PhD Proposal: Deep Learning methods for human behavior monitoring

INRIA Sophia Antipolis, STARS group
2004, route des Lucioles, BP93
06902 Sophia Antipolis Cedex – France
http://www-sop.inria.fr/members/Francois.Bremond/

1. Scientific context

STARS group works on automatic video monitoring and human behavior understanding for health applications. The Deep Learning platform developed in STARS, detects mobile objects, tracks their trajectory and recognizes related behaviors predefined by experts. This platform contains several techniques for the detection of people and for the recognition of human postures/gestures using conventional cameras. However, there are scientific challenges in people tracking when dealing with real word scenes: cluttered scenes, handling wrong and incomplete person segmentation, handling static and dynamic occlusions, low contrasted objects, moving contextual objects (e.g. chairs), similar appearance of clothes among different people ...

Multiple Object Tracking (MOT) is a fundamental task that aims at associating the same objects across multiple frames in a video clip. A robust and accurate MOT algorithm is indispensable in broad applications, such as people monitoring and video surveillance. An end-to-end MOT algorithm can be divided into three different but closely related tasks; single frame detection of objects, short term tracking and long-term tracking of said objects, the latter two are usually merged together into a problem commonly known as data association. This gave rise to the dominant paradigm in MOT, tracking-by-detection, which first obtains bounding boxes by detection frame by frame, and then generates trajectories by associating the same objects between frames. While these tasks are part of the same MOT problem, they are often treated apart, either trained separately or the data association step is not a deep learning-based approach which hinders the whole process.

On top of the aforementioned issue of separated training, short term tracking and long-term tracking have the same objective (data association) but they have different inputs. Short term tracking deals with per frame feature representation of an object and long-term tracking needs to deal with a historic feature representation that encapsulates the myriad of changes of an object across a larger frame span. In other words, we need a memory that tracks said changes, that is differentiable and can back-propagate the information all the way up to the detection task.

2. General objectives of the PhD

This work consists in designing efficient People Joint Detection and Tracking algorithms. One potential approach could use differentiable Memory Banks to build a Deep Learning memory-based architecture that can be trained to learn a feature representation of a tracklet. Therefore,

the main difference with respect to the current state-of-the-art is that this MemoryTracker will be conceived to mitigate the loss of information from training separately both detection, short term tracking and long-term tracking tasks. Designing an efficient memory-based architecture is far from evident. Indeed, the first challenge is to be able to infer dense representations (i.e. tracklet vectors). To do so, we propose the use of ROI-alignment from the pipeline of deformable DETR detector. We also can take advantage of joint detection and short-term tracking by using 3D CNNs, this can allow us to have temporal and spatial information that is not available with vanilla 2D CNNs. The use of 3DCNNs can output more reliable tracklets over a small number of frames and use that information to better update the MemoryBank.
In addition to allowing a truly end-to-end pipeline, the MemoryTracker could overcome the batch training problem by storing the tracklet feature vector with an intra-batch loss and an out-of-batch loss. Both losses could be based on triplet loss functions that depend on the current input sequence (intra batch) and the following sequences (out-of-batch). However, while the features of the current frames are given to the detection pipeline, the features of the previous frames are given to the MemoryBank.

To validate the work, we will assess the proposed algorithms on video-monitoring applications and homecare videos from Nice Hospital and from public places, such as the ones in MOT20 https://motchallenge.net/data/MOT20/.

3. Pre-requisites

Master 2 (or Engineer) in Computer Vision or Mathematics,
With theoretical knowledge in Computer Vision, Mathematics, and Deep Learning (PyTorch, TensorFlow), and technical background in C++ and Python programming, Linux.

Place of PhD: Inria Sophia Antipolis

4. Schedule

1st year:

- Study the limitations of existing DL People Tracking algorithms.
- Proposing a new approach for People Tracking using Joint Detection and Tracking.

2nd year:

- Start to Improve the proposed DL People Tracking approach.
- Writing papers

3rd year:

- Evaluate, improve and optimize proposed DL People Tracking approach.
- Writing papers and PhD manuscript.

5. Contact

Francois.Bremond@inria.fr

6. Bibliography

1. JD. Zuniga, Ujjwal and F. Bremond. DeTracker: A Joint Detection and Tracking Framework. In Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP 2022, Virtual, February 6-8, 2022.
2. Berclaz, J., Fleuret, F., Turetken, E., and Fua, P. (2011). Multiple object tracking using k-shortest paths optimization. IEEE transactions on pattern analysis and machine intelligence.
3. Bergmann, P., Meinhardt, T., and Leal-Taix´e, L. (2019). Tracking without bells and whistles. In the IEEE International Conference on Computer Vision (ICCV).
4. Chu, P. and Ling, H. (2019). Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6171–6180.
5. Dehghan, A., Modiri Assari, S., and Shah, M. (2015). Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4091–4099.
6. Dong, X. and Shen, J. (2018). Triplet loss in siamese network for object tracking. In Proceedings of the European Conference on Computer Vision (ECCV), pages 459–474.
7. Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., and Cucchiara, R. (2018). Learning to detect and track visible and occluded body joints in a virtual world. In European Conference on Computer Vision (ECCV).
8. Feichtenhofer, C., Pinz, A., and Zisserman, A. (2017). Detect to track and track to detect. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 3057–3065.
9. Feng, W., Hu, Z., Wu, W., Yan, J., and Ouyang, W. (2019). Multi-object tracking with multiple cues and switcheraware classification. CoRR, abs/1901.06129.
10. He, K., Gkioxari, G., Doll´ar, P., and Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969.