Siamese Instance Search for Tracking

Lina Zhu

June 10 ,2018

Abstract

In this article, we present a fundamental tracker that differs from the most advanced trackers: we do not apply any model updates, no occlusion detection, no combination of trackers, no geometric matching, and still provide the most advanced technology tracking performance. The presented tracker simply matches the initial patch in the candidate's first frame of the target in a new frame and returns the most similar patches through the learning matching function. The strength of the match function comes from a general extensive training, ie without any target data, using a conjoined deep neural network we designed for tracking. It turns out that the learned matching function is very powerful, very simple tracker builds on it, creates a conjoined instance search tracker, SINT, which only uses the original observation of the target from the first frame, enough to achieve the best performance . In addition, we show that the proposed tracker even allows the target to re-identify the absence of a complete video shot after the target.

1. Introduction

The heart of many tracking algorithms is the function of the image and input matching frames of the target. Matching functions for tracking ideally provide for good matching even if objects in the video are occluded, change their size, rotate in and out of the plane or experience uneven illumination, camera motion, and other disturbing factors. One approach is to model each of these distortions by introducing affine transformations explicitly in matching[2], probability matching[1], feature images, illumination invariants[3], and occlusion detection[4] . While a clear matching mechanism may be well-suited to solve a distortion, it is very likely that disturb another one. In this work, we propose to learn to match rather than explicitly model the matching of specific distortions. More specifically, we suggest that we learn from external videos that contain various interference factors. However, these invariance invariances are not explicitly modeled. If the set of external videos is large enough, the goal is to learn a universal matching function prior. In this article, we simply match the first goal with the candidate in a new framework frame and return the most similar one of the learning matching functions, without updating the target, the tracker combination, the occlusion detection, and so on. Figure 1 illustrates the tracking algorithm. We summarize the contribution of this work as follows: First, we propose to learn a generic matching function from external video data tracking in order to robustly process the video sequence of the common appearance changes that the object may experience. The learned function can be applied as it is without any adaptation, the previous new tracking video cannot see the target object. Secondly, based on the universal matching function learned on this basis, we propose a tracker that achieves the most advanced tracking performance. The presented tracker is completely different from the most advanced tracker tracker. We do not apply model updates, do not apply occlusion detection, no tracker combination, no geometric matching, and so on. In each frame, the tracker only needs to find the best patch that matches the candidate's initial patch with the target in the first frame of the learning match function. Third, to learn the matching function, we use a dual-stream concatenation network, which we specifically design for tracking. Further, the instant model is updated without any drift that one would expect, and the proposed tracker allows successful target objects to re-identify after the goal is absent for a long period of time, such as a full shot.

2. Design evaluation

We first validate our design choices of the network. In this sets of experiments, box refinement is not considered. Network tuned generically on external video data vs. network pre-tuned on ImageNet vs. network fine tuned target-specifically on first frame In this experiment, we show the effectiveness of the Siamese network tuned on external data. To that end, we compare the Siamese fine tuned AlexNet-style network using ALOV with the ImageNet pretuned AlexNet and the Siamese fine tuned network using the training pairs gathered in the first frame. In this comparison, all three use a single layer fc6 for feature representation. As shown in the rows (a)-(c) of Table 1, the Siamese fine

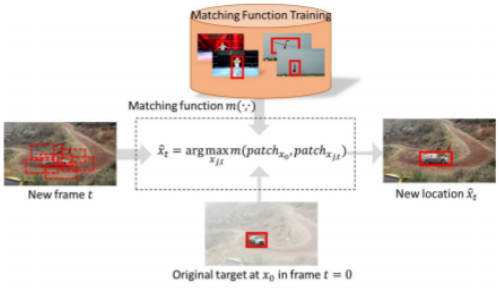

Figure 1. The tracker simply finds the patch that matches best to the original patch of the target in the first frame, using a learned matching function.

tuned network using ALOV (c) significantly improves over the pre-tuned net (a), while fine tuning on the first frame (b) gives a marginal improvement. We conclude that Siamese networks fine tuned using large amount of external data are to be preferred. To max pool or not to max pool? We now examine our design choice of having no maxing pooling layers in the network. As shown in Table 1, (d) vs. (a) and (e) vs. (c), including max pooling layers deteriorates accuracy, as expected due to the reduction of the resolution of the feature maps which causes poor localization. when no max pooling layers are included, the success rate improvement is higher at higher intersection-over-union overlap ratios, see Table 2. We conclude that max pooling layers are not necessary for our Siamese invariance network with small AlexNet-style architecture.

Table 1. Evaluation of different architectural and design choices of the Siamese invariance network for tracking on the OTB dataset.

	AUC	Prec@20
(a)pretrained-alexnet-fc6	42.8	66.3
(b)firstframe-Siamese-finetuned-alexnet-fc6	44.0	67.9
(c)Siamese-finetuned-alexnet-fc6	47.4	72.0
(d)pretrained-alexnet-fc6-nomaxpooling	50.0	70.8
(e)Siamese-finetuned-nomaxpooling	53.9	74.8
(f)Siamese-finetuned-conv45fc6-nomaxpooling	55.0	76.2
(g)Siamese-finetuned-vgg16	59.2	83.6

Table 2. Success rates (sr) of the tracker at three intersection-over-union overlap ratios for different network architectures.

	sr@0.3	sr@0.5	sr@0.7
pretrained-alexnet-fc6	68.3	46.2	19.6
pretrained-alexnet-nomaxpooling	75.3	58.1	32.6
Siamese-finetuned-alexnet-fc6	74.6	56.2	25.4
Siamese-finetuned-alexnet-nomaxpooling	79.3	67.6	38.8

3. Conclusion

This work presents Siamese INstance search Tracker, SINT. It tracks the target, simply by matching the initial target in the first frame with candidates in a new frame and returns the most similar one by a learned matching function.The strength of the tracker comes from the powerful matching function, which is the focus of the work. We take extra care that there is absolutely no overlap between the training videos and any of the videos for evaluation. Namely, we do not aim to do any pre-learning of the tracking targets. Once learned, the matching function is used as is, without any adapting, to track arbitrary, previously unseen targets. It turns out the matching function is very effective in coping with common appearance variations an object can have in videos. The simple tracker built upon the matching function, reaches state-of-the-art performance on OTB, without updating the target, tracker combination, occlusion detection and alike.

References

[1] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, 2000. 1

[2] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981. 1

[3] H. T. Nguyen and A. W. Smeulders. Robust tracking using foreground-background texture discrimination. *IJCV*, 2006. 1

[4] J. Pan and B. Hu. Robust occlusion handling in object tracking. In *CVPR*, 2007. 1