Assignment 1 Computer Vision course

Spadetto Matteo[214352]

Universita degli Studi di Trento, via Sommarive 9, IT
matteo.spadetto-1@studenti.unitn.it

Abstract. The goal of this assignment is to provide details for three object tracking methods. The chosen methods are ECO, MDNet, RT-MDNet in particular, and SiameseFC, with SiamMask.

Keywords: Object tracking · Deep tracking · ECO · MDNet · RT-MDNet · SiameseFC · SiamMask.

1 Introduction to object tracking

An important branch of computer vision science is developing nowadays in order to detect objects and tracking them. These algorithms are used mainly in Human-Machine interfaces (HMI), video surveillance, robot navigation (including driverless cars) and more. For their relevant importance in the world and their high level of growth, these technologies are continuously implemented by the community to find better solutions to the problems found in this decade. The most critical issues are:

- Fast Motion (FM);
- Occlusion (OCC);
- Illumination Variation (IV);
- Motion Blur (MB);
- Deformation (DEF);
- Scale Variations (SV);
- Background Clutter (BC);
- Low Resolution (LR);
- In-Plane Rotation (IPR);
- Out-of-Plane Rotation (OPR);
- Out-of-View (OV).

In order to solve them, a lot of new techniques and methods was developed. This assignment has the intent of make an overview on object tracking methods used, focusing in three of them. It is important to say that there is not a defined approach, with a single method, to overcome all issues of object tracking but a list of them, some good for some features and some others for different ones. This is due to the large application possibilities in the real world, the field in witch computer vision technologies are mainly involved.

2 Object tracking methods

After a little preview of object tracking science, an overview of the different methods treat in the next pages is necessary. In table 1 and [Fig.1] a summary of the main properties of all the methods is given. The purpose of this assignment is to get details of the most performing methods available, one for each main classes (that will be exposed next) of tracker was chosen.

Table 1. Main properties for ECO, RT-MDNet and SiamMask trackers

Tracker	Class	Baseline tracker	Scale estimation	Offline training	Online learning	FPS
ECO	CFT	CCOT	Yes	No	Yes	8
RT-MDNet	Deep	MDNet	Yes	Yes	Yes	1
SiamMask	Deep	SiameseFC	Yes	Yes	No	58

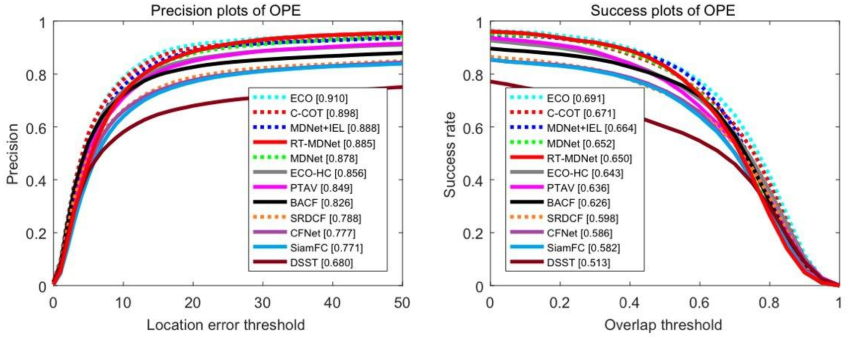

Fig. 1. Quantitative results on OTB2015 from which it is possible to find precision and success of the ECO, RT-MDNet and SiameseFC

3 Data-sets used for validation

In object tracking, different methods of validation are used in order to produce different benchmarks for the community. These challenges built, in the last years, a standard for object trackers evaluation defining specific requirements. The main two of them are:

- Visual Object Tracking (VOT) challenges;
- Object Tracking Benchmark (OBT).

In VOT challenges, composed by five different categories (for different tracking domains), trackers are tested with common datasets for each category and an evaluation kit, then results and codes are shared publicly and available at the VOT site (`http://www.votchallenge.net/`). The two main classes in which the challenge is divided are:

- Single-camera, single-target, model-free, casual-trackers, applied to short-term tracking (short-term means that is not required that trackers are able to re-detect target in case of loss of the object);
- Single-camera, single-target, model-free, applied to long-term tracking (long-term means that is required re-detection of the target in case of loss of the object).

Dataset are focused to stress algorithms, and results are based on accuracy, robustness and expected average overlap (EAO). In general, the provided datasets are composed by videos going from 41 to 1500 frames in different pixel resolutions (some for short-term and some for long-term challenges) at 30fps. Scenarios are very different in order to test motion change, occlusions, illumination change, camera motion and size change. Examples of videos are in picture [Fig.2]. OTB evaluates trackers on two different bases producing results for both precision and success:

- Temporal Robustness Evaluation (TRE);
- Spatial Robustness Evaluation (SRE).

It is also important to remember that there are a lot of challenges and benchmarks in object tracking world as UAV, TempleColor, and others producing different results and benchmarks.

4 Source codes

For all the methods described in this assignment it is possible to find the open-source code in the relative GitHub repositories. The source code is fully downloadable and includes all the necessary material to make trackers work in different operative systems and program languages. With these informations is possible to perform demos, training or simply have more details on how these methods are giving successful results thanks to the guides provide in the "README.md" file of the projects.
It is possible to find the source codes of the treated trackers at the following links:

- ECO: `https://github.com/martin-danelljan/ECO`;
- MDNet: `https://github.com/HyeonseobNam/MDNet`;
- RT-MDNet: `https://github.com/IlchaeJung/RT-MDNet`;
- SiameseFC: `https://github.com/bertinetto/siamese-fc`;
- SiamMask: `https://github.com/foolwood/SiamMask`.

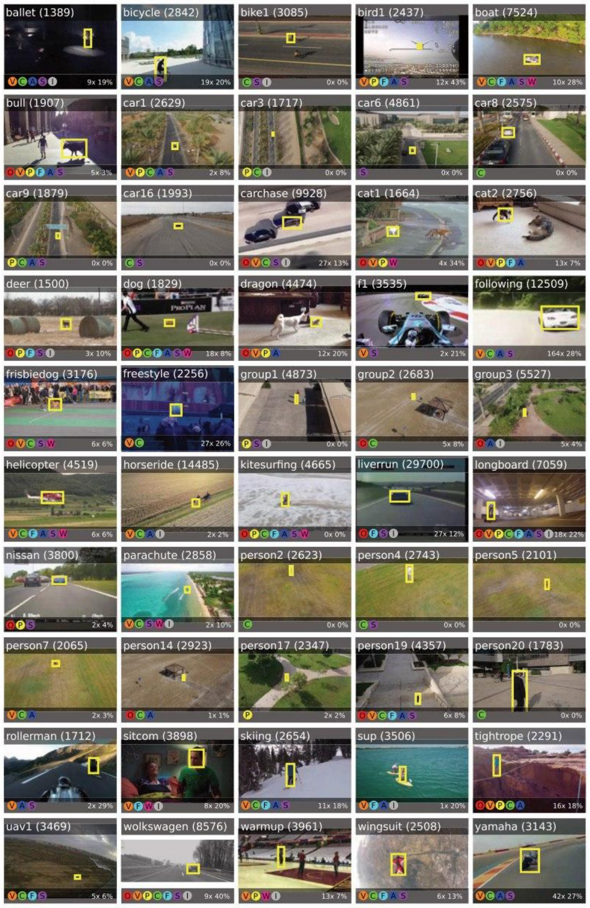

Fig. 2. LBT50 dataset provided for VOT challenge, a frame for each video. (V) Out-of-view, (O) Full occlusion, (C) Camera motion, (P) Partial occlusion, (F) Fast motion, (S) Scale change, (I) Similar objects, (W) Viewpoint change, (A) Aspect ration change.

5 ECO, RT-MDNet and SiamMask methods

In object tracking is possible to categorize trackers as generative vs. discriminative, single-object vs. multiple-object and online vs. offline learning. It is also possible to separate trackers in Correlation-Filter base Trackers (CFTs) and Non-CFTs (NCFTs), this last difference deserves a little explanation.
CFTs trackers, initially not very efficient for online tracking, are now efficient for adaptive tracking thanks to Minimum Output of Sum of Squared Error (MOSSE), computed in Fourier domain. This kind of trackers work in frequency domain in order to get more speed in computations. They follow the "tracking-by-detection" approach. Correlation filters are generated at the first frame for

a specific target position. In the tracking phase, position is estimated using the position of the object in the previous frame. Then, thanks to feature extraction methods, it is possible to build a feature map starting from the initial input patch. It is also possible to apply other types of filters in order to perform a greater tracking. The target is, at the end, updated with the new features and correlation filters.

In this methods convolution operations are substituted by correlation ones, in frequency domain, and confidence map is computed in spatial domain. These methods have issues in orientation, scale adaption and shape because of changes in time and the selection of an efficient representation. Another important fact is that if the target is lost than it cannot be recovered. One of the best trackers of this class, as it will be shown later is ECO.

NCFTs trackers are usually divided in sub-classes:

- Patch learning trackers;
- Sparsity trackers;
- Superpixel trackers;
- Graph trackers;
- Multiple-instance-learning trackers;
- Part-based trackers;
- Siamese-based trackers.

For this assignment purpose only patch-learning and siamese-based tracker will be treat.

In patch-learning trackers target and background patches are exploited, these methods are trained with positive and negative samples and then tested with others. The highest response gives the object position. In particular, a Multi-Domain Network (MDNet) uses shared layers, those exploit generic target representation and one fully convolutional (FC) layer, that identifies the target with binary classification. Samples are generated on previous target location and FC layer is learned on first frame. When tracking fails a weight, adaption for FC layers is performed with again positive and negative samples, but from the current short-term interval.

Siamese-based deep trackers use matching mechanisms. The learning methods give the general object appearance variations. In this method the network tries to match the target with samples, searching for similarities between patches. SiameseMask is one of these trackers.

5.1 ECO

ECO is part of the family of the CFTs, in particular, of the Discriminative Correlation Filters (DCFs) based methods. DCFs trackers use robust scale estimation, multi-dimensional features, non-linear kernels, more complex learning models and long-term memory components to improve both accuracy and tracking speed. This method, unfortunately, requires larger models introducing the problem of over-fitting, with the risk of decreasing performances.To reduce risk

of over-fitting ECO introduce a factorized convolution formulation using the 80% less model parameters in case of deep features.
Another aspect is the use of a generative model of sample distribution in order to boost the diversity avoiding the requirement of a large sample set. In DCF approach the weights are typically set to goes exponentially:

$$\alpha_{\mathrm{j}} \sim (1-\gamma)^{M-j} \tag{1}$$

Where:

- j is the frame;
- γ is the learning rate;
- α_{j} is the sample with the smallest weight.

If the number of samples has reached a maximum limit the sample with the smallest weight α_{j} is replaced, requiring a large sample limit for a representative sample set. The ECO approach instead is based on joint probability distribution $p(x, y)$ of the sample feature maps x and corresponding outputs y. The idea is to search for the filter that minimises the excepted correlation error. Finally, the expected loss is approximated.

$$E(f) = \mathbb{E}\Bigg\{||S_{\mathrm{f}}\{x\} - y||^2_{L^2}\Bigg\} + \sum_{d=1}^{D} ||wf^d||^2_{L^2} \tag{2}$$

The model update idea is to use a sparse updating scheme, more common in Non-DCF (NDCF) tracker. In order to not require too much complexity, for every defined interval N_s a filter update is performed. So the N_s parameter determines how much often the filter is updated, where a $N_s = 1$ means one update for each frame increasing N_s the level of computation required decreases. Finally, a less frequent update of the filter stabilizes the learning especially in that scenarios where a new sample is affected by sudden changes (rotations, deformations, clutter and occlusions). In terms of computability the bottleneck is in the learning step.
ECO is a re-visitation of DCF methods to reduce complexity and over-fitting increasing accuracy and speed, reducing learning complexity and the memory required for it. ECO was one of the VOT2016, UAV123, OTB2015 and Temple-Color winners and it was published in CVPR in 2017.

5.2 RT-MDNet

MDNet is a visual tracking algorithm based on representation from a discriminatively trained Convolutional Neural Network (CNN). In CNN the biggest problem is to collect a large amount of training data and for this reason the Multi-Domain Network (MDNet) occurred. MDNet learns the shared representation of targets from multiple annotated video sequences and each of them is handled as separate domain. This approach consists in separate branches of domain-specific layers for binary classification, sharing at the end the common informations collected from all the sequences for generic representation learning. The training of each domain is performed iteratively and separately while shared layers are updated in every iteration.

The main reason of success of this method is given by the multi-domain learning framework, that divides domain-independent informations from domain-specific one, and then, the CNN, pre-trained by multi-domain learning, is update online to learn domain-specific information in an adaptive way.

The learning algorithm uses the so called Stochastic Gradient Descent (SGD) method where in the $k^{th} <$ iteration the network is updated based on a mini-batch consisting of the training samples where only a single branch is enabled. Then the process is iterated until the network converges or the maximum number of iterations defined is reached. It is possible to separate the update process in two components: the long-term updates, performed in regular intervals, and the short-term updates, applied whenever failures are detected. Another pro of MDNet is the use of bounding box regression techniques to improve target localization accuracy training a model to predict the target position.

With this approach the network is pre-trained offline while the domain-specific and the connected layers are fine-tuned online.

CNNs are effective in object tracking but most of them are very slow for a lot of applications and also MDNet suffers from this problem. In addition its algorithms are not optimized to distinguish target instances across multiple domains. In order to avoid this problems a new real-time object tracker is proposed: RT-MDNet. With this method a Region of Interest Alignment (RoIAlign) layer is introduced to extract target representation from a fully convolutional feature map. To maintain the representation capacity, the network has an high resolution feature map and enlarge the receptive field of each activation. It also uses an instance embedding loss in pre-training stage that runs near the MDNet binary foreground/background classification loss. Using a multi-task loss it is possible to discriminate object instances with similar semantics across multiple domains more efficiently.

The RoIAlign layer accelerate features extraction maintaining quality in representations but these features are inherently coarse compared to the ones from individual proposal bounding box. For this reason a denser fully convolutional feature map is computed and the field of each activation is enlarged. RoIAlign uses nearby grid points to compute the interpolated value so it is important to chose the right interval of the sampled points in order to not lose informations. In RT-MDNet this is done adaptively from the shared dense map grid.

The loss term enforce target objects in different domains to be embedded far from each other in a shared feature space and gives the possibility to learn discriminative representations of the unseen target objects in new test sequences.

RT-MDNet architecture is shown in [Fig.3].

MDNet was one of the winner of OTB2013, OTB2015, VOT2014 challenges and it was published in CVPR in 2016.

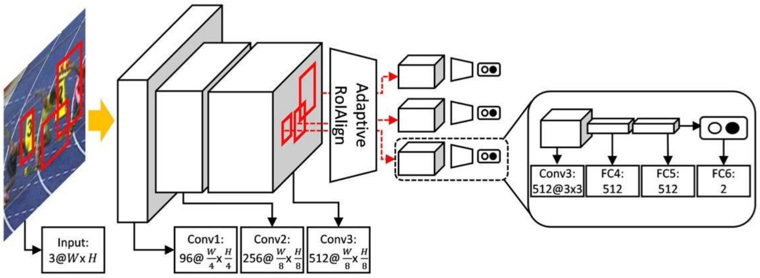

Fig. 3. RT-MDNet network architecture of the proposed tracking algorithm.

5.3 SiamMask

SiamMask is part of Siamese object tracking family and it is able to both target tracking and semi-supervised target segmentation with a single and real-time approach.
SiamMask uses the SiameseFC concept to implement them. SiameseFC is an offline-trained fully convolutional network which compares an exemplar image against a search image to get a dense response map. Here the goal is for the biggest value of the response map to match with the target location in the search area x. To allow each spatial element for the response map, the so called Response of a candidate Window (RoW), to encode more informations about the object, depth-wise cross-correlation is used producing a multi-channel response map. In SiamMask it is possible for the RoW of a fully-convolutional Siamese network to encode the needed informations to build a pixel-wise binary mask adding extra branch and loss to the ones of the Siamese trackers. During training steps each RoW is labelled with a ground-truth binary label and associated with a pixel-wise ground-truth mask. The classifiers indicate if a pixel is part of the object or not. During mask representation step this approach allows every pixel classifier to works with data contained in the entire RoW in order to have a complete view of its corresponding candidate window in x. This is of fundamental importance to disambiguate instances similar to the target. Finally to produce a more accurate object mask, low and high resolution features are merged using refinement modules.
The bounding boxes in SiamMask are of three types, shown in [Fig.4]:

- Axis-aligned bounding rectangle;
- Rotated minimum bounding rectangle;
- Automatic bounding box generation.

During tracking, SiamMask evaluates frame per frame without adaption. The output mask is selected by using location with highest score in the classification branch, then the output of the mask branch is binarized at a certain threshold. SiamMask architecture is shown in [Fig.5].

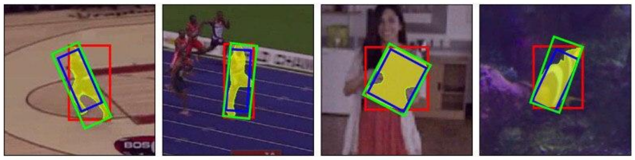

Fig. 4. Three different methods of representation in SiamMask.

SiameseFC was one of the winner of OTB2013, VOT2014, VOT2015, VOT2016 and it was published in ECCV in 2016.

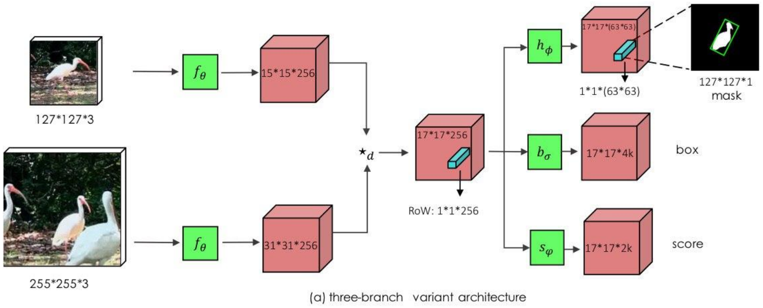

Fig. 5. SiamMask schematic representation of three-branch variant architecture.

6 Final evaluation

In this assignment different methods for object tracking are explained covering the main different strategies used to avoid the issues treated in the introduction paragraph.
The different methods have different major capabilities:

- ECO: it is a very precise and high success rate tracker but not very fast (8fps). It is good for not very fast features detection but with very high precision requirements. ECO improved C-COT issues and it is recognized as on of the most efficient trackers nowadays in terms of precision;
- RT-MDNet: it is NCFT tracker. This assignment has shown as NCFTs based methods are not so accurate as CFTs ones. RT-MDNet improved a lot the problems of MDNet with a real-time implementation;
- SiamMask: it is a very high speed tracker even losing some precision and success rate. This method is great for very fast features detection and it is able to provide a more real representation of the object with three types of representation.

In conclusion, as mentioned before, there are a lot of good methods nowadays in object tracking science and the choice of using one of them is related to the application. It is possible to identify some methods greater than the others but, due to the fact that they have very different bases, results can be different. For this reason the assignment bases the studies on the benchmarks, which provide a standard for object tracking science.
All references used in this assignment ([1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]) are listed below.

References

1. Benchmark results. https://github.com/foolwood/benchmark_results, last accessed 02/11/19
2. Eco github. https://github.com/martin-danelljan/ECO, last accessed 02/11/19
3. Eco homepage. http://www.cvl.isy.liu.se/research/objrec/visualtracking/ecotrack/index.html, last accessed 02/11/19
4. Eco supplemental material. http://openaccess.thecvf.com/content_cvpr_2017/ supplemental/
Danelljan_ECO_Efficient_Convolution_2017_CVPR_supplemental.pdf, last accessed 02/11/19
5. Ecopaper. http://openaccess.thecvf.com/content_cvpr_2017/papers/
Danelljan_ECO_Efficient_Convolution_CVPR_2017_paper.pdf, last accessed 02/11/19
6. Handcrafted and deep trackers. https://arxiv.org/pdf/1812.07368.pdf, last accessed 02/11/19
7. Mdnet github. https://github.com/HyeonseobNam/MDNet, last accessed 02/11/19
8. Mdnet homepage. http://cvlab.postech.ac.kr/research/mdnet/, last accessed 02/11/19
9. Mdnet paper. https://arxiv.org/pdf/1510.07945v2.pdf, last accessed 02/11/19
10. Mdnet vot2015. http://votchallenge.net/vot2015/download/
presentation_Hyeonseob.pdf, last accessed 02/11/19
11. Otb homepage. https://ieeexplore.ieee.org/document/7001050, Last accessed 02/11/19
12. Rt-mdnet github. https://github.com/IlchaeJung/RT-MDNet, last accessed 02/11/19
13. Rt-mdnet paper. http://openaccess.thecvf.com/content_ECCV_2018/papers
/Ilchae_Jung_Real-Time_MDNet_ECCV_2018_paper.pdf, last accessed 02/11/19
14. Siamesefc github. https://github.com/bertinetto/siamese-fc, last accessed 02/11/19
15. Siamesefc paper. http://www.robots.ox.ac.uk/ luca/siamese-fc.html, last accessed 02/11/19
16. Siammask github. https://github.com/foolwood/SiamMask, last accessed 02/11/19
17. Siammask homepage. http://www.robots.ox.ac.uk/ qwang/SiamMask/, last accessed 02/11/19
18. Siammask paper. https://arxiv.org/pdf/1812.05050.pdf, last accessed 02/11/19
19. Vot2019 challenge results. http://openaccess.thecvf.com/content_ICCVW$_2$019/
papers/VOT/Kristan_The_Seventh_Visual_Object_Tracking_VOT
2019*_Challenge_Results_ICCVW_*2019*_paper.pdf, lastaccessed*02/11/19