ENSEMBLING OBJECT DETECTORS FOR IMAGE AND VIDEO DATA ANALYSIS

Kateryna Chumachenko[†], *Jenni Raitoharju*[†§], *Alexandros Iosifidis*[⋆], *Moncef Gabbouj*[†]

[†]Tampere University, Faculty of Information Technology and Communication Sciences, Finland
[§] Finnish Environment Institute, Programme for Environmental Information, Finland
[⋆]Aarhus University, Department of Electrical and Computer Engineering, Denmark

ABSTRACT

In this paper, we propose a method for ensembling the outputs of multiple object detectors for improving detection performance and precision of bounding boxes on image data. We further extend it to video data by proposing a two-stage tracking-based scheme for detection refinement. The proposed method can be used as a standalone approach for improving object detection performance, or as a part of a framework for faster bounding box annotation in unseen datasets, assuming that the objects of interest are those present in some common public datasets.

Index Terms— object detection, bounding box annotation, ensemble models

1. INTRODUCTION

Driven by the broad availability of an extensive amount of datasets in different domains, object detection has become one of the most widely used tools within the field of computer vision in recent years, finding applications in various areas, such as video surveillance [1], medical diagnostics [2], historical image analysis [3], and industrial applications [4]. Despite the huge progress made in the field of object detection in the recent years, not much attention has been paid to the generalization ability of the object detection methods: the developed algorithms assume that the training and test data come from the same distribution, and the test performance is reported on data from the same dataset used for training. For this reason, it is always preferable to train the methods on data collected directly from the domain of application, even if the classes of interest are present in some public datasets. Nevertheless, the models trained on public datasets are still often used in industrial applications. This is primarily due to the fact that obtaining training data from the domain of interest is generally both expensive and time-consuming, as it requires a significant amount of manual labour related to the annotation of the data specific to the application.

Several methods have been proposed for speeding up the bounding box annotation process on images [5, 6, 7]. Most of them still require large amounts of manual annotations, corrections, and retraining of the models. However, it is often the case that the classes of interest belong to the set of common classes present in public datasets, such as people, vehicles, and animals. Using widely-available object detection methods pre-trained on such public datasets has the potential to significantly reduce the amount of manual labour required for the annotation process. In this work, we aim to take a step towards utilizing such methods in a way that improves the object detection performance and reduces the annotation time.

This work was supported by Business Finland under the project 5G Vertical Integrated Industry for Massive Automation (5G-VIIMA).

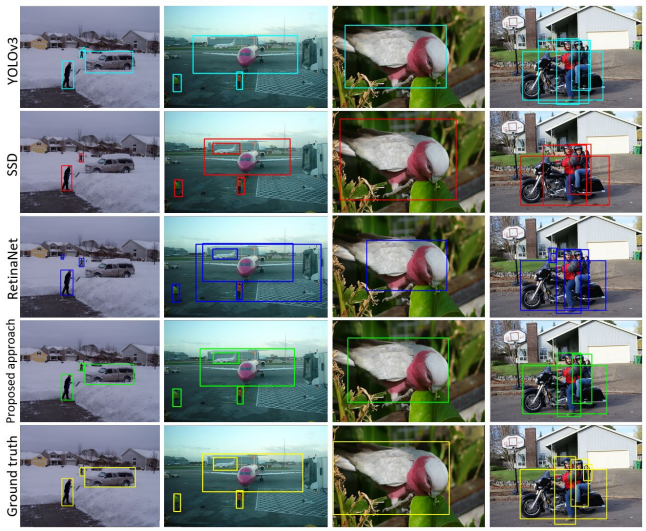

Fig. 1. Examples of the detection results of base detectors, proposed method, and ground truth bounding boxes on PASCAL VOC dataset.

Despite the rapid development of object detection methods [8, 9, 10, 11], only a very limited number of works have focused on creating ensembles of those for improved detection performance [12, 13, 14]. Moreover, to the best of our knowledge, none of the existing approaches aims at improving the precision of resulting bounding boxes, although precision can play a crucial role in a variety of applications, primarily including tracking and re-identification problems.

In this paper, we aim to take a step towards improving object detection performance and precision in image and video data by proposing a method for ensembling multiple object detection methods. In addition, we propose a weighting scheme with regression-based weights learnt from a small number of images, as well as an extension for video data utilizing temporal information.

2. RELATED WORK

Significant progress has been made in the field of object detection in recent years [8, 9, 10, 11], dominated by deep learning-based methods. To improve the object detection performance, several approaches for combining information from multiple object detectors have been proposed recently. In [13], a cascade of two face detectors was proposed to reduce the number of false-positive detections. Several metrics for evaluating the diversity and correlation of detectors were used to select the best pair of detectors in [14]. In [12], an

SVM classifier was used to select the suitable detection method for speed considerations. A learning to rank based approach, intended to rank the more relevant detections higher, was proposed in [15]. Notably, in all the above-mentioned approaches, the combination is aimed at the selection of the best detection. In our method, we first use contextual information of detectors to select the relevant detections and subsequently improve the precision of the final bounding box by taking a weighted average of the coordinates with the weights learned from a small subset of images of the target dataset.

The need for data annotation in the application domain motivates the emergence of various methods for speeding up the bounding box annotation process. A common approach to annotating large-scale datasets is using crowd-sourced annotations [5]. Since this approach exposes the data to public, it cannot be used for applications where the data needs to be kept private. Another approach relies on self-training [6, 7]. There, the general methodology is to first annotate a set of images manually and use them for the training of an object detector. The trained detector is subsequently used for producing the bounding boxes for the rest of the dataset. The obtained detections are manually refined, and the process continues until perfect annotation is achieved. Although minimizing to some extent the needed workload, the self-training approaches still require a high number of images to be annotated. In addition, a certain amount of time is needed for training the object detector.

3. PROPOSED METHODS

The main motivation behind the proposed methods lies in the assumption that distinct object detectors achieve different levels of performance on different data, and extraction of useful information from each method has potential for improving the detection accuracy and precision of bounding boxes. The proposed approach can be used as a standalone methodology for improving object detection performance or as a part of a framework for creating bounding box annotations. We rely on three state-of-the-art object detectors: SSD [9], YOLOv3 [10], and RetinaNet [11]. The choice of these detectors is based on their fast execution and good object detection performance reported in the literature. Here, one should note that the method can be equally well applied to any set of object detectors.

3.1. Fusion by non-maximum suppression

First, we consider the straightforward way of fusion by non-maximum suppression that will serve as a baseline for our work. Here, we treat all obtained detections as coming from one object detector and apply non-maximum suppression to suppress non-confident duplicates of detections identified to be the true positive ones. First, all bounding boxes detected in the image are sorted according to their confidence scores. The most confident bounding box is selected as the true detection. Then, the Intersection over Union $IoU = \frac{Area\ of\ overlap}{Area\ of\ union}$ between the true detection's bounding box and every other detection's bounding box is calculated. The bounding boxes with higher IoU (above a certain threshold θ, equal to 0.5 in our work) are identified as belonging to the same object as the true bounding box and removed from the list of detections. The process continues from the next most confident bounding box out of the remaining ones and is performed for each class separately.

Although providing a reasonable level of performance improvement, this approach suffers from several limitations. Firstly, the final bounding boxes are simply bounding boxes detected by one of the base detectors, limiting the potential of fusing the knowledge of multiple methods. Secondly, the use of confidence score of the detector as a metric for deciding the quality of bounding boxes is questionable. The common interpretation of the confidence score is the probability of the bounding box to be the true positive one. However, such interpretation does not have a strong theoretical basis, as confidence scores of different object detection methods, being learned parameters, can reflect significantly different true-positive rates. In other words, the confidence score scales of different object detection approaches are not calibrated, and some methods tend to produce high-confidence detections that are false-positives, while others output true positive detections with low confidence scores.

3.2. Proposed method for ensembling bounding boxes

Taking a step towards more meaningful use of the knowledge provided by several object detectors, we propose the following approach that adjusts the final selection of bounding boxes based on how likely they are to correspond to true detections. First, the previously described IoU-based merging is employed to identify the bounding boxes belonging to the same object. However, instead of discarding all the bounding boxes corresponding to the same object as the most confident bounding box, we keep track of the source detectors for each bounding box. From the set of bounding boxes obtained from each detector, we select only the one having the highest IoU with the most confident bounding box. As a result, we obtain a set of detections, each described by 1-3 bounding boxes. At this point, we can observe that objects detected by only one of the detectors are generally false-positives, so out of the obtained set of detections, we discard the ones described by only one bounding box. In order to obtain the final detections, the selected bounding boxes for each detection need to be combined. To exploit the knowledge present in each of these bounding boxes, we propose to fuse them as a weighted combination with learned weights.

3.2.1. Improving precision of bounding boxes

In order to improve the precision of the bounding box coordinates, we further extend the proposed approach by taking a weighted average value of the coordinates of each of the vertices of the bounding boxes identified to belong to the same object. We weight each coordinate by the corresponding confidence score of the detection and normalize the resulting value by the sum of the confidence scores so that more confident detectors put more weight to the final output. Direct use of confidence scores would suffer from the limitations caused by the scores of different detectors not being calibrated. To address this issue, we propose a scheme for re-weighting the output of each detector that results in better calibration of confidence scores, and, therefore, in a more meaningful combination of the bounding boxes. Assuming that we have obtained a scalar weight w^j for the j^{th} detector, the new coordinates and confidence scores are calculated as

$$\hat{c_i} = \frac{\sum_{j=1}^{D} s_i^j w^j c_i^j}{\sum_{i=1}^{D} s_i^j w^j} \qquad \text{and} \qquad \hat{s_i} = \frac{\sum_{j=1}^{D} w^j s_i^j}{\sum_{j=1}^{D} w^j}, \tag{1}$$

where $\hat{c_i}$ and $\hat{s_i}$, are the refined coordinate and updated confidence score of i^{th} detection, D is the number of detectors, w^j is the weight of j^{th} detector, and s_i^j and c_i^j are the score and coordinate of i^{th} detection of j^{th} detector.

Appropriate weights w^j for each detector cannot be determined empirically. Therefore, we formulate a regression problem to learn these weights from a low number of manually annotated images. Let us denote by $\mathbf{b}_i^j = [x_{i1}^j, y_{i1}^j, x_{i2}^j, y_{i2}^j]$ a vector of coordinates

representing the i^{th} bounding box of detector j, where x_{i1}^j, y_{i1}^j, x_{i2}^j and y_{i2}^j are the coordinates of two of the bounding box corners. Let the confidence score corresponding to this bounding box be s_i^j and $\mathbf{g}_i = [g_{x1}, g_{y1}, g_{x2}, g_{y2}]$ be the corresponding groundtruth vector of coordinates. Assuming that we are operating with D distinct object detection methods, each i^{th} detection $\mathbf{X}_i$ can be represented by $D \times 4$ matrix $\mathbf{X}_i = [\mathbf{b}_i^1 s_i^1; \mathbf{b}_i^2 s_i^2; ...; \mathbf{b}_i^D s_i^D]$ of confidence score-scaled bounding box coordinates obtained from D detectors. Therefore, our goal is to find a $D \times 1$ dimensional set of weights $\mathbf{w} = [w^1, w^2, ..., w^D]^T$ that would satisfy the following criterion for each detection i:

$$\mathbf{w}^T \mathbf{X}_i = \mathbf{g}_i \Rightarrow \quad [w^1, w^2, ..., w^D] \quad \times \begin{bmatrix} \mathbf{b}_i^1 s_i^1 \\ \mathbf{b}_i^2 s_i^2 \\ ... \\ \mathbf{b}_i^N s_i^D \end{bmatrix} = \quad \mathbf{g}_i,$$

$$[w^1, w^2, ..., w^D] \quad \times \begin{bmatrix} x_{i1}^1 s_i^1 & y_{i1}^1 s_i^1 & x_{i2}^1 s_i^1 & y_{i2}^1 s_i^1 \\ x_{i1}^2 s_i^2 & y_{i1}^2 s_i^2 & x_{i2}^2 s_i^2 & y_{i2}^2 s_i^2 \\ ... & ... & ... & ... \\ x_{i1}^D s_i^D & y_{i1}^D s_i^D & x_{i2}^D s_i^D & y_{i2}^D s_i^D \end{bmatrix} = \\ = \begin{bmatrix} g_{x1} & g_{y1} & g_{x2} & g_{y2} \end{bmatrix}. \tag{2}$$

To obtain the solution to the problem, we optimize it iteratively by means of Stochastic Gradient Descent optimized over the Mean Squared Error (MSE), defined as

$$\mathcal{L}_{MSE} = \frac{1}{4n} \sum_{i=1}^{n} \sum_{d=1}^{4} (\mathbf{g}_{i,d} - \hat{\mathbf{g}}_{i,d})^2, \tag{3}$$

where n is the number of training samples, $\mathbf{g}_{i,d}$ is the d^{th} coordinate of the ground truth bounding box of i^{th} sample, and $\hat{\mathbf{g}}_{i,d}$ is the corresponding predicted coordinate.

3.2.2. Creating detection-ground truth pairs

The described process requires the creation of a training set of detection-ground truth pairs $\{\mathbf{X}_i; \mathbf{g}_i\}$. For this purpose, we first manually annotate a low number of images with the ground truth bounding boxes (100 in our experiments), as annotation of such a small number of images is not time-consuming, and apply the base object detection methods on these images. To reduce the amount of manual labour required for manually assigning each bounding box to the corresponding ground truth box, for forming the $\{\mathbf{X}_i; \mathbf{g}_i\}$ pairs we follow the following process: for each ground truth bounding box $\mathbf{g}_i$ we find all the overlapping bounding boxes of the corresponding class obtained by different detectors and select the ones having the IoU higher than a certain threshold θ. From this set, for each of the detectors we select the bounding box having the highest IoU with the ground truth box, and the rest of the bounding boxes are discarded. The obtained bounding boxes are then used to create the matrix $\mathbf{X}_i$ corresponding to the ground truth coordinates $\mathbf{g}_i$. In case one of the detectors did not produce a detection that would correspond to the groundtruth, we simply set the corresponding bounding box in $\mathbf{X}_i$ to zeros, i.e., $\mathbf{b}_i^j = [0, 0, 0, 0]$. We discard redundant duplicate detections and false-positive detections that did not match with any of the groundtruth bounding boxes, since the re-weighting of bounding boxes cannot fix the presence of incorrect detections as such.

3.3. Exploiting temporal information in videos

An additional step can be taken to further improve the performance on video data, by taking advantage of temporal information. We can safely assume that objects in a video sequence do not appear randomly at different frames, but follow certain trajectories of considerable time length. Therefore, we can enrich the set of detections obtained using the proposed ensembling approach with the ones that were likely to be missed and reduce the number of false-positive ones by following a two-stage tracking-based approach.

In the first stage, to obtain a set of detections that were likely missed by the object detectors, we apply a set of correlational trackers [16]. At the first frame, an object tracker is initialized from each of the detections of that frame and they are tracked throughout the video. At each subsequent frame, the tracked bounding boxes are matched with the detections of the frame following the IoU-based matching process. A successful match is defined by an IoU exceeding 0.5, or 0.4 in the case that the tracklet was initialized in the past 3 frames, due to the assumption that the shape of the object changes significantly when entering the field of view, leading to higher differences between the bounding boxes detected by the detectors and the one tracked by the tracker. For the detections not matched with any of the tracklets, a new tracker is initialized, and objects that were tracked successfully but for which no detections were found are continued to be tracked unless one of the following holds:

1. the tracklet was only matched with detections for a small number of frames (in our experiments we set this number to 5) and then missed for more than 5 frames,
2. the tracklet was not matched with any of the detections for more than 50 frames.

The first rule is intended for discarding tracklets that are likely to be initialized from false-positive detections, and the second rule is for discarding tracklets corresponding to objects that are no more present in the scene, but for which tracking continued. When an object that was tracked, but not matched with detections for a certain number of frames is rematched with a detection again, the bounding boxes predicted by the tracker at the frames where no detection happened are added to the set of detections.

At the second stage, a multi-object tracker [17] is applied to reduce the number of false-positive detections: using the set of detections enriched with the ones obtained from the tracker at the first stage, we identify the sequences of detections belonging to the same object. The resulting sequences that consist only of a small number of consecutive frames (less than 5 in our work) and thus the ones that are most likely corresponding to false-positive detections are discarded. Note that the choice of tracking methods here is dictated by their fast speed and any other tracking methods can be employed as well.

4. EXPERIMENTAL SETUP AND RESULTS

To assess the applicability of the proposed approaches to real-world problems and evaluate the ability of the base detectors to generalize to previously unseen data, we evaluate the algorithms on different datasets than they were trained on. We select three state-of-the-art object detectors as our base methods: SSD [9] trained on images resized to 512×512, YOLOv3 [10] on 416×416 images, and RetinaNet [11] on images rescaled such that the smaller side is equal to 800 pixels. All models are trained on MS-COCO dataset [18]. To evaluate the ability of the methods to generalize to previously unseen

Table 1. MAP results on PASCAL VOC dataset

		plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	monit.	TOTAL
IoU 0.5	**Ret.Net**	86.28	81.55	75.01	54.16	63.04	80.08	82.05	88.71	57.13	79.61	57.47	81.65	82.7	84.02	82.59	48.49	72.17	65.41	84.25	73.97	74.02
	SSD	82.8	80.14	72.09	52.55	51.78	81.56	77.75	88.49	56.24	76.64	61.32	81.47	83.84	82.35	79.39	41.62	70.75	65.82	87.85	71.13	72.28
	Yolov3	88.33	83.88	69.2	55.56	65.12	86.34	80.67	86.79	66.76	70.49	66.62	82.51	87.32	84.19	83.24	50.02	67.9	71.91	86.59	74.96	75.42
	NMS	90.92	86.57	**78.94**	63.79	**69.38**	**87.11**	84.88	**91.88**	**68.83**	82.43	66.72	86.84	89.14	87.05	86.16	**54.37**	**79.00**	**74.71**	90.1	78.65	79.87
	Our	**90.97**	**87.08**	78.59	**65.54**	68.45	87.08	**84.92**	91.85	68.47	**84.16**	**67.14**	**87.48**	**89.94**	**87.46**	**86.57**	53.38	78.69	74.33	**90.38**	**79.06**	**80.08**
IoU 0.75	**Ret.Net**	60.66	50.38	48.36	24.06	31.87	71.06	59.31	71.03	33.62	60.35	33.86	59.62	58.45	59.14	46.75	19.08	44.44	50.53	67.11	50.95	50.05
	SSD	56.11	46.80	42.79	22.20	23.82	70.51	55.95	65.67	28.41	56.50	34.46	56.89	54.79	53.67	41.92	12.14	46.89	45.08	65.33	47.66	46.38
	Yolov3	56.67	51.29	43.31	22.29	38.19	76.59	56.87	65.98	38.36	52.12	38.87	60.42	60.53	55.10	51.57	18.93	41.49	52.59	63.37	48.36	49.64
	NMS	62.72	53.85	51.46	27.74	39.38	77.11	60.85	71.17	39.42	64.30	39.15	63.89	62.82	58.78	53.26	19.71	52.93	55.66	69.13	51.52	53.75
	Our	**67.43**	**59.32**	**54.43**	**31.65**	**42.69**	**78.32**	**64.41**	**77.37**	**43.70**	**66.41**	**43.55**	**69.74**	**66.24**	**64.43**	**56.60**	**22.84**	**55.21**	**60.05**	**72.39**	**60.04**	**57.84**
IoU 0.85	**Ret.Net**	36.75	28.57	24.17	7.56	10.12	56.36	34.60	48.50	14.55	33.28	17.4	40.01	36.60	31.75	20.83	5.64	21.65	32.27	45.52	21.81	28.40
	SSD	27.31	16.25	16.70	5.11	7.71	55.51	27.50	33.09	9.47	29.89	11.99	28.30	29.11	19.72	14.33	1.95	23.10	21.57	32.45	15.97	21.35
	Yolov3	18.68	19.32	15.90	5.04	11.11	53.56	27.69	35.84	13.42	22.75	12.28	28.06	30.99	19.52	20.91	4.17	15.14	24.72	32.64	15.53	21.36
	NMS	34.10	22.07	24.30	5.97	11.79	56.55	31.00	43.30	13.73	34.55	13.11	35.02	36.99	25.04	21.60	4.14	25.46	26.92	43.73	18.93	26.42
	Our	**37.65**	**30.51**	**27.78**	**10.53**	**15.13**	**63.76**	**40.11**	**50.58**	**19.13**	**37.47**	**21.04**	**44.65**	**43.4**	**35.38**	**27.08**	**6.05**	**29.91**	**37.05**	**46.58**	**23.1**	**32.35**

Table 2. MAP@0.5 results on video datasets.

	RetinaNet	SSD	YOLOv3	NMS	Our	Our-tr.
EPFL Camp	67.89	65.08	68.12	67.38	70.47	**70.70**
EPFL Lab-6	91.24	90.48	88.80	92.12	92.21	**92.53**
Campus Aud	79.97	76.39	71.34	80.1	81.14	**82.36**

data, we report the results on the intersection of classes with previously unseen PASCAL VOC dataset (training + test set) [19] and three video datasets: EPFL Campus-7 dataset, EPFL Lab-6 dataset [20], and Campus Auditorium dataset [21, 22]. In all video datasets, only the people class is considered, since the groundtruth data is available only for this class. We use the Mean Average Precision (MAP) as a metric for evaluation of the detection performance as defined in PASCAL VOC 2012 challenge [23, 24]. Out of all the obtained detections, we discard the ones with a confidence score below 0.05 and report the MAP at the default IoU of 0.5 for each class separately as well as the total MAP. Besides, to evaluate the effect of the proposed approach on the precision of the bounding boxes, we report MAP at higher IoUs of 0.75 and 0.85 for each class in the image dataset. In our experiments, we perform 5-fold cross-validation. In the image dataset, 100 samples are used for training the score re-weighting model in each of the folds, with the remaining 9863 images used for testing. In the video dataset, we split the videos into 5 continuous segments, in each of which the last 100 images are used for training with the rest of the frame sequence used for testing. This is done because the proposed tracking-based approach requires an uninterrupted video sequence. The same test sets are used for reporting the results of separate object detectors and the mean MAP value across 5 folds is reported. For learning the weights, the bounding box coordinates x and y are scaled by the image width and height, respectively. From the resulting pairs obtained from 100 annotated images, 30% are taken for validation, and the regression model is trained on the remaining 70% with a learning rate of 10^{-5} starting from zero-initialized weights. The MSE is calculated on the validation set at each iteration and training proceeds until MSE stops improving for a number of iterations, after which the weights resulting in the best performance are selected. We report separately the results of each object detection method, the proposed approach, and the proposed approach refined by the tracking-based refinement scheme in the video datasets. We also report the results obtained by applying solely the non-maximum suppression to the detectors' output to showcase that the improvement of the detection performance is caused by the re-weighting scheme to a large extent.

The results on the object detection methods as well as the proposed approaches on image and video datasets are presented in Tables 1 and 2, respectively, where the best MAP is highlighted in bold. On the PASCAL VOC dataset, the best overall performance among the base detectors is achieved by YOLOv3 at MAP@0.5, and by RetinaNet at MAP@0.75. This allows us to conclude that YOLOv3 is able to detect the presence of objects of interest in general better, but RetinaNet produces the bounding boxes that match with the ground truth boxes more closely. This observation also reinforces the motivation of our proposed approach - by combining outputs of several object detectors we can combine the fair detection ability of less precise detectors such as YOLOv3, but compensate the precision of bounding boxes by more precise detectors such as RetinaNet. We observe that the proposed approach outperforms the base detectors with all the IoU thresholds. At MAP@0.5 the improvements range from 0.52% for the dining table class up to 9.98% for the boat class. At this IoU threshold our proposed approach is performing on par with non-maximum suppression, with overall improvement of 0.21%. However, the performance differences are increased with the increase in IoU threshold, - we achieve 4.09% and 5.93% better performance on than non-maximum suppression on MAP@0.75 and MAP@0.85, respectively. Besides, we outperform all base detection methods on all IoU thresholds, leading to 4.66%, 7.79%, and 3.95% MAP improvements on IoU of 0.5, 0.75, and 0.85, respectively. Some example results on PASCAL VOC dataset can be seen in Figure 1. Note that for clarity only the detections with confidence score of at least 0.4 are shown. In the video dataset, the proposed regression-based approach achieves a significant improvement over all of the detectors on all three datasets, and applying the refinement process based on tracking pushes this improvement further, leading to an overall improvement of 2.58%, 1.29%, and 2.39% on EPFL Campus-7, EPFL Lab-6, and Campus Auditorium datasets, respectively.

5. CONCLUSIONS

We proposed a method for ensembling multiple object detectors that re-weights the confidence scores and bounding box coordinates, as well as exploits contextual information. The method resulted in better MAP scores compared to base detectors, where the gap is especially large on higher IoU thresholds. The extention of the proposed method for video data utilizing temporal information pushes the improvement in performance even further. The proposed methods can, therefore, be utilized directly for obtaining improved object detection results or as a part of a framework for creating annotations on new datasets.

6. REFERENCES

[1] H. Jung, M. Choi, J. Jung, J. Lee, S. Kwon, and W. Young Jung, "Resnet-based vehicle classification and localization in traffic surveillance systems," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 61–67.

[2] G. Hamed, M. Marey, S. Amin, and M. Tolba, "Deep learning in breast cancer detection and classification," in *Joint European-US Workshop on Applications of Invariance in Computer Vision*. Springer, 2020, pp. 322–333.

[3] K. Chumachenko, A. Männistö, A. Iosifidis, and J. Raitoharju, "Machine learning based analysis of finnish world war ii photographers," *IEEE Access*, vol. 8, pp. 144184–144196, 2020.

[4] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved yolo-v3 model," *Computers and Electronics in Agriculture*, vol. 157, pp. 417–426, 2019.

[5] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *AAAI Conference on Artificial Intelligence Workshops*, 2012.

[6] B. Adhikari, J. Peltomaki, J. Puura, and H. Huttunen, "Faster bounding box annotation for object detection in indoor scenes," in *European Workshop on Visual Information Processing*. IEEE, 2018, pp. 1–6.

[7] V. Wong, M. Ferguson, K. Law, and Y. Lee, "An assistive learning workflow on annotating images for object detection," in *IEEE International Conference on Big Data*. IEEE, 2019, pp. 1962–1970.

[8] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: a survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg, "Ssd: single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[12] H. Zhou, B. Gao, and J. Wu, "Adaptive feeding: achieving fast and accurate detections by adaptively combining object detectors," in *IEEE International Conference on Computer Vision*, 2017, pp. 3505–3513.

[13] D. Marčetić, T. Hrkać, and S. Ribarić, "Two-stage cascade model for unconstrained face detection," in *International Workshop on Sensing, Processing and Learning for Intelligent Machines*. IEEE, 2016, pp. 1–4.

[14] S. Yang, A. Wiliem, and B. Lovell, "It takes two to tango: cascading off-the-shelf face detectors," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 535–543.

[15] S. Karaoglu, Y. Liu, and T. Gevers, "Detect2rank: combining object detectors using learning to rank," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 233–248, 2015.

[16] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference*. BMVA Press, 2014.

[17] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE International Conference on Image Processing*. IEEE, 2017, pp. 3645–3649.

[18] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft coco: common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[19] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[20] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2007.

[21] Y. Xu, X. Liu, L. Qin, and S. Zhu, "Cross-view people tracking by scene-centered spatio-temporal parsing," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4299–4305.

[22] Y. Xu, X. Liu, Y. Liu, and S. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4256–4265.

[23] J. Cartucho, R. Ventura, and M. Veloso, "Robust object recognition through symbiotic deep learning in mobile robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2018, pp. 2336–2341.

[24] J. Cartucho, "map (mean average precision)," 2018, [Online]. Available: https://github.com/Cartucho/mAP.