Adversarial Learning for Visual Tracking Research Idea

Emanuel Di Nardo[1][0000−0002−6589−9323]

University of Milan, Milan MI 20122, Italy

Abstract. The doctoral research activity[1] mainly focuses on methodologies in the field of computer vision. In particular, the work is focused on designing, developing and validating novel approaches, also based on deep learning methodologies, for visual tracking. Visual tracking in video sequences has always been a main topic in computer vision and interesting results have been obtained by approaches based on Support Vector Machine, Siamese Networks and Discrete Correlation Filters. However, these techniques are limited due to the low discriminative ability of the used features for object detection. In his research activities, Emanuel Di Nardo proposes a novel approach, based on Generative Adversarial Networks for feature extraction or regression. In particular, using Generative Adversarial Networks we are able to characterize the elements to be traced in the scene and make them easier to recognize.

Keywords: Deep Learning · Adversarial Learning · Feature Extraction · Visual Tracking

1 Introduction

Visual tracking in video sequences has always been a topic that arouses the attention of the scientific community. It consists in detecting and following an object that moves in a scene. The object will inevitably undergo modifications during its movement (Fig. 1). It happens both because of the displacement itself being free of constraints and because the scene itself, which could present obstacles between the camera and the object and still due to problems caused by the acquisition conditions as in the case of non-ideal lighting. Therefore it is needed to use techniques that are defined as *robust* with a fair compromise of accuracy. There are many challenges for this kind of task and one is VOT (Visual Object Tracking) [7]. Usually, in this context the following three parameters are taken into account:

1. **Accuracy.** Mean Overlap between the target and the ground truth
2. **Robustness.** How many times the target is lost
3. **Expected Average Overlap (EAO).** Mean of accuracy over multiple video sequences with the same visual properties. It combines accuracy and robustness

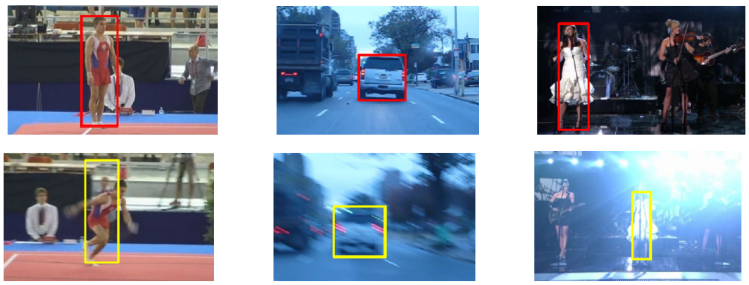

Fig. 1. Tracking examples under different conditions [5]. It is possible to observe in (row 1) normal appearence and in (row 2) appearance changes, rigid target with a motion blur, scale changes and illumination problems

In the VOT challenge, it is possible to identify two kinds of tracking called *short-term* and *long-term*. In the former, an object is always visible in the scene and it is possible to detect it in each frame. In the latter, the target could not be present in the scene for long time due to a total occlusion because it came out of the scene. In this case, the tracker can not report any position for the object, but it can provide a confidence score that the object is not present.

2 Related Work

Most of the works in visual tracking are compared on various challenges such as VOT [7] and MOT (Multiple Object Tracking) [8]. On the one hand, a strong evolution of techniques is based on the template matching of the whole object [9] [10] [11]. On the other hand, other approaches tend to take into account the movement and to estimate the possible position in which the target is located, relying for example on the optical flow [12] [13]. Other methodologies called part-based tend to scan the areas close to the initial target by estimating which points have the greatest probability that there is a target or a part of it [14] [15]. Nowadays, most trackers use approaches based on artificial Neural Networks (NN) at various stages of the tracking process. Some use them to have meaningful features that can be representative of the object [16] [17]. In recent years, moreover, techniques based on Siamese networks have emerged, which see two parallel networks that work together to estimate the location of the object in the scene [5] [17] [18]. Other techniques use filters that allow, through a domain transformation, to discriminate the probable position in a robust and a highly-efficient way [19] [20] [23]. Some methodologies mix all these approaches together to be more and more precise [21] [22]. The approaches based on Deep Learning [16] [17] [5] [22] use a pre-trained Neural Network on known datasets for classifying objects in the images. A recent technique is VITAL [24]. It uses

[1] Ph.D. supervisor: Angelo Ciaramella (University of Naples Parthenope); co-supervisor: Fabio Narducci (University of Naples Parthenope).

a Generative Adversarial Networks to generate a mask that represents the most relevant features in the image, based on the input target. Another recent approach uses compressive sensing for trajectory tracking [26] in order to reduce the image complexity and be able to know where the target is moving on.

3 Research Idea

The main objective of the research activity is the introduction of a novel approach based on Generative Adversarial Networks (GANs) [1] for visual tracking. In a first scenario a GAN is used for tracking. Usually, the adversarial networks are used to generate samples that are as close as possible to the real ones. This is possible thanks to the property they have to learn the distribution of the data they want to generate. This type of approach leads to the generation of a latent space for the representation of the data. Here, the idea is to use this property for extracting representative samples (i.e., features). Some studies reported this approach combined with autoencoders [27] [28]. In particular, the generator network encodes the vector representation extracted form the latent space. In a second scenario GANs could be used directly to perform a regression operation [29] [30] to define where an object is found, or at least, as a support to the localization through probable positions. Differently from what has already been done [24], in the proposed approach GAN should be able to extract a meaningful representation of the object with a concrete reconstruction of the target localization in the image instead of a simple dropout mask that suggests what are the areas that are more sensible to the input. Furthermore, we want the ability of domain adaptation of GANs to be discriminating for this activity without recurring, as it happens in [24] and [25] to pre-training on the task of tracking. Another aspect that should not be underestimated is that of scene and the object regularizing. In this context, GANs could be used to remove alterations in images to make tracking easier or even generating objects in positions different from those known to better estimate how an object can be changed over time.

4 Planning

Phase One A generative network is built to perform and study the segmentation of the image trying to localize a target object in an image. The investigation aims to generate an image that is related to the ground truth used in the discriminator network. As the first point, the activity tries to relate an image and a target that is visible in it. The first experiments are conducted using the model in Fig. 2 a generative model with two inputs (the image and the target) that are encoded individually. In the end, they are concatenated on the feature dimension to achieve an association between them. Further methodologies are in development to enforce the relationship between the image and the target. In this context, it has to face some problems. The first one is the multi-domain property that the network tries to approximate because it is not trained on a set of objects that belong to the same category, but on a large variety of them. On

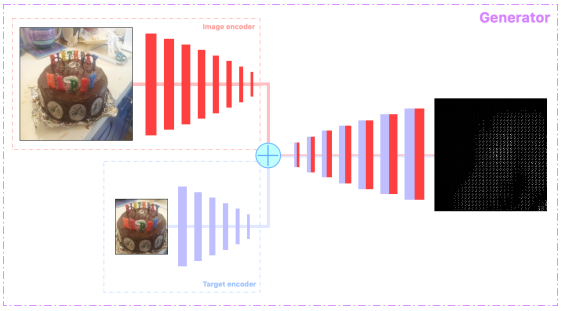

Fig. 2. First generator model studied in the research activity

the other hand, the segmentation purpose should help to normalize this behavior because the objective function is calibrated to work on a less complex solution. Another problem is related to the segmentation quality. It is possible that the result is not accurate with a *degeneration* to a kind of output that can be more similar to a heat-map.

Phase Two The segmentation obtained from the first step can help to understand what is the discriminatory effect of the learned latent space. It can be used trying to work only on the encoding of the input without bringing it to the generative output in a pure autoencoder fashion. The main challenge is on the usage of only encoded features because space on which it is mapped could be lost some important properties if partially described. Another important investigation can be done on the discriminator side of the GAN. Usually, it is used only in the generative learning step and not in the operative phase, but it learns to extract characteristic features of real data. It is an important property that could be used in a verification step to avoid low-quality output or in a distraction-aware manner.

Future proposals In addition to GAN based plan, other research paths can be invested in future investigation. It involves analyzing the techniques based on dictionary learning, compressive sensing [26] and ODE [3] that appear to be a valid alternative to the classical operations found in all tracking algorithms.

The shown planning would achieve a strong representation of a generic object and, consequently, learning also the localization in the space. It can be used as *tracking-by-detection* solution in a short-term challenge. It gives the possibility to join the VOT challenge to evaluate the research work using the benchmark tools provided in the competition and validate the effectiveness of the developed model.

5 Conclusions

Aim of this study is the introduction of innovative methodologies for visual tracking in the field of computer vision. In particular, it could be noted that

new elements that are currently used in different fields with excellent results could give a strong boost to the current research status.

These should be validated comparing with proposed state-of-the-art methodologies to understand how much they characterize or not. From here, it is possible to analyze how to use the whole adversarial model as a detector of objects without going through other methods. This study could also lead to using different characterization techniques from convolutional networks. In particular, in [31] the proposed approach deviates from the classic CNN and that appears to be a good alternative to prevent the number of parameters from exploding.

References

1. Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.
2. E. J. Candes, J. Romberg and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," in IEEE Transactions on Information Theory, vol. 52, no. 2, pp. 489-509, Feb. 2006. doi: 10.1109/TIT.2005.862083
3. Ricky T. Q. Chen and Yulia Rubanova and Jesse Bettencourt and David Duvenaud. Neural Ordinary Differential Equations, 2018.
4. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.
5. Bertinetto, Luca, et al. "Fully-convolutional siamese networks for object tracking." European conference on computer vision. Springer, Cham, 2016.
6. Lukezic, Alan, et al. "Discriminative correlation filter with channel and spatial reliability." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
7. Kristan, Matej, et al. "The sixth visual object tracking vot2018 challenge results." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
8. Dendorfer, Patrick, et al. "CVPR19 Tracking and Detection Challenge: How crowded can it get?." arXiv preprint arXiv:1906.04567 (2019).
9. Chen, W., Cao, L., Zhang, J., Huang, K.: An adaptive combination of multiple features for robust tracking in real scene. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 129–136, December 2013
10. Gao, J., Xing, J., Hu, W., Zhang, X.: Graph embedding based semi-supervised discriminative tracker. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 145–152, December 2013
11. Maresca, Mario Edoardo, and Alfredo Petrosino. "Matrioska: A multi-level approach to fast tracking by learning." International Conference on Image Analysis and Processing. Springer, Berlin, Heidelberg, 2013.
12. Wendel, Andreas, Sabine Sternig, and Martin Godec. "Robustifying the flock of trackers." 16th Computer Vision Winter Workshop. Citeseer. 2011.
13. Maresca, Mario Edoardo, and Alfredo Petrosino. "Clustering local motion estimates for robust and efficient object tracking." European Conference on Computer Vision. Springer, Cham, 2014.
14. Lukežič, Alan, Luka Čehovin Zajc, and Matej Kristan. "Deformable parts correlation filters for robust visual tracking." IEEE transactions on cybernetics 48.6 (2017): 1849-1861.

15. Battistone, Francesco, Alfredo Petrosino, and Vincenzo Santopietro. "Watch out: Embedded video tracking with BST for unmanned aerial vehicles." Journal of Signal Processing Systems 90.6 (2018): 891-900.
16. Sun, Chong, et al. "Learning spatial-aware regressions for visual tracking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
17. Li, Bo, et al. "High performance visual tracking with siamese region proposal network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
18. Li, Yuhong, and Xiaofan Zhang. "SiamVGG: Visual Tracking using Deeper Siamese Networks." arXiv preprint arXiv:1902.02804 (2019).
19. Kiani Galoogahi, Hamed, Ashton Fagg, and Simon Lucey. "Learning background-aware correlation filters for visual tracking." Proceedings of the IEEE International Conference on Computer Vision. 2017.
20. Lukezic, Alan, et al. "Discriminative correlation filter with channel and spatial reliability." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
21. Li, Feng, et al. "Learning spatial-temporal regularized correlation filters for visual tracking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
22. Valmadre, Jack, et al. "End-to-end representation learning for correlation filter based tracking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
23. Yun, Sangdoo, et al. "Action-decision networks for visual tracking with deep reinforcement learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
24. Song, Yibing et al. "VITAL: VIsual Tracking via Adversarial Learning." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): n. pag. Crossref. Web.
25. Nam, Hyeonseob, and Bohyung Han. "Learning Multi-Domain Convolutional Neural Networks for Visual Tracking." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): n. pag. Crossref. Web.
26. Kracunov, Marijana, Milica Bastrica, and Jovana Tesovic. "Object Tracking in Video Signals Using Compressive Sensing." 2019 8th Mediterranean Conference on Embedded Computing (MECO) (2019): n. pag. Crossref. Web.
27. Pinho, Eduardo, and Carlos Costa. "Feature Learning with Adversarial Networks for Concept Detection in Medical Images: UA. PT Bioinformatics at ImageCLEF 2018." CLEF (Working Notes). 2018.
28. Sohn, Kihyuk, Honglak Lee, and Xinchen Yan. "Learning structured output representation using deep conditional generative models." Advances in neural information processing systems. 2015.
29. Olmschenk, Greg, Zhigang Zhu, and Hao Tang. "Generalizing Semi-Supervised Generative Adversarial Networks to Regression Using Feature Contrasting." Computer Vision and Image Understanding 186 (2019): 1–12. Crossref. Web.
30. Karan Aggarwal and Matthieu Kirchmeyer and Pranjul Yadav and S. Sathiya Keerthi and Patrick Gallinari. Regression with Conditional GAN. 2019
31. Ullah, Ihsan, and Alfredo Petrosino. "A strict pyramidal deep neural network for action recognition." International Conference on Image Analysis and Processing. Springer, Cham, 2015.