Assessing Post-Detection Filters for a Generic Pedestrian Detector in a Tracking-By-Detection Scheme

Volker Eiselein, Erik Bochinski and Thomas Sikora
Communication Systems Group, Technische Universität Berlin
{eiselein,bochinski}@nue.tu-berlin.de

Abstract

Tracking-by-detection becomes more and more popular for visual pedestrian tracking applications. However, it requires accurate and reliable detections in order to obtain good results. In this work, we propose two different post-detection filters designed to enhance the performance of custom person detectors. Using a popular deformable-parts-based pedestrian detector as a baseline, a detailed comparison over multiple test videos is performed and the gain of both algorithms is proven. Further analysis shows that the improved detection outcomes also lead to improved tracking results. We thus found that the usage of the proposed post-detection filters is recommendable as they do not impose a high computational load and are not limited to a specific detector method.

1. Introduction

The extraction of semantic information from videos is of major interest for many applications, especially for videos containing objects of special interest, such as humans. The process of tracking these objects over multiple frames allows deriving descriptions of their actions and potential relationships and is known as multi-target tracking and a key technology to applications such as e.g. forensics, sports or traffic analysis and activity recognition. It is also important in the detection of specific events such as loitering or anomalous trajectories, which are common use cases in the surveillance domain. In its usual definition, multi-target tracking seeks to identify both an unknown number of targets in a video and their respective paths.

For this work, we focus on visual pedestrian tracking in surveillance scenarios which is commonly based on automatic pedestrian detection and a tracking algorithm.

Pedestrian detectors are manifold and usually focus on gradient distributions in the image. Hand-crafted feature representations include the well-known histograms of oriented gradients as proposed in [4] and a huge number of extensions [9, 6, 20]. With the rise of modern deep learning techniques, convolutional neural networks have also been applied to pedestrian detection and shown their ability to obtain good results [12, 2].

For visual tracking, a huge number of approaches have been proposed which often exploit color features [16], gradients [25] or a fusion of features [15] in order to identify an object's position in the next frame. Other methods use correlation [5] or kernel approaches [13] in order to derive robust object representations and combine the features found to tracks. An overview of different techniques for person tracking can be found in [23].

Visual trackers, however, often suffer from the drifting problem which means that the image features extracted for training may contain background information and thus, with every update step over time, the object representation in the tracker can become worse. Nonetheless, not performing an update of the object representation is also undesirable as an object can appear very differently over a video depending on pose, local illumination, contrast and so on.

This dilemma led to the application of tracking-by-detection methods which have gained significant interest in the community. Among these methods are batch algorithms [19] which take a sequence of frames together with their respective detections and estimate the tracks for this time frame. Another option is the usage of on-line methods [3, 7] which estimate the tracks according to the available information in each frame but lack knowledge about detections in future time steps. Therefore, these methods can be very sensitive to missed detections. As in general setups, the number of pedestrians in the scene is unknown, the tracking algorithm must provide a way of dealing with both false positive and false negative detections.

One of the first works which coupled the detection and tracking processes in order to obtain a better tracking performance was [1] where a Gaussian process latent variable model is used to improve hypotheses for human pose in subsequent frames. In [11], the authors combine dense appearance-based likelihood maps with spatial priors from a

particle filter. Other approaches have been presented by [8] where crowd density estimates serve as information prior and in [24] where a second detector is trained in an unsupervised manner on the results of a first detection step.

While these works contributed to the overall understanding of the relation of object detection and tracking, it must be said that they mostly imply a very tight connection of these two components in a framework which imposes hard constraints on both system parts. In case of a proprietary detector library, the source code might not be available and the interface would be required to return internal detector values (e.g. pixelwise scoremaps in the case of [11]). Other problems to be mentioned are the need for re-training in [24] and thus a potentially much higher run-time.

In this work, we propose two post-detection filters for generic tracking-by-detection frameworks in order to enhance the detection quality and thus improve the tracking process. The first filter uses a hysteresis thresholding concept in order to adapt to lower-scoring detections. The second filter applies optical flow and actively identifies potentially missed detections from previous frames.

Both filters work on the detection level, i.e. they do not require access to internal detector data such as score maps but can be applied regardless of the detector used. They can be seen as semi-trackers which introduce additional knowledge into the detection process and are fast to compute.

The structure of this paper is as follows: Sections 2.1 and 2.2 present two filter approaches improving the detection quality of a custom pedestrian detector. Section 3 shows experimental results of both methods and evaluates implications on the detection threshold. It also shows how the improved detection results lead to better tracking outcomes. Section 4 concludes the paper.

2. Post-Detection Filters for Person Detectors

Object detectors produce pairs of bounding boxes and confidence scores of objects of interest on a per-frame basis, thus ignoring the spatio-temporal relations between them. Applying an increasing threshold to the confidence scores leads to achieving fewer false-positives at the cost of more false negatives. Depending on pose, illumination etc., the same person produces different scores in subsequent frames. Using a low threshold, a pedestrian may be detected in all frames, but a lot of false detections in other areas of the images are produced, while using a high threshold results fewer detections and more false negatives. Configuring these thresholds can be tedious and good values may vary over different videos.

In this work, we show both a *passive* and an *active* post-detection filter in order to improve the detection quality before applying a tracking method. While the passive filter relies on lower-scoring detection candidates contained in the detector result set, the active approach seeks to identify additional detection candidates the detector has missed. Both methods have been designed to work on the detection level after a potential non-maxima suppression step as e.g. in [9], thus ensuring the highest possible level of independence and modularity, and should be applicable in a wide range of scenarios.

2.1. Passive Detection Filtering Using Temporal Hysteresis Thresholding

As mentioned above, parametrization of a pedestrian detector in order to find a suitable compromise between false positives and false negatives can be challenging. The passive filtering technique used in this work circumvents this dilemma by using both a high threshold σ_h and a low threshold σ_l in a hysteresis-like manner:

1. At time t, all high-scoring detections $\{d_{h_0}, ..., d_{h_{N-1}}\}$ in frame I^t with score $S_{h_i} > \sigma_h$ are extracted and added to the set of results D^t.

2. All low-scoring detections $\{d_{l_0}, ..., d_{l_{N-1}}\}$ are gathered by thresholding the detections of frame I^t with σ_l and compared to the high-scoring detections of the previous frame. For each high-scoring detection $d_{h_i}^{t-1} \in D^{t-1}$ from the previous frame with no matching counterpart $d_{h_j}^t \in D^t$ where $IOU(d_{h_i}^{t-1}, d_{h_j}^t) \leq \sigma_{iou}$, the highest scoring detection d_{l_i} fulfilling $IOU(d_{h_i}^{t-1}, d_{l_i}^t) \geq \sigma_{iou}$ is extracted ($\sigma_{iou} = 0.1$ in this work). In this term, $IOU(d_i, d_j)$ denotes the intersection-over-union of the bounding boxes of detections d_i, d_j. All d_{l_i} found are added to the set of results D^t.

3. Report D^t as final detections for frame I^t. It contains thus the high-scoring detections from the current frame and low-scoring detections with a significant overlap to previously received detections.

Figure 1 illustrates the principle of this scheme. It can be seen that almost all low-confidence detections are ignored while only the true positives are accepted.

In some cases, a false detection can be propagated infinitely by this scheme, e.g. when a pedestrian leaves the scene near a permanently reappearing low-confidence detection. To prevent this, an additional maximum propagation time t_{MAX} is used to limit the maximum number of subsequent low-confidence scores of $d_{l,i}$. This passive post-detection filter can be implemented very efficiently and its computational complexity is negligible compared to the detector step.

2.2. Active Detection Filtering Using Optical Flow

Different from the previously shown *passive* detection filter, it is also possible to design an *active* detection filter

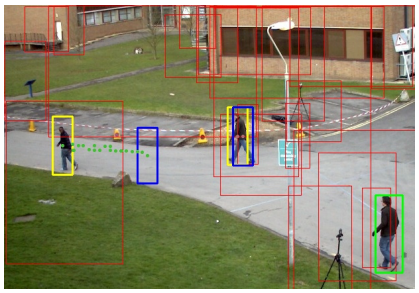

Figure 1. Example of passive hysteresis thresholding in the PETS09-S2L1 sequence. Green boxes: high confidence detections; Yellow boxes: current low-confidence detection validated from previous frames (dotted line indicates position in interim frames); Red boxes: discarded low-confidence detections.

which searches actively for candidate detections not identified by the detector, regardless of their score. We implement this active filtering approach using sparse optical flow.

Using the previous and the current image frames I^{t-1}, I^t, sparse local optical flow information $V^{t-1,t}(x,y)$ can be derived. We use a pyramidal implementation of [17] but other methods such as [21] could also be used. With the region of interest for every detection in $D^{t-1} = \{d_0, \dots d_{N-1}\}$, the propagated position $\hat{d}_i^t$ of a detection in the current frame is

$$\hat{d}_i^t = V^{t-1,t}(d_i^{t-1}) = d_i^{t-1} + \mathbf{v}_i^{t-1,t} \tag{1}$$

with $\mathbf{v}_i^{t-1,t}$ as the local displacement for d_i^{t-1}.

This gives the propagated detection set $\widehat{D}^t = \left\{\hat{d}_0^t, \dots, \hat{d}_{N-1}^t\right\}$ which contains current position estimates of all detections from the last frame.

In order to identify targets which existed in previous frames but have not been found in the current frame, a comparison between the two detection sets D^t and $\widehat{D}^t$ is performed. We use the spatial overlap of the respective regions of interest (IOU) but other options could be image information of detections (e.g. color / gradient distribution).

As a result of this step, all propagated detections are removed from $\widehat{D}^t$ if a matching candidate in D^t is found (we use the criterion $IOU > 0.5$), thus yielding the filtered set of $\widehat{D}^t_{filter}$. This step is especially important because two measurements for one object would violate fundamental assumptions for tracking systems.

Now, detections in the current frame can be "filled up" with propagated detections from $\widehat{D}^t_{filter}$ and the resulting detection set is

$$D^t_{final} = \widehat{D}^t_{filtered} \cup D^t. \tag{2}$$

In principle, this concept of propagating detections from previous images into new ones can be done for arbitrary numbers of frames t_{MAX}, e.g. for $t_{MAX} = 2$, a detection in frame t would be propagated into the frames $t+1$ and $t+2$ and so on for greater values of t_{MAX}. The double filter inhibits too many false positives but on the other hand, the gain is limited due to saturation effects.

The filter run-time depends on the number of propagations and detections in the video and the optical flow implementation. In standard scenarios, the computational load is low compared to the detector and can easily be parallelized.

3. Experimental Results

For evaluation, we use a set of very different video sequences from the well-known CAVIAR[1] (1. "EnterExitCrossingPaths1cor", 2. "WalkByShop1cor", 3. "ThreePastShop1cor", 4. "ThreePastShop2cor"), Pets2009 [10] and Parking Lot [22] datasets and the Clear metrics [14] computed by the development kit of the MOT challenge [18]. The baseline detections were obtained from DPM v5 implementation[2] using the VOC2007 model. Following [2], a minimum IOU of 0.2 instead of 0.5 is used for true positives in order to account for inaccuracies in the ground truth of some sequences.

Figure 3 shows N-MODA and N-MODP values for different detector thresholds and their filtered counterparts. In case of the passive filtering, this threshold represents σ_h for the high-confidence detections. The low-confidence threshold σ_l (passive) and maximum propagation times t_{MAX} (active / passive) are chosen for best performance.

For most applications, the detection accuracy N-MODA is very important. It is noticed that for this measure, both methods consistently outperform the baseline detection for a wide range of detector thresholds. Only for very low thresholds outside of the working range of the detector, the filtered results are slightly worse because too many false detections are propagated. The possible range of suitable detection thresholds is increased by the filters which means that the detector configuration becomes easier.

The peak accuracy of the active filtering approach is, with one exception, slightly better than the passive filtering. A reason for this is that the size of the detection bounding boxes can change considerably after a couple of propagations while the size of the active-filtered regions of interest remains the same.

For the PETS09-S2L1 sequence, the gain for active filtering is lower than for other videos which can be explained by the lamp post in the middle of the scene. Related occlusions inhibit correct optical flow estimates.

The N-MODP values, describing the spatial accuracy of the detected bounding boxes, are mostly worse than in the baseline case. This behaviour is indeed intuitive since already baseline detections have a certain position noise com-

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

[2] https://github.com/rbgirshick/voc-dpm

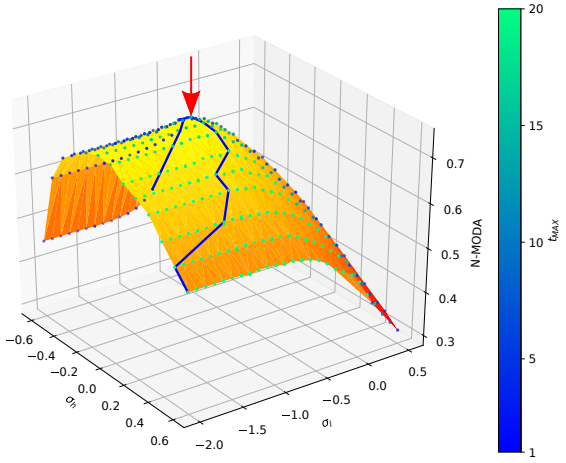

Figure 2. N-MODA values of the passive filter depending on the thresholds σ_h, σ_l and the best propagation time t_{MAX} for sequence PL1. The red arrow marks the peak, the blue line the maximum for each threshold σ_h.

pared to the ground truth. Both filters can suffer from drifting effects and thus increase this noise level.

Figure 2 visualizes the performance of the passive filtering scheme depending on its parameters for the exemplary Parking Lot 1 sequence. It can be seen that for all low σ_l thresholds and high maximum propagation times t_{MAX} roughly optimal N-MODA values are anticipated, making this method easy to parametrize. The best σ_h threshold is relatively easy to find, however if a maximal gain is desired, a little fine-tuning for σ_l and t_{MAX} is required. Even without these configurations, the overall gain for the best σ_h compared to the baseline is considerably stable.

Table 1 summarizes the N-MODA / N-MODP values for the filtering schemes with their parameters and the unfiltered baseline method. The previously mentioned minimum IOU of 0.2 for correct true positives generally leads to reduced N-MODP values. Note that for all experiments higher thresholds for the filtered results can be used in comparison to the baseline, leading to fewer false positives. The gains again show the improvements of the two filtering schemes compared to the baseline.

Table 2 shows tracking results using a Gaussian mixture probability hypothesis density filter based on [7]. This tracker has been chosen as an example for a tracking-by-detection system not using additional image information.

Generally, the improved detection quality leads to better tracking results, too. Due to the low-pass properties of the tracker, however, this effect is not perceived on every video but can be huge for some cases. The tracker parameters are the same for all comparisons and have been optimized on the baseline results for a fair comparison thus indicating that higher gains could be achieved for the filters.

4. Conclusion

In this work we presented an in-depth analysis of two post-detection filters in order to enhance the performance of a pedestrian detector and compared their performance on a range of different video sequences to a popular state-of-the-art detector. It was shown that a simple, hysteresis-based detection filter can improve the detector performance considerably but is outperformed by an active detection filter using sparse optical flow.

The improvements on the detection level have been validated and confirmed also for tracking using a tracking-by-detection algorithm on the respective detection results. Both filters are real-time capable for standard scenarios and suitable for custom pedestrian detectors.

Future work will comprise an extension of this analysis on other detectors and the adaptation of new post-detection filter strategies using visual information from the image.

5. Acknowledgements

The research leading to these results has received funding from the European Communitys FP7 under grant agreement number 607480 (LASIE).

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[2] E. Bochinski, V. Eiselein, and T. Sikora. Training a convolutional neural network for multi-class object detection using solely virtual world data. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 278–285, Colorado Springs, CO, USA, Aug. 2016.

[3] M. D. Breitenstein, F. Reichlin, B. Leibe, E. K. Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, pages 1515–1522, 2009.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 2, pages 886–893, 2005.

[5] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014.

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8):1532–1545, 2014.

[7] V. Eiselein, D. Arp, M. Pätzold, and T. Sikora. Real-time multi-human tracking using a probability hypothesis density filter and multiple detectors. In *AVSS*, pages 325–330, 2012.

[8] V. Eiselein, H. Fradi, I. Keller, T. Sikora, and J.-L. Dugelay. Enhancing human detection using crowd density measures and an adaptive correction filter. In *10th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, Kraków, Polen, Aug. 2013. IEEE, IEEE Computer Society.

Sequence		Caviar 1	Caviar 2	Caviar 3	Caviar 4	PETS09	Parking Lot 1	Parking Lot 2
baseline (DPM)	σ	-0.5	-0.5	-0.3	-0.3	-0.5	-0.4	-0.4
	N-MODA	0.52	0.33	0.33	0.3	0.72	0.67	0.73
	N-MODP	**0.14**	**0.25**	0.36	**0.34**	**0.18**	0.19	**0.08**
passive filtering	$\sigma_h/\sigma_l/t_{MAX}$	-0.4/-0.8/13	-0.4/-0.8/12	-0.1/-0.6/20	-0.2/-0.8/20	-0.1/-0.9/9	-0.1/-0.8/17	-0.1/-1.0/16
	N-MODA	0.54	0.34	0.35	0.32	0.75	0.73	0.81
	Gain	0.04	0.03	0.06	0.07	0.04	0.09	0.11
	N-MODP	0.13	0.24	**0.37**	0.33	**0.18**	0.18	**0.08**
	Gain	-0.07	-0.04	0.03	-0.06	0	-0.05	0
active filtering	σ/t_{MAX}	-0.4/13	-0.3/19	0.1/20	0/20	-0.2/3	0/18	-0.1/13
	N-MODA	**0.58**	**0.39**	**0.37**	**0.35**	**0.76**	**0.76**	**0.85**
	Gain	0.12	0.18	0.12	0.17	0.06	0.13	0.16
	N-MODP	0.12	0.23	**0.37**	0.31	**0.18**	**0.20**	**0.08**
	Gain	-0.14	-0.08	0.03	-0.09	0	0.05	0

Table 1. Detection metrics for both filters with respective parameters and unfiltered baseline. Gain denotes the respective improvements.

Sequence		Caviar 1	Caviar 2	Caviar 3	Caviar 4	PETS09	Parking Lot 1	Parking Lot 2
baseline (DPM)	N-MOTA	0.44	0.29	0.25	**0.28**	**0.67**	0.51	0.72
	N-MOTP	**0.14**	**0.25**	**0.37**	**0.37**	0.12	**0.29**	**0.06**
passive filtering	N-MOTA	0.46	0.30	0.27	**0.28**	0.62	0.59	0.71
	Gain	0.05	0.03	0.08	0	-0.07	0.16	-0.01
	N-MOTP	**0.14**	0.24	0.34	0.36	**0.13**	**0.29**	**0.06**
	Gain	0	-0.04	-0.08	-0.03	0.08	0	0
active filtering	N-MOTA	**0.53**	**0.35**	**0.30**	**0.28**	**0.67**	**0.69**	**0.77**
	Gain	0.2	0.21	0.2	0	0	0.35	0.07
	N-MOTP	0.10	0.21	0.35	0.33	0.12	0.26	**0.06**
	Gain	-0.29	-0.16	-0.05	-0.11	0	-0.1	0

Table 2. Comparison of tracking metrics for both filter methods to the unfiltered baseline. Gain denotes the respective improvements.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[10] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1–6. IEEE, 2009.

[11] A. Gepperth, E. Sattarov, B. Heisele, and S. A. R. Flores. Robust visual pedestrian detection by tight coupling to tracking. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1935–1940, Oct 2014.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[13] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015.

[14] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336, 2009.

[15] X. Lan, A. J. Ma, and P. C. Yuen. Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201, 2014.

[16] P. Liang, E. Blasch, and H. Ling. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Transactions on Image Processing*, 24(12):5630–5644, 2015.

[17] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence (IJCAI 1981)*, pages 674–679, 1981.

[18] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831.

[19] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):58–72, Jan 2014.

[20] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2014.

[21] T. Senst, V. Eiselein, and T. Sikora. Robust local optical flow for feature tracking. *Transactions on Circuits and Systems for Video Technology*, 09(99), 2012.

[22] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1815–1821. IEEE, 2012.

[23] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1442–1468, 2014.

[24] X. Wang, G. Hua, and T. X. Han. Detection by detections: Non-parametric detector adaptation for a video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 350–357, June 2012.

[25] B.-F. Wu, C.-C. Kao, C.-L. Jen, Y.-F. Li, Y.-H. Chen, and J.-H. Juang. A relative-discriminative-histogram-of-oriented-gradients-based particle filter approach to vehicle occlusion handling and tracking. *IEEE Transactions on Industrial Electronics*, 61(8):4228–4237, 2014.

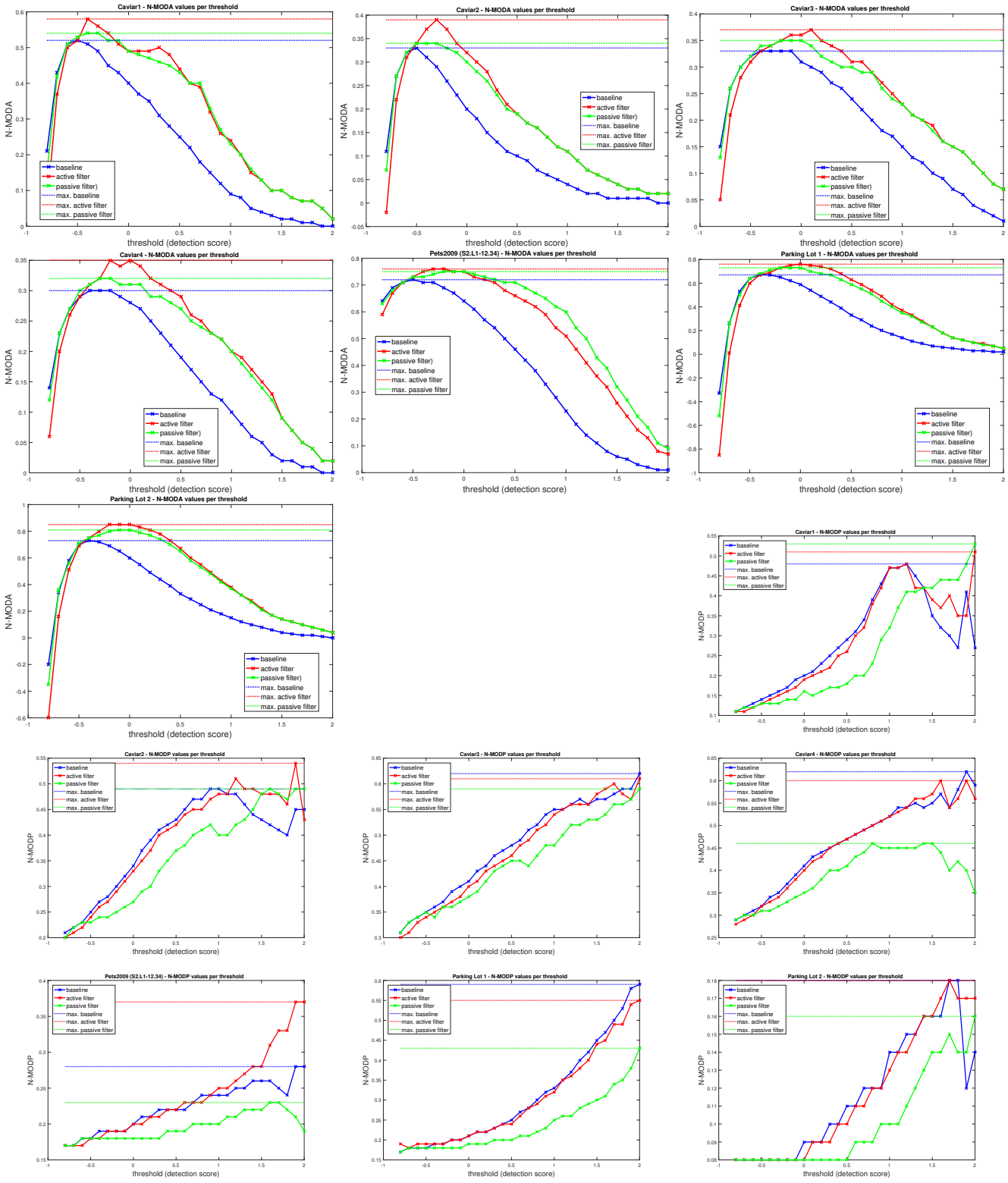

Figure 3. Best N-MODA and their respective N-MODP values for different detector thresholds for all test sequences.