Big Brother can you find, classify, detect and track us ?

Marc Décombas [1] *, Jean Benoit Delbrouck* [1]

TCTS Lab - University of Mons, Belgium
{marc.decombas, jeanbenoit.delbrouck} @gmail.com

2. Project objectives (max. 1 page): providing the rationale for the proposed project.

In this project, we will build a system that can detect, recognize objects or humans and describe them as much as possible on video. Objects may be moving as well as the people coming in and out of the visual field of the camera(s). Our project will be split into three main tasks :

- detection and tracking
- people re-identification
- image/video captioning

The system should work in real time and should be able to detect people and follow them, re-identify them when they come back in the field and give a textual description of what each people is doing.

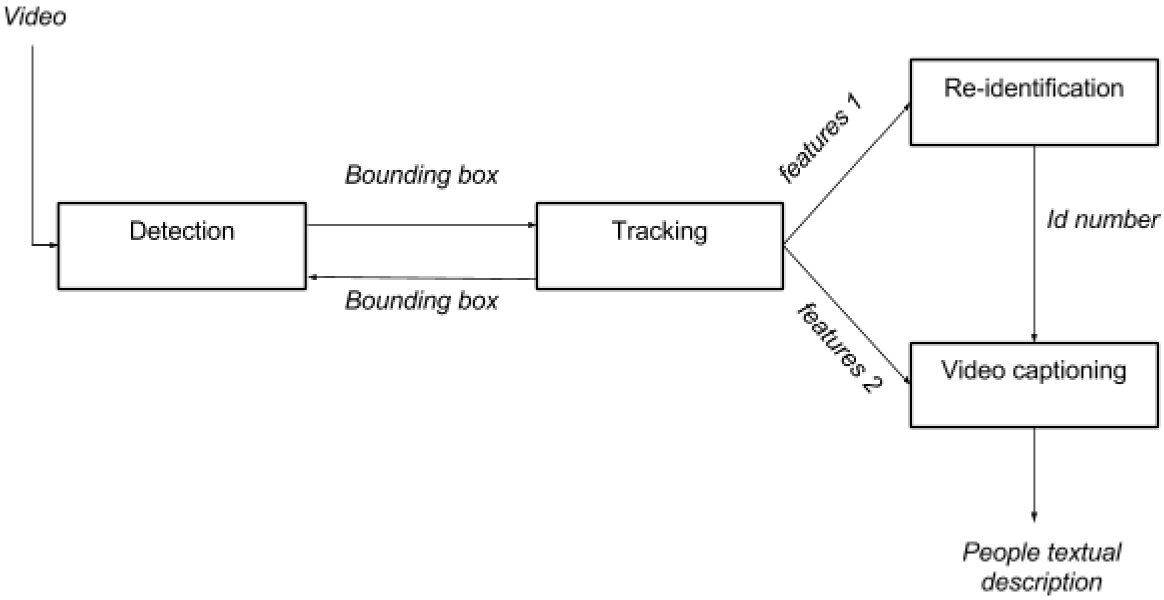
3. Background information

Deep neural networks became really popular in image classification since 2012 with the breakthrough of Alex Krizhevsky [3]. He proposes a deep neural network that achieves top-1 and top-5 test set error rates of 37.5% and 17.0% compared to previous work that gives 45.7% at top-1 and 25.7% at top-5 [4] on Imagenet - ILSVRC-2010 . Since then, deep learning has also improved results for tasks like detection [5] [6], tracking [7] [8],people re-identification [9][10].

Real time video captioning [21][22] has recently been made through Hierarchical Recurrent Neural Network (H-RNN). It consist of a RNN for language modeling, a multimodal layer [24] for integrating information from different sources and an attention model [23]. Cells use for natural language processing are usually LSTM or GRU.

4. Detailed technical description (max. 3 pages):

The global project can be described as follow :



From a video, a detection algorithm will extract people position but also objects that could be useful to describe and differentiate it from one another. The detection algorithm will be used to initialize the tracking algorithm that will generate bounding box for the next frames. These bounding boxes will be compared with the proposal bounding box from the detectors to obtain a more robust object detectors in the video. The trackers will also gives temporal correlation between the bounding boxes. From the tracks obtained, it will be possible to extract features (spatial and temporal) to define a unique id for each person. This way, it will be possible to re-identify people that appears several times in the video field of the camera(s). With this unique id and others features extracted from the tracks, it will be possible to do video captioning and realize a textual description of each person.

In more details, for the detection part, the system in [6] will be used due to its high performance to detect and it is faster than the real time. Indeed, it accurately classifies objects in the visual modality on 20 classes with VOC 2007+2012 dataset [11] but most the classes are not useful. A new training will be done on specific classes based on the new dataset of [12].

This detection will be improved by adding recurrent neural networks [13] which can take into account the temporal nature of the data. Results based on this idea are presented in [14] and can be done with the darknet framework [15]. Temporal information about the position of the objects can be obtained by the help of the tracking part.

For the tracking part, we will use [8] and improve the system to manage multi-object tracking. The tracker can be initialized and improved by the detection part. The work in [16] has given the best results in the VOT challenge [17]. It will be interesting to see if it is possible to integrate some ideas in [8] while keeping real time tracking. Moreover, Youtube just shared a new dataset [18]

For re-identification, it will be possible to start with [10] and try to use the temporal information provided by trackers and the detection. As in [8], siamese network can be used to do re-identification [19]. It will be studied a way to integrate this work in [10] and [8]. As it is also possible to obtain information about the project that wears people, we will see how much this high level information is useful.

For the image/video captioning, the plan is to use the Visual Genome dataset to train and compare different DNN based systems. CNN, RNN/LSTM and Siamese networks will be considered in the first place for this task.

5. Work plan and implementation schedule (max. 1 page): a tentative timetable detailing the work to be done during the workshop.

Workpackage	Week 1	Week 2	Week 3	Week 4
Detection & Tracking	X	X	X	
Re-Identification	X	X	X	
Video captioning	X	X	X	
Integration			X	X

6. Benefits of the research

The expected outcome of this project is an end to end solution to detect, track, identify and re-identify people in order to obtain an individual video captioning. This solution could work as a web service and in real time.

A secondary outcome will be a state of the art framework that combines different open source solutions. This way, it will be possible to see how far are we to have a solution like big brother.

7. Profile team

- CV Marc Decombas (leader)
- CV Jean Benoit Delbrouck (leader)
- Researchers needed :
    - Background in deep learning or computer vision
    - knowledge in detection, tracking, visual descriptor, re-identification, CBIR
    - Knowledge in docker, API, server

8.References

[1] Rise of Surveillance Camera Installed Base Slows". 5 May 2016. Retrieved 5 January 2017.

[2] "CCTV in London" (PDF). Retrieved 2009-07-22.

[3] A. KRIZHEVSKY, I. SUTSKEVER and G. HINTON, Imagenet classification with deep convolutional neural networks. In : *Advances in neural information processing systems*. 2012. p. 1097-1105.

[4] J. Sánchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1665–1672. IEEE, 2011.

[5] REDMON, Joseph, DIVVALA, Santosh, GIRSHICK, Ross, *et al.* You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015.

[6] REDMON, Joseph et FARHADI, Ali. YOLO9000: Better, Faster, Stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[7] D. Held, and Thrun, Sebastian and Savarese, Silvio, Learning to Track at 100 FPS with Deep Regression Networks, European Conference Computer Vision (ECCV), 2016

[8] BERTINETTO, Luca, VALMADRE, Jack, HENRIQUES, João F., *et al.* Fully-convolutional siamese networks for object tracking. In : *European Conference on Computer Vision*. Springer International Publishing, 2016. p. 850-865.

[9] CAMPS, Octavia, GOU, Mengran, HEBBLE, Tom, *et al.* From the Lab to the Real World: Re-Identification in an Airport Camera Network.

[10] XIAO, Tong, LI, Hongsheng, OUYANG, Wanli, *et al.* Learning deep feature representations with domain guided dropout for person re-identification. *arXiv preprint arXiv:1604.07528*, 2016.

[11] http://host.robots.ox.ac.uk/pascal/VOC/

[12] http://visualgenome.org/

[13] http://pjreddie.com/darknet/rnns-in-darknet/

[14] NING, Guanghan, ZHANG, Zhi, HUANG, Chen, *et al.* Spatially Supervised Recurrent Convolutional Neural Networks for Visual Object Tracking. *arXiv preprint arXiv:1607.05781*, 2016.

[15] https://github.com/pjreddie/darknet

[16] NAM, Hyeonseob, BAEK, Mooyeol, et HAN, Bohyung. Modeling and propagating cnns in a tree structure for visual tracking. *arXiv preprint arXiv:1608.07242*, 2016.

[17] http://www.votchallenge.net/

[18] ZHANG, Li, XIANG, Tao, et GONG, Shaogang. Learning a discriminative null space for person re-identification. In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. p. 1239-1248.

[19] https://research.googleblog.com/2017/02/advancing-research-on-video.html?m=1

[20] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, Yueting Zhuang; Hierarchical Recurrent Neural Encoder for Video Representation With Application to Captioning ; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1029-1038

[21] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, Yueting Zhuang; Hierarchical Recurrent Neural Encoder for Video Representation With Application to Captioning ; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1029-1038

[22] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, Wei Xu; Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4584-4593

[23] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations, 2015.

[24] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). ICLR, 2015.