Mathematics for Computer Vision (MCV 2018)
February 15-16th 2018, Trento, ITALY
Session 1: Mathematics for Machine Learning

Learning to Track

Oswald Lanz

February 15, 2018

Abstract In this talk we will revisit earlier work on people/object tracking and show how it can be framed into a coherent picture within a modern deep learning approach. We will consider invariance and equivariance as mathematical principles to derive our recurrent deep network architecture for multi-object tracking.
Keywords: Object Tracking, Deep Learning.

1 Introduction

Back in '03 when I started my PhD at the Fondazione Bruno Kessler (the ITC-irst at that time), I found my research inserted in the context of the PEACH Project [1]. The project objective was that of studying and experimenting with advanced technologies that can enhance cultural heritage appreciation. I was intrigued by the opportunities of tracing visiting patterns in an exhibition space (such as a museum), and I targeted my research at scalable algorithms for people tracking in camera networks. In particular, I focussed on occlusion handling as a key challenge in tracking individuals that move within groups [2], and how it can be scaled to large environments through task decomposition [3]. I focussed on inference and did not care much about features and representations of the patterns (people) to track. I implemented a rigid shape model to represent the body of a person, and used color histograms to characterize the appearance of each target [4]. I modeled photometric distortions (due to uneven illumination) as a function of position in the scene and learned its parameters from data [5]. I coded the influence of other targets (group members) and scene hotspots (exhibits) as a function of their relative position and pose [6]. We recently also used pose estimation [7] in combination with tracking to detect groups of interacting people [8].

Much of the research done at that time shares similar structure: hand-crafted features and shallow inference on top of them, with domain knowledge encoded in their design. This has substantially changed with the rise of deep learning: Data should explain it all. And indeed, it does so a lot better on most computer vision problems and in AI. Modern approaches model the process transforming raw input signal to desired outputs with a single global function, that is, a composite network of parametric differentiable modules that are numerically optimized all the way down to the raw signal input to reproduce output responses from annotated samples. Put in relation to the above, inference is no longer constrained by the sub-optimal design of features, representations and process models, the network is trained to develop its own internal representations to best fit the data with the given capacity. With large enough datasets and computing resources, it is not too surprising that they outperform (sometimes beating the human) on complex recognition tasks in supervised settings. However, when the underlying process has some understood structure such as in multi-object tracking, then this structure should reflect the design of deep architectures for end-to-end training to be effective.

Overview. In this presentation I will revisit earlier work on multi-object tracking in the attempt to frame it into a modern deep learning framework. I will set the proper context by overviewing recent work on object detection and tracking, and put a methodological focus on recurrent neural networks for sequences and - more in general - structured output prediction. I will present few works from the state of the art in more detail, and

Oswald Lanz
Fondazione Bruno Kessler, Trento, Italy. E-mail: {lanz}@fbk.eu

show how they can be put into a coherent picture for addressing multi-object tracking with a novel deep learning approach. I will try to give proper methodological (when possible mathematical) justification that lead to the development of the approach.

2 Literature

Object detection and tracking has been extensively studied in computer vision. We will glimpse over some works representative of the recent progress in the field, to the detail useful to introduce the technical part of the talk.

Object detection. Convolutional neural nets (CNNs) have set the state of the art in object detection from images. Any tracking system needs a detection component. We will analyze SSD [9] that predicts object bounding boxes and object category scores from images with a single network, in a single shot. Others have investigated object detection in video [10], and people detection from multiple views [11].

Similarity learning. Tracking in contrast to detection, seeks for correlations across frames to follow a target in an image sequence. Given an exemplar image of a target, a similarity function can be used for online tracking: with each new incoming frame the target is searched around the previous location based on a similarity score. A similarity function can be learned from data in the form of a siamese network, that is, two branches of CNNs with shared parameters [12]. Siamese networks can also be trained to provide features that are tightly coupled to correlation filters for tracking [13], and to jointly track and detect multiple targets [14].

Tracking by detection. The currently predominant approach to multi-object tracking is tracking by detection. In a first step, object detectors provide potential locations of the objects of interest in the form of bounding boxes. Then, the task of multi-object tracking translates into a data association problem where the bounding boxes are assigned to trajectories that describe the path of individual object instances over time. In online methods, the association to bounding boxes in the incoming frame is often formulated as bipartite graph matching and solved via the Hungarian algorithm [15].

Structured prediction. Recurrent networks are popular with sequence prediction tasks such as image/video captioning, text/speech generation, machine translation. Long-short term memory (LSTM) or gated recurrent unit (GRU) are the most effective to account for long-term dependencies, especially when they use attention [16]. Naturally any video has sequence structure and recurrent models have been developed for multi-object tracking [17] and trajectory prediction [18]. In some domains there is even richer output structure to exploit, examples are tree-structured LSTM network in natural language processing [19] and graph neural networks [20] for semantic object parsing [21].

Higher-order tracking. Online tracking methods typically update object locations sequentially, using current estimates and information from the next frame. To increase robustness and accuracy one can look ahead several frames to update the current estimates. Multiple hypotheses tracking propagates object location hypotheses over a future time window, by collecting future object detections in a tree structure and pruning unsupported branches at the current time [22].

Global optimization. Multi-object tracking can be solved globally by constructing a data association graph and solving the associated network flow problem [23]. Recent work [24] replaces hand-crafted graph potentials with layered parametric functions and translates network flow into a differentiable function of potentials, in a way the resulting architecture can be trained end-to-end via backpropagation. Instead of using graphs to represent trajectories discretely, the authors in [25] perform data association and trajectory model fitting in a discrete-continuous energy minimization.

3 Preliminaries

Definition 1 (Video objects) A set of video frame sequences $\mathbf{V} = \{v(t) \in I\!R^{N_v \times M_v \times 3} \,|\, t \in [T_v]\}_v$ paired with annotations $\mathbf{L} = \{(c, i, t, x) \in \mathcal{C} \times I\!N \times [T_v] \times [N_v, M_v]^2\}_v$, where $\mathcal{C}$ is the object class label set, $i \in I\!N$ indexes an object instance, T_v is the length of video v and N_v, M_v is its spatial resolution, $[T] = 1, 2, \ldots, T$.

Video frames are photometric measurements of physical objects $\{O_i\}$ composing a scene that undergo a complex digital image formation process. The resulting random variable is $v(t) = g(X(\{O_i\}, t) + s) + r$ where g is a partial model of the process, X is the object deformations, and r, s are noise terms for object deformations and image formation residuals. g encompasses radiometric and geometric processes by which 2D images of 3D objects are formed. Annotations group objects into semantic categories and associate measurements of each object instance with their bounding boxes envelopes on the image plane. Video objects are typically provided by object tracking benchmarks in form of a training set.

Definition 2 (Object tracking) Given video objects $\mathbf{VL}$, object tracking over an unseen video q produces structured predictions $f(q \,|\, \mathbf{VL}) = \langle \mathcal{N}, \mathcal{E} \rangle_q$ with node set $\mathcal{N}$ and edge set $\mathcal{E}$. A node $(c, t, x) \in \mathcal{C} \times [T_q] \times [N_q, M_q]^2$ represents the location x of a video object with class label c in $q(t)$. An edge $e \in [\#\mathcal{N}]^2$ encodes three types of dependencies among node pairs: same

object instance, occlusion, other interaction. In online tracking f produces a t-directed acyclic graph.

Output graphs can have diverse structure. A set of tracklets (node sequences) is produced with 1st order methods, when data association is sought among adjacent frames with object interactions neglected, or not exposed. Occlusion is interaction at appearance level with one object blocking measurement(s) of the object(s) behind, and thus injects a set of directed edges in the graph. Other dependencies that make the output structure denser is correlated motion, when people interact in groups and/or with hotspots in the scene, or when scene obstacles interfer with object trajectories. An output graph can be seen as a symbolic-numeric representation of the data, q, produced by g (image formation process) and X (objects deformation process).

Definition 3 (Learning to track) To realize object tracking with deep learning, a network f of parametric differentiable modules is optimized to reproduce video objects $\mathbf{VL}$. With a differentiable loss $\mathcal{L}$ quantifying prediction accuracy, the $\theta_{\mathbf{VL}} = \text{argmax}\, \mathcal{L}(\mathbf{L}, f(\mathbf{V} \,|\, \theta))$ is found by numerical optimization. Predictions on a test video $q \notin \mathbf{V}$ are then computed with fixing parameters, by $\langle N, E \rangle_q = f(q \,|\, \theta_{\mathbf{VL}})$.

Most research is about designing $f, \mathcal{L}$ in a way the iterative mini-batch optimization $\theta \leftarrow \theta - \eta \cdot \nabla \mathcal{L}(\mathbf{l}_i, f(\mathbf{v}_i \,|\, \theta))$ scales to large video objects ($\{\mathbf{vl}\}_i$ is a partition of $\mathbf{VL}$) in terms of memory footprint, computation time and prediction accuracy. Arguably, generalization capability is strongly related to how well f resembles the structure of the underlying generative process $g \circ X$, and a best model would be of minimum capacity (fewest parameters) while most expressive (lowest residuals).

Conjecture 1 (Generalization) *A deep network for object tracking should have siamese structure at the bottom layers and recurrent structure at the top. It should be compositional in the latent embedding connecting the two, with a built-in gating mechanism controlling object interactions. These include occlusion that is realized in the perspective projection generating the data.*

Our multi-object tracking approach develops alongside the conjecture. Mathematical foundations and Information Theory of deep learning are discussed in [26,27].

References

1. O. Stock and M. Zancanaro. *PEACH - Intelligent Interfaces for Museum Visits*. Springer, 2007.
2. O. Lanz. Approximate bayesian multibody tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1436–1449, 2006.
3. T. Hu, S. Messelodi, and O. Lanz. Dynamic task decomposition for decentralized object tracking in complex scenes. *Comput. Vis. Image Underst.*, 134:89–104, 2015.
4. O. Lanz, P. Chippendale, and R. Brunelli. An appearance-based particle filter for visual tracking in smart rooms. In *Int. Eval. Workshops CLEAR*, 2007.
5. S. Mutlu, T. Hu, and O. Lanz. Learning the scene illumination for color-based people tracking in dynamic environment. In *ICIAP*, 2013.
6. G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: Towards socially and personality aware visual surveillance. In *ACM Workshop MPVA*, 2010.
7. Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe. A multi-task learning framework for head pose estimation under target motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(6):1070–1083, 2016.
8. J. Varadarajan, R. Subramanian, S. Rota Bulò, N. Ahuja, O. Lanz, and E. Ricci. Joint estimation of human pose and conversational groups from social scenes. *Int. J. Comput. Vis.*, 2017.
9. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S.E. Reed, C.-Y. Fu, and A.C. Berg. SSD: single shot multibox detector. In *ECCV*, 2016.
10. Y. Lu, C. Lu, and C.-K. Tang. Online video object detection using association lstm. In *ICCV*, 2017.
11. P. Baque, F. Fleuret, and P. Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *ICCV*, 2017.
12. L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, and P.H.S. Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, 2016.
13. J. Valmadre, L. Bertinetto, J.F. Henriques, A. Vedaldi, and P.H.S. Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, 2017.
14. C. Feichtenhofer, A. Pinz, and A. Zisserman. Detect to track and track to detect. In *ICCV*, 2017.
15. C. Huang, Y. Li, and R. Nevatia. Multiple target tracking by learning-based hierarchical association of detection responses. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(4):898–910, 2013.
16. K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R.S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.
17. A. Milan, S.H. Rezatofighi, A.R. Dick, I.D. Reid, and K. Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI*, 2017.
18. A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, F.-F. Li, and S. Savarese. Social LSTM: human trajectory prediction in crowded spaces. In *CVPR*, 2016.
19. K.S. Tai, R. Socher, and C.D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*, 2015.
20. F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009.
21. X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan. Semantic object parsing with graph LSTM. *arXiv:1603.07063*, 2016.
22. C. Kim, F. Li, A. Ciptadi, and J.M. Rehg. Multiple hypothesis tracking revisited. In *ICCV*, 2015.
23. L. Zhang, Y. Li, and R. Nevatia. Global data association multi-object tracking using network flows. In *CVPR*, 2008.
24. S. Schulter, P. Vernaza, W. Choi, and M. Chandraker. Deep network flow for multi-object tracking. In *CVPR*, 2017.
25. A. Milan, K. Schindler, and S. Roth. Multi-target tracking by discrete-continuous energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):2054–2068, 2016.
26. R. Vidal, J. Bruna, R. Giryes, and S. Soatto. Mathematics of neural networks. *arXiv:1712.04741*, 2017.
27. R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv:1703.00810*, 2017.