask encoder and mask decoder to minimize binary cross entropy (BCE) loss. Similarly to [14] we have used each sample multiple times, the final loss is therefore:

$$\begin{aligned} loss = & BCE(P(I_1, M_1, I_2), M_2) \\ & + BCE(P(I_2, M_2, I_1), M_1), \end{aligned} \tag{1}$$

where the tuple (I_1, M_1, I_2, M_2) constitutes a training sample, $P(\cdot, \cdot, \cdot)$ denotes the prediction model.

The model was trained on 40000 unique samples in batches of 16 samples for 33 epochs, each lasting for 400 iterations. We have used Adam optimizer [19] with initial learning rate 0.0001 that was decreased by half every 10 epochs.

3.2 Tracking

Similarly to [4], we have evaluated our segmentation model within a state-of-the-art discriminative tracker, presented in [6] where the IoU bounding box refinement [8] is replaced by our model for region and scale estimation. Both components share the same ResNet50 [17] feature extractor.

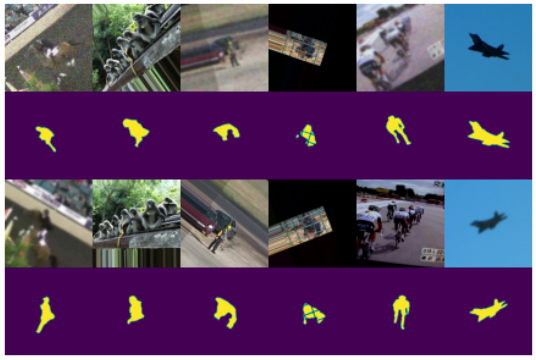


Figure 3: Examples of training samples, each sample column is composed of two augmented samples with corresponding segmentation masks.

Tracker	EAO	A	R	Overlap
ATOM [6]	0.267	0.467	0.708	0.386
SiamMask[9]	0.321	0.624	0.648	0.405
D3S[4]	0.439	0.699	0.769	0.508
Ours	0.416	0.696	0.735	0.578

Table 1: Results for VOT2020 baseline and unsupervised experiments. For all measures higher value is better.

In each frame the target is approximately localized using the discriminative component, then a query patch is extracted and fed into the segmentation model together with the anchor patch (from first frame) and its mask. The resulting mask is used to estimate the position of the target as well as its size. The size is used to adjust trackers internal image sampling scale. We assume that the scale changes according to a random-walk motion model that we estimate using Kalman filter [20]. The uncertainty of a prediction is estimated by using the co-segmentation in reverse: using the predicted mask and query image as reference the model predicts the mask in the anchor frame and compares it to the initialization mask. The overlap between those masks is used to determine prediction uncertainty.

4 Experimental evaluation

We have performed our experiments on a recent VOT2020 dataset [21]. The dataset contains 60 diverse sequences, the target object in each sequence is annotated with a segmentation mask. The overall results for the experiment with multiple restarts and the unsupervised single run experiment are shown in Table 4. Scores in the table show that our method is performing comparably to the D3S [4] and significantly outperforms another recent segmentation tracker SiamMask [9] as well as ATOM [6] from which it was derived. More importantly, the results also show that our method improves results of ATOM in terms of accuracy as well as robustness.

Comparison between our method and D3S [4] is also interesting because the former can be viewed a generalization of the latter. In D3S the mask is directly down-sampled to the feature size, the generative model and its application are tuned by hand. Our method performs down-sampling and matching in a fully differentiable manner, allowing the network to take care of the exact details of which data is important. This requires a larger network, at the moment we have been very liberal with its size, but we believe that the number of weights can be reduced with further experiments. Still, it is worth noting that our model was trained with around 2% of data used by D3S and in 20% of the time (around 4 hours).

Even though the overall performance of D3S and our method are similar, the results on per-sequence basis show some conceptual differences. Figure 4 shows frames of selected sequences. In general D3S better estimates regions with highly deformable objects on a clearly different background (Figure 4, sequences gymnastics1 and iceskater1). Our method is more conservative and does not segment every limb in such cases. On the other hand it can handle occlusions with similar objects better because of the more verbose correspondence description (Figure 4, sequences girl, flamingo1, and nature).

The main weaknesses of the proposed method is the static anchor template and the fragile scale estimation approach. If the scene will change too much from the first frame, the match will become unreliable. While our method does not tend to explode the segmentation as frequently as D3S, it is much more prone to collapse: because the difference between the anchor and query are too great, the segmentation only captures part of the object (hence wrong scale) or confuses it with with a similar object with better matching appearance (Figure 4, sequence fish2). This shows the need for a scale estimation that is decoupled from segmentation as well as some kind of anchor update mechanism.

5 Conclusion

We have presented our preliminary work with integration of a co-segmentation model into a visual tracker. We have modified a co-segmentation model [14] and integrated into a state-of-the-art discriminative tracker. The results on a challenging VOT2020 dataset show that our proposed tracker outperforms a tracker that it was derived from [6] and performs comparably to a very recent method that integrates segmentation into tracking in a conceptually similar way [4].

We believe that our findings open up new opportunities for more accurate and robust tracking using segmentation. In the future we plan to investigate how to reliably update the anchor image to keep it relevant in the long run or to even use multiple anchor images. We will also investigate how to estimate object's scale directly in the model and how to resolve ambiguities and distractions by observing the context from the past frames.

Acknowledgement: This research was supported by the ARRS grant Z2-1866.

References

[1] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J. Kämäräinen, L. Cehovin Zajc, O. Drbohlav, A. Lukezic, A. Berg, A. Eldesokey, J. Käpylä, G. Fernández, and et al. The seventh visual object tracking

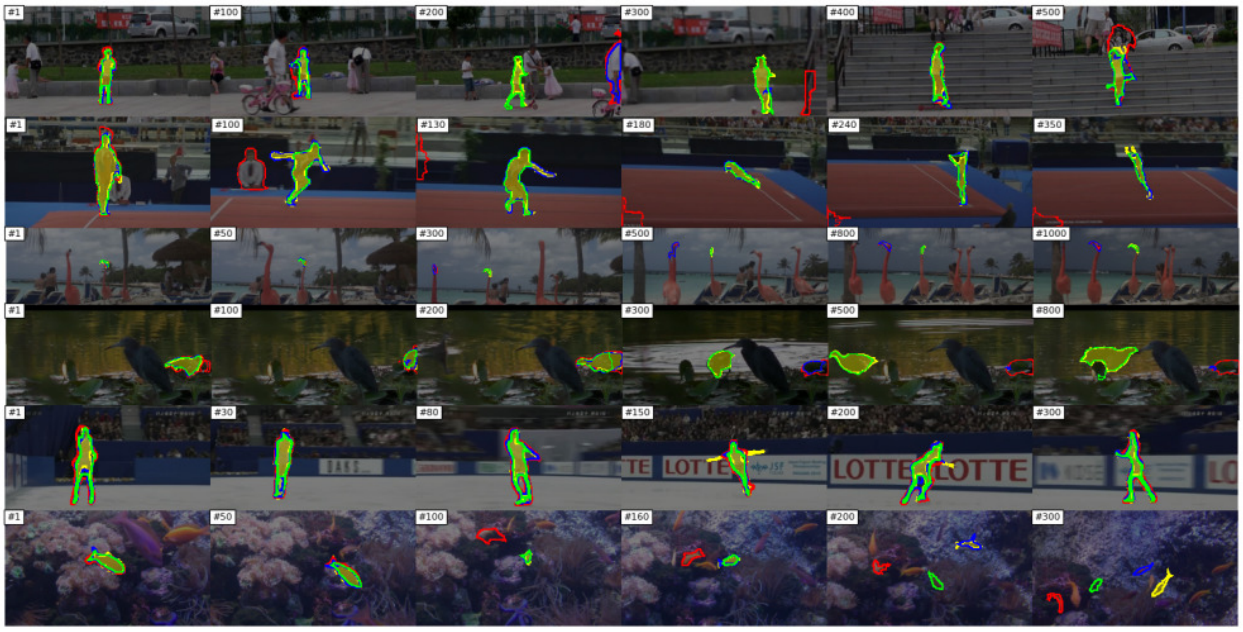


Figure 4: Qualitative evaluation on selected sequences from VOT2020: girl, gymnastics1, flamingo1, nature, iceskater1, and fish2. Groundtruth is shown with yellow, tracker results with red (SiamMask [9]), blue (D3S [4]) and green (ours). All images were dimmed for better region visibility.

vot2019 challenge results. In *ICCVW*, pages 2206–2241, 2019.

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, pages 850–865. Springer, 2016.

[3] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, pages 6638–6646, 2017.

[4] Alan Lukezic, Jiri Matas, and Matej Kristan. D3s-a discriminative single shot segmentation tracker. In *CVPR*, pages 7133–7142, 2020.

[5] Helmut Grabner, Michael Grabner, and Horst Bischof. Real-time tracking via on-line boosting. In *BMVC*, 2006.

[6] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, pages 4660–4669, 2019.

[7] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 2015.

[8] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018.

[9] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, 2019.

[10] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, pages 784–799, 2018.

[11] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *CVPR*, volume 1, pages 993–1000. IEEE, 2006.

[12] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[13] Sara Vicente, Vladimir Kolmogorov, and Carsten Rother. Cosegmentation revisited: Models and optimization. In *ECCV*, pages 465–479. Springer, 2010.

[14] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. In *ACCV*, pages 638–653. Springer, 2018.

[15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.

[20] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

[21] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J. Kämäräinen, M. Danelljan, L. Cehovin Zajc, A. Lukezic, O. Drbohlav, L. He, Y. Zhang, S. Yan, J. Yang, G. Fernández, and et al. The eight visual object tracking vot2020 challenge results. In *ECCVW*, 2020.