Improving Neural Non-Maximum Suppression for Object Detection by Exploiting Interest-Point Detectors

Charalampos Symeonidis, Ioannis Mademlis, Nikos Nikolaidis, Ioannis Pitas

AIIA Lab, Department of Informatics

Aristotle University of Thessaloniki Greece

{charsyme, imademlis, nnik, pitas}@csd.auth.gr

INTRODUCTION

- Non-maximum suppression (NMS) is a final refinement step in almost every visual object detector.
- The problem that NMS attempts to solve arises from the tendency of many detectors to output multiple, neighbouring candidate object ROIs for a single given visible object.
- The default algorithm (Greedy NMS) is fairly simple and suffers from drawbacks, due to its need for manual tuning.
- The fixed IoU threshold of GreedyNMS can lead to failure in certain cases, e.g. wide suppression may remove detections that cover objects with lower scores, while too low a threshold is unable to suppress duplicate detections.
- NMS has been improved using deep neural networks that learn how to solve a spatial overlap-based detections rescoring task in a supervised manner. The aim is to decrease the score of those that cover an object which has already been detected.
- In this paper, a method is presented that improves neural NMS performance by augmenting the representation of each input detection. This is performed by extracting interest-points within each detection ROI and using the Frame Moments Descriptor (FMoD) for exploiting the statistical dispersion of their spatial distribution to create an appearance-based candidate ROI representation.

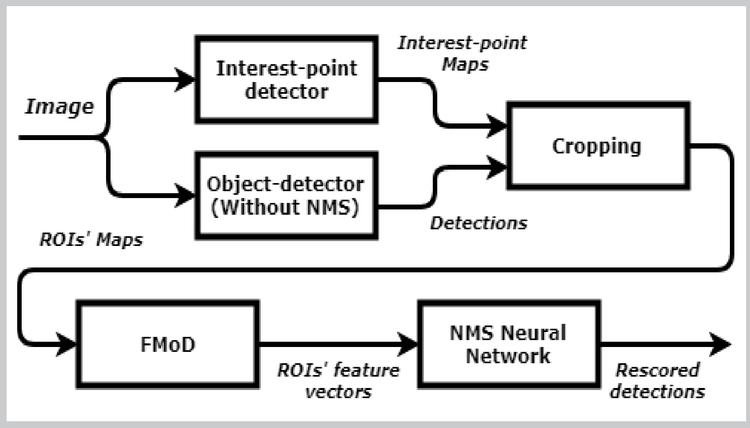
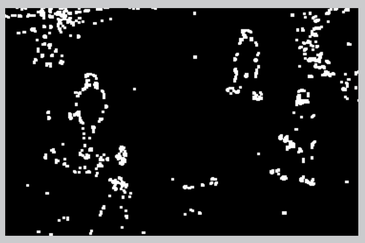


Figure: Pipeline of the proposed method.

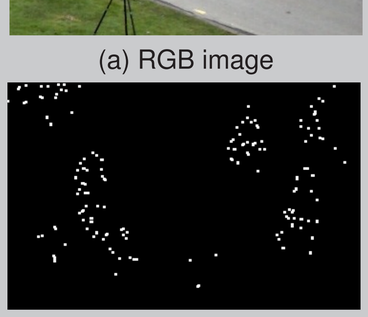
INTEREST-POINT DETECTION

- Interest-point detectors, such as SIFT, FAST and AKAZE, can detect locations on an RGB image that possess useful properties.
- In interest-point maps, those pixel that correspond to the location of a detected interest-point contain an integer luminance value, correlated to its magnitude.
- Alternatively, edge maps were also generated using the Scharr operator.

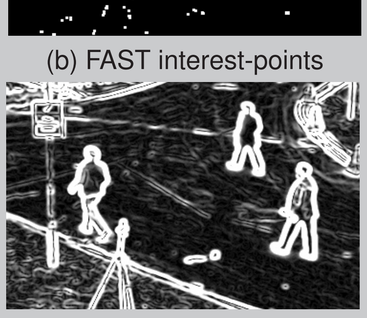
(a) RGB image



(b) FAST interest-points



(c) AKAZE interest-points



(d) Edge map

Figure: Interest-points extracted from FAST and AKAZE detectors, along with the corresponding edge map.

ROI REPRESENTATION

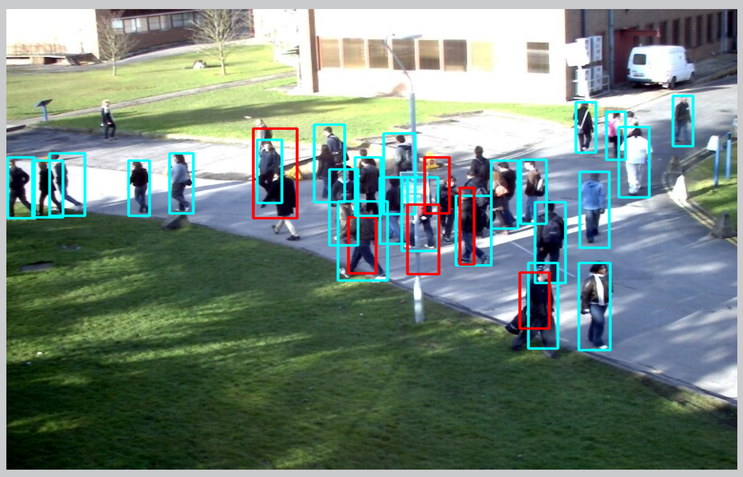
- The overall spatial distribution of interest-points within the candidate ROIs seems to align with the silhouettes of the detected objects.
- This distribution can be exploited as a candidate ROI appearance-based discriminant factor for identifying complete vs partial object silhouettes.

ROI REPRESENTATION (CONTINUE)

- The Frame Moments Descriptor (FMoD) [1] has been adopted for capturing the spatial distribution of the interest-points within the ROI interest-point map.
- The representation vector of a candidate ROI contains statistical attributes, such as horizontal/vertical mean/skew/st.deviation, etc. Those attributes are computed under a spatial pyramid partitioning scheme.

EMPIRICAL EVALUATION

- The evaluation was performed on three object detection datasets, i.e. COCO, PETS and Okutama-Action using only the class “person”. Faster-RCNN, [2], YOLOV3 detectors were selected.
- All datasets contain images with crowded areas, where many persons occlude each other, making GreedyNMS highly unsuitable.
- State-of-the-art GossipNet [3] was selected as the testbed for implementing and evaluating the proposed method.
- GossipNet is a neural network capable of jointly processing input detections and rescoring them.
- In GossipNet, the appearance of candidate detection is ignored and only their spatial interrelationships are exploited.
- The proposed approach was evaluated by feeding into GossipNet the appearance-based FMoD description vector of a detection.
- Comparisons were made against GreedyNMS, OpenCVNMS, and the default GossipNet.



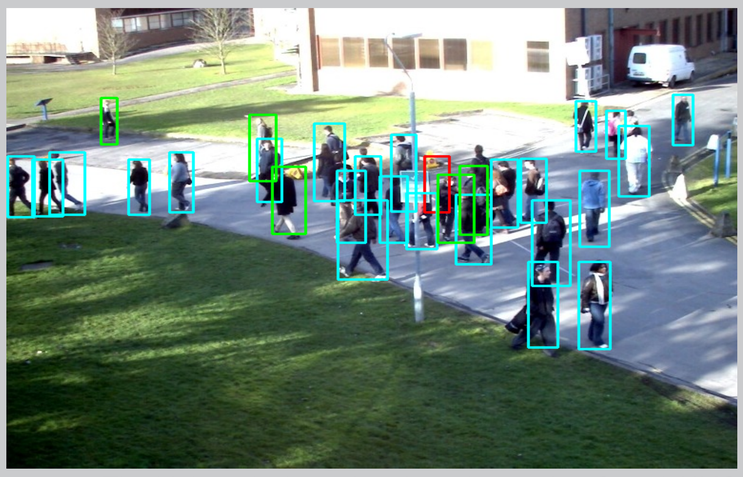


Figure: The highest (30) scoring detections using GreedyNMS (top) and the proposed method (bottom). Blue: True Positive(TP) detections assigned to the same objects between the two methods. Red: False Positive(FP) detections. Green: TP detections assigned to objects detected using only the corresponding method.

EMPIRICAL EVALUATION: COCO

- FMoD statistics from the candidate ROIs’ AKAZE interest-point maps and from their edge maps increase GossipNet’s Average Precision (AP) by a small amount.

Method	Minival AP	Minitest AP
Greedy NMS IoU>0.5	65.6%	65.0%
OpenCV NMS IoU>0.5	65.6%	65.1%
Default GossipNet_128	67.3%	66.8%
AKAZE_FMoD_128	67.6%	67.0%
FAST_FMoD_128	67.5%	66.8%
EdgeMap_FMoD_128	**67.8%**	**67.2%**
Improvement	***+0.5%***	***+0.4%***

EMPIRICAL EVALUATION: PETS

- FMoD statistics both from the interest-point maps and from the edge maps increase the AP of the default GossipNet.
- Description vectors created using FAST and SIFT interest-point maps achieve the best AP in all conducted experiments.

Method	AP
Greedy NMS IoU $>$ 0.4	76.4%
Greedy NMS IoU $>$ 0.5	73.0%
OpenCV NMS IoU $>$ 0.4	76.3%
OpenCV NMS IoU $>$ 0.5	72.4%
Default GossipNet_90	83.4%
Default GossipNet_128	84.3%
FAST_FMoD_90	**86.4%**
AKAZE_FMoD_90	84.8%
EdgeMap_FMoD_90	85.5%
Improvement	***+2.1%***

EMPIRICAL EVALUATION: OKUTAMA-ACTION

- Though Okutama-Action may not suffer from cluttered ground-truth detections, the best proposed method variant surpasses both GreedyNMS and the default GossipNet by 2% in AP.
- The appearance information extracted from each ROI helps to reduce the scores of False Positive detections that are not perceived as “double” detections of an already detected object and, thus, are unaffected by default GossipNet.

Method	AP
Greedy NMS IoU $>$ 0.4	70.7%
Greedy NMS IoU $>$ 0.5	71.4%
Default GossipNet_128	71.9%
FAST_FMoD_128	**73.9%**
EdgeMap_FMoD_128	73.8%
Improvement	***+2.0%***

CONCLUSIONS

- In this work, neural NMS performance was augmented by feeding the network additional information extracted from within each candidate ROI.
- The deviation in 2D distribution between the interest-points or edges detected inside a ROI enclosing the actual object entirely, and one that only captures it partially, is exploited as a discriminant factor.
- The empirical evaluation on three public person detection datasets leads to state-of-the-art results, at a small computational overhead.
- Overall, the proposed method significantly enhances the operation of neural NMS by forcing it to implicitly learn how to solve a appearance-based binary classification problem(complete vs partial object silhouette), on top of the typical ROI overlap-based detections rescoring.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union’s European Union Horizon 2020 research and innovation programme under grant agreement No 731667 (MULTIDRONE).

REFERENCES

[1] I. Mademlis, N. Nikolaidis, and I. Pitas. Stereoscopic video description for key-frame extraction in movie summarization. In *Proceedings of the EURASIP European Signal Processing Conference (EUSIPCO)*. IEEE, 2015.

[2] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2009.

[3] J. H. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.