IMPROVING NEURAL NON-MAXIMUM SUPPRESSION FOR OBJECT DETECTION BY EXPLOITING INTEREST-POINT DETECTORS

Charalampos Symeonidis, Ioannis Mademlis, Nikos Nikolaidis and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
Email: {charsyme, imademlis, nnik, pitas}@csd.auth.gr

ABSTRACT

Non-maximum suppression (NMS) is a post-processing step in almost every visual object detector. Its goal is to drastically prune the number of overlapping detected candidate regions-of-interest (ROIs) and replace them with a single, more spatially accurate detection. The default algorithm (Greedy NMS) is fairly simple and suffers from drawbacks, due to its need for manual tuning. Recently, NMS has been improved using deep neural networks that learn how to solve a spatial overlap-based detections rescoring task in a supervised manner, where only ROI coordinates are exploited as input. In this paper, neural NMS performance is augmented by feeding the network additional information extracted from the appearance of each candidate ROI. This information captures statistical properties regarding the spatial distribution of interest-points detected within the corresponding image region. Thus, the deviation in 2D distribution between the interest-points detected inside a ROI that encloses the actual object entirely, and within one that only captures it partially, is exploited as a discriminant factor, with the NMS network being implicitly forced to also learn how to solve an additional, appearance-based binary classification task (complete vs partial object silhouettes). The empirical evaluation on three public person detection datasets leads to state-of-the-art results, at a small computational overhead.

Index Terms— object detection, non-maximum suppression, interest point detection, appearance-based classification

1. INTRODUCTION

Object detection is a long-standing, fundamental problem in computer vision. It consists in generating bounding boxes (in 2D pixel coordinates) for objects detected on-image that belong to pre-specified object classes, as well as in assigning classification scores to them. State-of-the-art object detectors are two-stage algorithms, initially creating object proposals for input images using a method such as selective search or a deep neural network, then resampling pixels, or extracting features from these proposals using Convolutional Neural Networks (CNNs), and finally using a classifier to determine the existence and the class of any object in each proposal. Faster R-CNN [1], a widely-used end-to-end neural object detector, is the top-performing algorithm in this category.

The second type of mainstream neural object detectors are end-to-end systems, i.e., networks that directly process raw images and output bounding boxes/regions-of-interest (ROIs), without an intermediate object proposition step, having the task to both classify and localize objects. Thus, each candidate detected object ROI is composed of a class, a set of spatial pixel coordinates and a confidence score. The end-to-end nature of similar single-stage object detectors significantly lowers their runtime requirements. YOLOv3 [2] is one the most promising algorithms in this category.

Almost all object detectors, either two-stage or single-stage, incorporate a final refinement step, i.e., *Non-Maximum Suppression* (NMS), where spatially overlapping detected ROIs are merged / filtered. The problem it attempts to solve arises from the tendency of many detectors to output multiple, neighbouring candidate object ROIs for a single given visible object, due to their implicit sliding-window nature. Below, the term "detection" is employed for these candidate detected object ROIs, prior to applying NMS, instead of the final post-processed/refined results, as it is commonly used in relevant literature.

The de facto standard in NMS for object detection is GreedyNMS [3]. It selects high-scoring detections and deletes less confident neighbours, since they most likely cover the same object. Its simplicity, speed and unexpectedly good behaviour in most cases, make it competitive against proposed alternatives, since rapid execution is of the utmost importance in NMS. An Intersection-over-Union (IOU) threshold determines which less-confident neighbors are suppressed by a detection. Most NMS algorithms, including GreedyNMS, do not make any extra effort to jointly process the ROIs and assign one detection per object. In additon, this fixed IOU threshold leads GreedyNMS to failure in certain cases. For instance, wide suppression may remove detections that cover objects with lower scores, while too low a threshold is unable

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).

to suppress duplicate detections.

In recent years, a number of alternatives or refinements to GreedyNMS have emerged. The most advanced NMS algorithms are neural networks that either refine the output of simpler NMS methods, or directly process the detector's results, completely replacing GreedyNMS in the latter case. For instance, GossipNet [4] is a neural network capable of jointly processing input detections, i.e., all the output candidate ROIs of an object detector for a given image, and rescoring them, in order to handle cases where the standard Non-Maximum Suppression algorithm fails. Typically, the appearance of candidate ROIs is ignored and only their spatial interrelationships are exploited.

In this paper, a method is presented that improves neural NMS performance by augmenting the representation of each input detection, so as to help the network in its rescoring task. This is performed by extracting interest-points within each detection ROI and exploiting the statistical dispersion of their spatial distribution to create an appearance-based candidate ROI representation. This representation is fused with the one constructed automatically by the network, thus improving the system's overall precision.

Interest-points, such as SIFT [5], FAST [6] or AKAZE [7], can be located very fast in an image nowadays. In the NMS case, their representational power stems from the fact that they lie mainly along an object's silhouette, since most raw detections outputted by a detector already cover part of an object. Normally, the scene background may contain a high number of interest-points, e.g., due to texture or illumination variability, therefore they do not convey semantic information on their own. However, such background image regions are not typically included in the detections that NMS processes. Thus, in this case, their spatial distribution tends to capture the shape of the object part lying within the ROI.

Our hypothesis is that proper representations of spatial interest-point maps computed on the candidate detection ROIs (i.e., the raw object detector output) may be fed to a neural NMS network in order to easily enhance its performance. In this paper, we implement and empirically evaluate an algorithm designed to test the above hypothesis, relying on off-the-shelf, fast, hand-crafted interest-point detectors/image descriptors and a state-of-the-art neural NMS network. To the best of our knowledge, the concept of interest-point maps has not been previously exploited for augmenting NMS performance, or for assisting neural object detection in general. The results empirically validate our hypothesis by showcasing a significant boost in average precision on three public person detection datasets, in comparison to the state-of-the-art baseline neural NMS network that we modified, thus opening up promising avenues for further research.

2. RELATED WORK

NMS, an essential part of computer vision for decades, is widely used in object detection [1]. Several algorithms have been proposed over time, based either on modified versions of GreedyNMS, or on entirely new approaches. In [8], the authors demonstrate that a GreedyNMS algorithm for person detection improves performance in face detection. Their method selects a bounding box with the maximum detection score and its neighboring boxes are suppressed using a predefined overlap threshold. The authors in [9] proposed a clustering approach that provides globally optimal solutions, in a relaxed problem formulation. However, the results do not indicate significant improvements over GreedyNMS.

The recently presented IoU-Net [10] is a supervised neural network that learns a suitable IoU threshold from the training data; this is then used in the typical GreedyNMS algorithm. Relation Network [11] has also been proposed, which processes a set of objects simultaneously, allowing to model relations between their appearance and their geometry.

Few works have explored true end-to-end learning that considers NMS. In [12, 13], NMS is included in training, thus the classifier is made aware of the NMS process which is employed during testing. Although conceptually correct, this does not make NMS itself learnable. SoftNMS [14] decays the detection scores of all other neighbors as a continuous function of their overlap with the higher-scored ROI, instead of eliminating all lower-scored surrounding ROIs.

In [4], the authors propose GossipNet, a deep feed-forward neural network that performs state-of-the-art NMS using only detection coordinates and their scores as input (candidate ROI appearance is not considered). Its architecture is based on a repeating set of fully-connected layers (such a set is called a "block"). Each successive block refines the encoded representation of all detections, by taking into account their respective spatial neighbors. The network's task is to jointly process all input image detections, so as not to directly prune them, but to rescore them. The aim is to decrease the score of those that cover an object which has already been detected. After rescoring, simple thresholding on the modified score is sufficient to significantly reduce the set of detections. During inference, the network input is a zero vector per ROI (as a trivial ROI representation) and a small set of pairwise features relating each candidate detection with its spatial neighbors, computed using properties such as the location of the two ROIs, their IoU, etc.

3. NEURAL NON-MAXIMUM SUPPRESSION EXPLOITING INTEREST-POINT DETECTORS

In this paper, non-trivial initial input is provided to a neural NMS network for each detection, without inducing significant computational overhead. The goal is to augment the internal representation of each candidate ROI using its appearance and, thus, increase network performance while retaining computational efficiency.

Interest-point detectors, e.g., a corner detector and/or a scale-space extrema detector, such as SIFT, FAST and AKAZE, can detect locations on an RGB image (in pixel co-

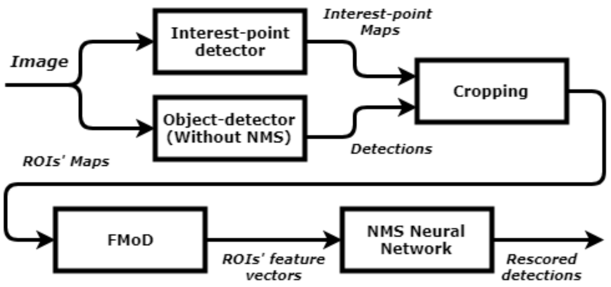


Fig. 1. Pipeline of the proposed method.

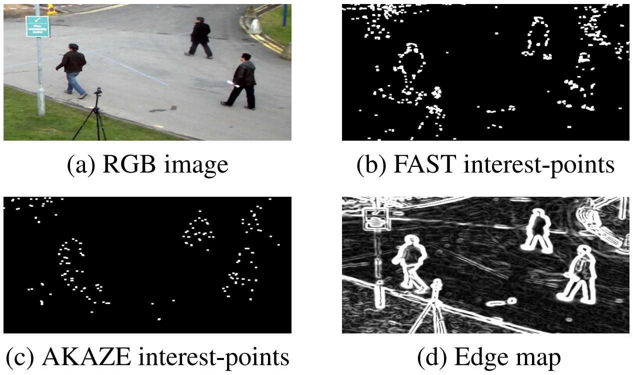
ordinates) that possess useful properties (such as invariance to several transformations) and may be employed for several different tasks. In many cases, a magnitude and/or an orientation are also computed for each interest-point, along with its location. Maps depicting detected interest-points in an example image are shown in Figure 2. Such interest-point maps can easily be obtained by constructing an initially blank, one-channel image, having dimensions in pixels identical to those of the original RGB image. Then, each pixel corresponding to the location of a detected interest-point can be set to an integer luminance value (in the interval [0, 255]), correlated to the latter's magnitude.

Extracting image interest-points with modern computers is rapid and efficient. This is a vital quality for NMS algorithms, since they constitute only a post-processing step in the overall object detection system and are expected to execute quickly. More significantly, when restricting interest-point detection in the candidate object ROIs typically fed to NMS algorithms, their overall spatial distribution within these ROIs seems to align with the silhouettes of the detected objects, as shown in Figure 2. Therefore, this distribution can be exploited as a candidate ROI appearance-based discriminant factor for identifying complete vs partial object silhouettes.

The proposed method consists in compactly capturing this distribution, using a hand-crafted, rapidly computable image descriptor, and employing it as an initial detection representation fed to a neural NMS network. This takes advantage of the fact that NMS inputs are image regions known to partially enclose visible object silhouettes, instead of, e.g., depicting the background. As far as we can tell, this serendipitous fact has not been previously noticed or exploited for augmenting NMS performance, or assisting neural object detection in general. Note that, alternatively, edge-maps of the candidate object ROIs (also rapidly computable) may be used instead of interest-point maps; our observations hold in this scenario too.

The Frame Moments Descriptor (FMoD) has been adopted for achieving this task. FMoD was originally devised in a global [15] and in a local [16] variant (LMoD), respectively applied to movie [17] and activity video [18, 19, 20] summarization via key-frame extraction. Typically, FMoD and LMoD capture informative image statistics from various available image channels (e.g., luminance, color/hue, optical flow magnitude, edge map, and/or stereoscopic disparity), both in a global and in various local scales, under a spatial pyramid video frame partitioning scheme.

In this paper, for the special use-case of describing a ROI interest-point map instead of a typical image/video frame, only the luminance channel is employed. The intent is to compactly capture the spatial distribution of the interest-points within the ROI interest-point map in a single numerical description vector.


(a) RGB image

(b) FAST interest-points

(c) AKAZE interest-points

(d) Edge map

Fig. 2. Interest-points extracted from FAST and AKAZE detectors, along with the corresponding edge map. The RGB image is a cropped sample from the PETS dataset.

In the simplest case (spatial pyramid depth equal to 1), the final 18-dimensional description vector for each detection contains statistical attributes, such as horizontal/vertical/vectorized block mean/st. deviation/skew, etc. However, FMoD may be separately computed for different input image regions, under a spatial pyramid partitioning scheme. Thus, pyramid depth equal to 2 may also be attempted, resulting in an aggregate 90-dimensional description vector.

4. EMPIRICAL EVALUATION

State-of-the-art GossipNet [4] was selected as the testbed for implementing and evaluating the proposed method, in the context of the highly industry-relevant person detection task. GossipNet is normally fed a zero vector as a trivial, initial representation of a detection. Thus, the actual representation of a candidate object ROI is built in stages, as information flows across the network layers, by exploiting mainly the pairwise features that capture the spatial interrelations between the candidate object ROI and its neighbors. The proposed approach was evaluated by feeding into GossipNet the ROI appearance-based description vector outlined in Section 3, instead of a zero vector.

The evaluation was performed on three object detection datasets, i.e., PETS [21], COCO [22] and Okutama-Action

[23], using only the class "person". All datasets contain images with crowded areas, where many visible persons occlude each other, making GreedyNMS highly unsuitable. The candidate detection ROIs for each dataset (before NMS post-processing) were extracted using the object detector framework reported in the corresponding baseline/competing paper. An exception is Okutama-Action, where the candidate detection ROIs were extracted using YOLOv3, since this dataset has not been previously used in any relevant work about NMS. The lower runtime requirements of YOLOv3, compared to other state-of-the-art detectors, was the principal factor behind our choice. To evaluate the proposed method, various different interest-point maps were extracted for each candidate detection ROI (using FAST, SIFT and AKAZE interest-point detectors) and, subsequently, each one was separately described with FMoD. Experiments were conducted by either taking the magnitude of interest-points into account, or by using only their location and, thus, generating binary maps.

Alternatively, edge maps were generated (using the Scharr operator) and fed as input to the FMoD description algorithm, instead of interest-point maps. In both cases, the proposed method takes advantage of the fact that NMS inputs are image regions known to partially enclose visible object silhouettes, instead of, e.g., depicting the background.

Finally, empirical evaluation was performed and comparisons were made against GreedyNMS, OpenCV[1] NMS, and the default GossipNet. All experiments were conducted using an NVIDIA GTX 1080Ti GPU and an INTEL 6900K CPU. The reported results are measured in average precision (AP) at 0.5 IoU. The GossipNet's architecture and training parameters were set as the authors suggested in [4] and no further study was made towards that direction. However, experiments were conducted using different variations of the detections' description vector size.

4.1. PETS

Though PETS is a relatively small dataset, it contains images with diverse levels of occlusion. The training set, the test set and the detections that were used are the same with those in [25]. In addition, the detections are reduced as proposed in [4] by a GreedyNMS of 0.8 IoU, due to their large number and the inability to be handled simultaneously by one GPU. GossipNet contained 8 blocks and was trained for $3 \cdot 10^4$ iterations, setting 10^{-3} as learning rate and it was decreased by 0.1 every 10^4 iterations. This combination of parameters were also suggested by [4]. All models using our proposed method, have a 90-dimensional detection representation. In Table 1, different variations of the proposed algorithm are compared against GreedyNMs, OpenCV NMS and the default GossipNet. FMoD statistics both from the interest-point maps and from the edge maps seem to 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_128	84.3%
Default GossipNet_90	83.4%
FAST_FMoD_90	**86.4%**
SIFT_FMoD_90	85.7%
AKAZE_FMoD_90	84.8%
EdgeMap_FMoD_90	85.5%
Improvement	***+2.1%***

Table 1. Comparison between different variations of the proposed method against competing ones, in the PETS test set. The last line depicts the improvement (in AP) achieved by the best proposed method variant against the best competing one.

4.2. COCO PERSON

COCO is a large dataset consisting 82,783 images for training and 40,504 images for validation/testing. Although it contains 80 labeled classes, only the "person" class was used for evaluating the proposed method. The same candidate detections, which were extracted using Faster R-CNN, and the same subsets of the validation set as in [4] were employed. The first subset, referred to as "minival", contains 5K images while the second subset, referred to as "minitest", contains 35K images.

GossipNet consisted of 8 blocks and it was trained for $2 \cdot 10^6$ iterations. The representation vectors consisted of 128 features in all variations. Each detection's representation vector in GossipNet was initialized with the representation vectors extracted from FMoD. The FMoD representation vectors consisted of 90 features, so each vector was padded with 38 zeros. The learning rate was set to 10^{-4} for the first 10^6 iterations, which later decreased to 10^{-5}.

As Table 2 shows, FMoD statistics from the candidate ROIs' AKAZE interest-point maps and from their edge maps increase GossipNet's AP by a small amount, both on minival set and minitest set.

4.3. Okutama-Action

Okutama-Action is an aerial UAV video dataset for aerial-view, concurrent human action detection, consisting of 43 minute-long fully-annotated sequences with 12 action classes. For our task, we only used ROIs with the class label "human". We employed YOLOv3 as our main object detector. The model was pre-trained on the COCO dataset and fine-tuned for $5 \cdot 10^4$ iterations using images of 832x832 resolution of Okutama-Action training set as input. The initial learning rate was set to 10^{-3} and it decreased by 0.1 at $5 \cdot 10^3$, $20 \cdot 10^3$ and $40 \cdot 10^3$ iterations.

[1] An implementation based on [24]

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%***

Table 2. Comparison between different variations of the proposed method against competing ones, in the COCO minival and minitest sets. The last line depicts the improvement (in AP) achieved by the best proposed method variant against the best competing one.

GossipNet consisted of 8 blocks and it was trained for $3 \cdot 10^4$ iterations. The representation vectors consisted of 128 features in all variations. Each detections representation vector in GossipNet was initialized with the representation vectors extracted from FMoD. The FMoD representation vectors consisted of 90 features, so each vector was padded with 38 zeros. The learning rate was set to 10^{-3} and it was decreased by 0.1 every 10^4 iterations.

Table 3 indicates that, although Okutama-Action may not suffer from cluttered ground-truth detections (cluttered or overlapping objects increase the significance of NMS post-processing for achieving good results), the best proposed method variant, i.e., the one exploiting statistical properties of FAST interest-point maps, surpasses both GreedyNMS and the default GossipNet by 2% in AP. Moreover, the appearance information extracted from each ROI also 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%***

Table 3. Comparison between different variations of the proposed method against competing ones, in the Okutama-Action test set. The last line depicts the improvement (in AP) achieved by the best proposed method variant against the best competing one.

4.4. Discussion

Overall, the proposed method significantly enhances the operation of neural NMS by forcing it to implicitly learn how to solve an appearance-based binary classification problem (complete vs partial object silhouette), on top of the typical ROI overlap-based detections rescoring. The exception is the COCO Person dataset where the benefit is small (about 0.5% in AP compared to default GossipNet), mainly because there is too great an intra-class variance with regard to the appearance of person silhouettes, due to very high variability in view angles and/or camera-to-subject distance during data capture. However, due to similar reasons, default GossipNet itself also struggles to non-negligibly improve detection performance, in comparison to the gains it induces against Greedy NMS in other datasets.

5. CONCLUSIONS

NMS is the last step in a typical object detection system. In this paper, neural NMS performance was augmented by feeding the network additional information extracted from within each candidate ROI. This information captures statistical properties regarding the spatial distribution of interest-points detected within the corresponding ROI. Thus, the deviation in 2D distribution between the interest-points detected inside a ROI enclosing the actual object entirely, and one that only captures it partially, is exploited as a discriminant factor, by taking advantage of the fact that NMS inputs are image regions known to already partially enclose visible object silhouettes. Alternatively, edge maps of the candidate ROIs are employed and described in a similar fashion. The empirical evaluation on three public person detection datasets leads to state-of-the-art results, at a small computational overhead. Future work will involve more extensive evaluation (e.g., in multiclass problems), acceleration of the candidate ROI description process, further enhancement of the ROI representations fed to the neural NMS network, as well as combining them with discriminant, CNN-derived representations that are already computed by the detector.

6. REFERENCES

[1] S. Ren, He. K., Girshick. R., and J. Sun, "Faster R-CNN: Towards real-time object detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] 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*, vol. 32, pp. 1627–1645, 2009.

[4] 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.

[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1999.

[6] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105–119, 2010.

[7] P. F. Alcantarilla, J. Nuevo, and A Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2011.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[9] R. Rothe, M. Guillaumin, and L. Van Gool, "Non-maximum suppression for object detection by passing messages between windows," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2014.

[10] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2018.

[11] H. Hu, J. Gu, Z. Zhang, Dai. J., and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] L. Wan, D. Eigen, and R. Fergus, "End-to-end integration of a convolutional network, deformable parts model and non-maximum suppression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] P. Henderson and V. Ferrari, "End-to-end training of object class detectors for mean average precision," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2016.

[14] N. Bodla, B. Singh, R. Chellappa, and L.S. Davis, "Soft-NMS improving object detection with one line of code," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[15] 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)*. 2015, IEEE.

[16] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Compact video description and representation for automated summarization of human activities," in *Proceedings of the INNS Conference on Big Data*. Springer, 2016.

[17] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multimodal stereoscopic movie summarization conforming to narrative characteristics," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5828–5840, 2016.

[18] I. Mademlis, A. Tefas, and I. Pitas, "A salient dictionary learning framework for activity video summarization via key-frame extraction," *Information Sciences*, vol. 432, pp. 319 – 331, 2018.

[19] I. Mademlis, A. Tefas, and I. Pitas, "Regularized SVD-based video frame saliency for unsupervised activity video summarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[20] I. Mademlis, A. Tefas, and I. Pitas, "Greedy salient dictionary learning with optimal point reconstruction for activity video summarization," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018.

[21] J. M. Ferryman and A. Ellis, "PETS2010: Dataset and challenge," in *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2010.

[22] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2014, Springer.

[23] M. Barekatain, M. Marti, H. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[24] L. C. Zitnick and P. Dollár, "Edge Boxes: Locating object proposals from edges," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[25] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele, "Learning people detectors for tracking in crowded scenes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.