An Empirical Analysis of Visual Features for Multiple Object Tracking in Urban Scenes

Mehdi Miah, Justine Pepin, Nicolas Saunier and Guillaume-Alexandre Bilodeau

Polytechnique Montréal, Canada

Motivation

The most popular paradigm in multiple object tracking (MOT) is tracking-by-detection where objects of interest are first detected and then associated. For the association step, the appearance, spatial and motion information are jointly used. Consequently, it is not clear which visual features are the best. Our objective is to rank several visual features for MOT focused on urban scenes according to the performance of the detector.

Methodology

To measure the ability of a descriptor to correctly retrieve an object from a set of objects , we tried to link two bounding boxes referring to the same object throughout a video using some affinity measures. Working with the true boxes gives access to the true identity of the objects and prevents the apparition of biases such as detecting only large objects.
Moreover, to simulate detections, we introduced noise in two ways. Firstly, by adding a white Gaussian noise (parametrized by a standard deviation sigma) to the bounding boxes coordinates. Secondly, by skipping some frames (parametrized by a sampling step).

The visual descriptors are :

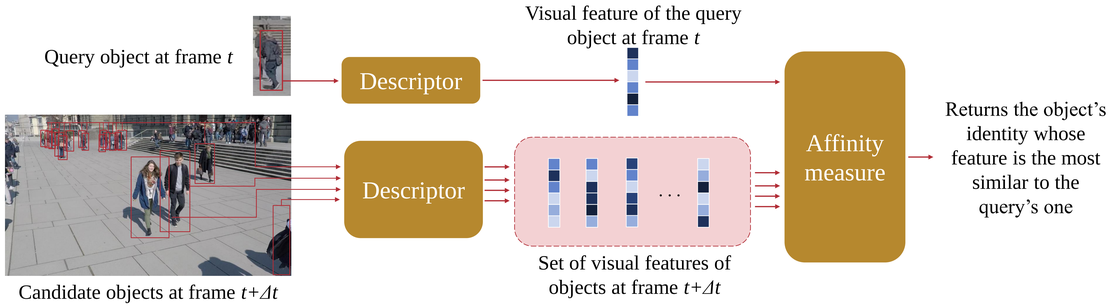
- color-based histogram descriptors : RGB (**RGB**) or grayscale (**GR**) ;
- histogram of oriented gradients (**HOG**) ;
- CNN-based features : ResNet-18 (**RSN**), VGG-19 (**VGG**), DenseNet-121 (**DNS**) and EfficientNet-B0 (**EFF**) representations ;
- re-identification (reID) descriptor : OSNet-AIN [1] for persons (**OSN**) and the model of [2] for vehicles (**VID**).

The affinity measures are :

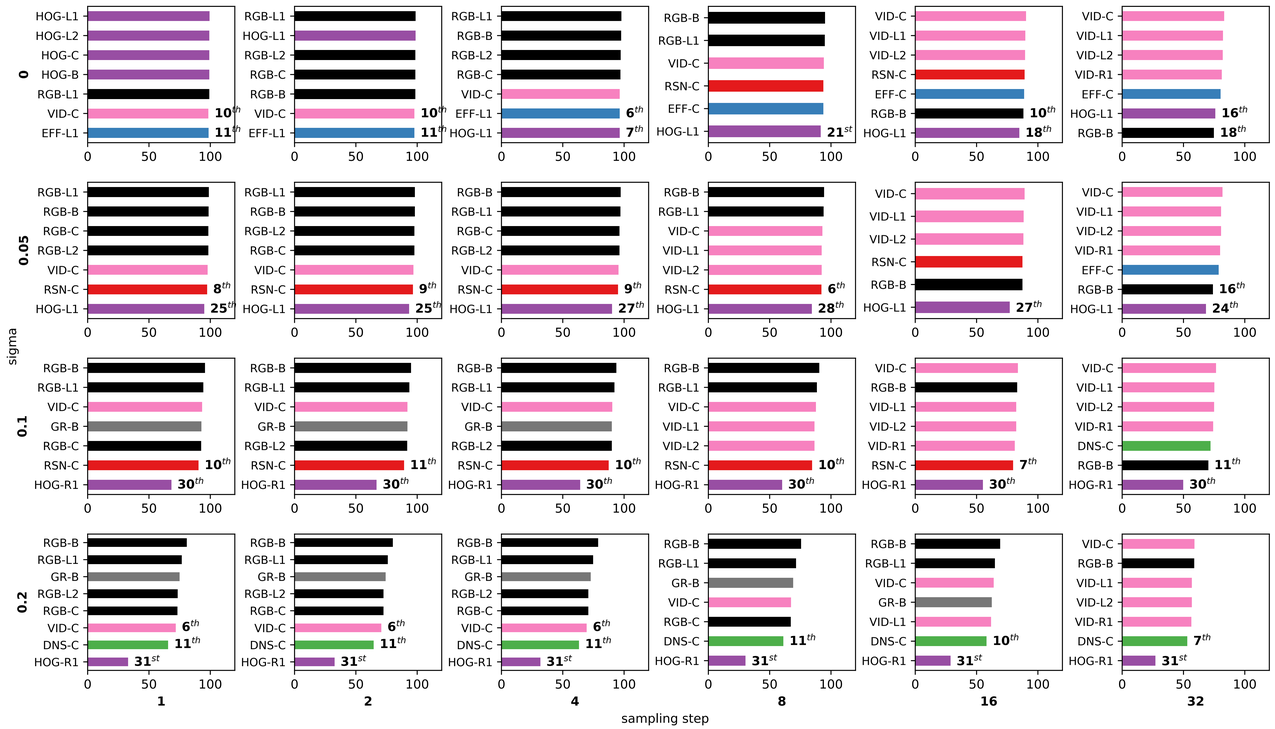
- L_1 and L_2 distances (**L1**, **L2**)
- Rank-1 counts (**R1**) ;
- Bhattacharyya distance (**B**, for histogram-based features) ;
- cosine similarity (**C**)

This leads to 35 pairs descriptor-affinity. We evaluated on four MOT datasets : DETRAC, UAVDT, MOT17 and WildTrack.

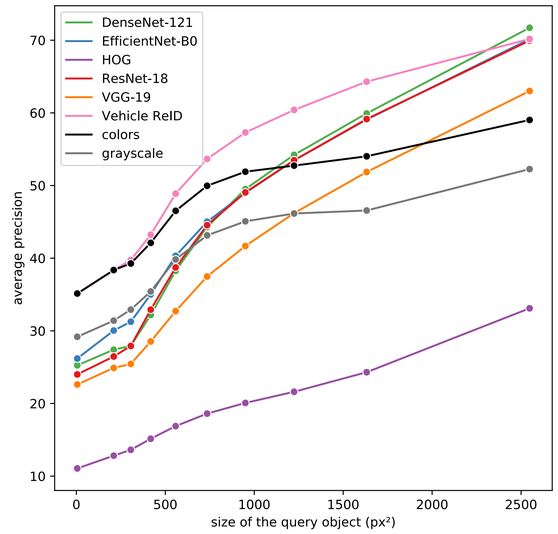
High-level explanation of the experimental methodology



Average precision on UAVDT



Effect of the query object size on UAVDT



Conclusion

ReID features with cosine similarity are one of the best descriptors for pedestrians and vehicles, regardless of the spatial precision and recall of the detector. When the boxes are not too noisy, color histograms with the Bhattacharyya distance are a good choice. Moreover, the size of objects matters on the choice of visual features.

References

[1] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning Generalisable Omni-Scale Representations for Person Re-Identification," *arXiv:1910.06827*, Oct. 2019.

[2] C.-W. Wu, C.-T. Liu, C.-E. Chiang, W.-C. Tu, and S.-Y. Chien, "Vehicle Re-Identification With the Space-Time Prior," in *CVPR - Workshops*, 2018.

Acknowledgements

Arcturus empower embedded

NSERC CRSNG