Detect2Rank : Combining Object Detectors Using Learning to Rank

Sezer Karaoglu, Yang Liu and Theo Gevers, *Member, IEEE,*

***Abstract*—Object detection is an important research area in the field of computer vision. Many detection algorithms have been proposed. However, each object detector relies on specific assumptions of the object appearance and imaging conditions. As a consequence, no algorithm can be considered universal. With the large variety of object detectors, the subsequent question is how to select and combine them.**

In this paper, we propose a framework to learn how to combine object detectors. The proposed method uses (single) detectors like DPM, CN and EES, and exploits their correlation by high level contextual features to yield a combined detection list.

Experiments on the PASCAL VOC07 and VOC10 datasets show that the proposed method significantly outperforms single object detectors, DPM (8.4%), CN (6.8%) and EES (17.0%) on VOC07 and DPM (6.5%), CN (5.5%) and EES (16.2%) on VOC10. We show with an experiment that there are no constraints on the type of the detector. The proposed method outperforms (2.4%) state-of-the-art object detector (RCNN) on VOC07 when RCNN is combined with other detectors used in this paper.

***Index Terms*—Object Detection, Fusion, Learning to rank**

I. Introduction

OBJECT detection is an active research area in the field of computer vision. Many detection algorithms have been proposed [1], [2], [3], [4], [5], [6], [7], [8]. Although these detection algorithms are successful for many detection tasks, they may be less accurate for some specific cases.

To gain more insight on the differences amongst detectors, Hoiem et al. [9] provide an extensive analysis on object detectors and their properties [9]. Their findings are that detectors perform well for common object appearances and common imaging conditions. Obviously, different design properties of the detectors (e.g. search strategy, features, and model presentation) influence the robustness of the methods to varying imaging conditions (e.g. occlusion, clutter, unusual views, and object size). For instance, detectors based on the sliding-window approach [1] using pre-defined window sizes and aspect ratios are good at finding likely object positions (rough object positions). However, they are less suited to detecting deformable objects precisely. Hoeim et al. [9] show that these types of detectors typically suffer from poor localization errors. Moreover, the large number of candidate regions to be considered limits the capability of sliding-window based object detection methods [2], [10]. Due to a large number of candidate regions (over $100K$ per image), it is not possible to perform object detection within an affordable time-frame while using strong classifiers [2]. The large number of candidate regions does not only restrict the classifier options but also influences the choice of the selected features. Extracting complex features from a prohibitively large number of sub-regions is not feasible due to its low efficiency [2], [7]. To avoid the limitations of a sliding-window approach, an object proposal method (selective search [7]), is integrated as a pre-processing step in current state-of-the-art techniques [5]. Selective search generates a significantly reduced set of candidate regions (around $2K$ per image). However, Hosang et al. [10] show that selective search generates candidate regions which are sensitive to changes in scale, illumination and geometrical transformations. This is because selective search is based on segmentation derived from superpixels which are unstable for small image deformations.

Besides the method to generate proper candidate regions for detection, the choice of features influences the robustness and discriminative power of the detectors. HOG-based templates are able to preserve the shape information [1], [4] of objects but are less suited for differentiating between visually similar categories such as cats and dogs. This limitation is addressed using color information in [3], following successful results of using color information in object recognition [11]. HOG-based object detection using color [3] is suited for object classes in which the intra-class color variation is low (e.g. potted plant and tv-monitor). However, the use of color negatively affects the detection accuracy for object classes in which the intra-class color variation is large (e.g. bottles and buses).

Finally, the chosen model and classifier drastically influences the performance of the detectors. In general, object detectors represent all positive samples of a given category as a whole [1], [3]. However, Malisiewicz and Efros [12] show that standard categories (e.g. train, car and bus) do not form coherent visual categories. Accordingly these methods are too generic. To address this issue Malisiewicz et al. [4] propose to train a separate linear SVM classifier for each positive sample in the training set. Gu et al. [13] show that using only one positive sample for training significantly reduces the generalization capacity. Hence, the detection performance of [4] deteriorates for uncommon object views.

As a consequence, no detection algorithm can be considered

S. Karaoglu is with the Computer Vision Group, University of Amsterdam, The Netherlands (e-mail: s.karaoglu@uva.nl).

Y. Liu is with the Computer Vision Group, University of Amsterdam, The Netherlands (e-mail: lawyoung529@gmail.com).

T. Gevers is with the Computer Vision Group, University of Amsterdam, The Netherlands, and also with the Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain (e-mail: th.gevers@uva.nl).

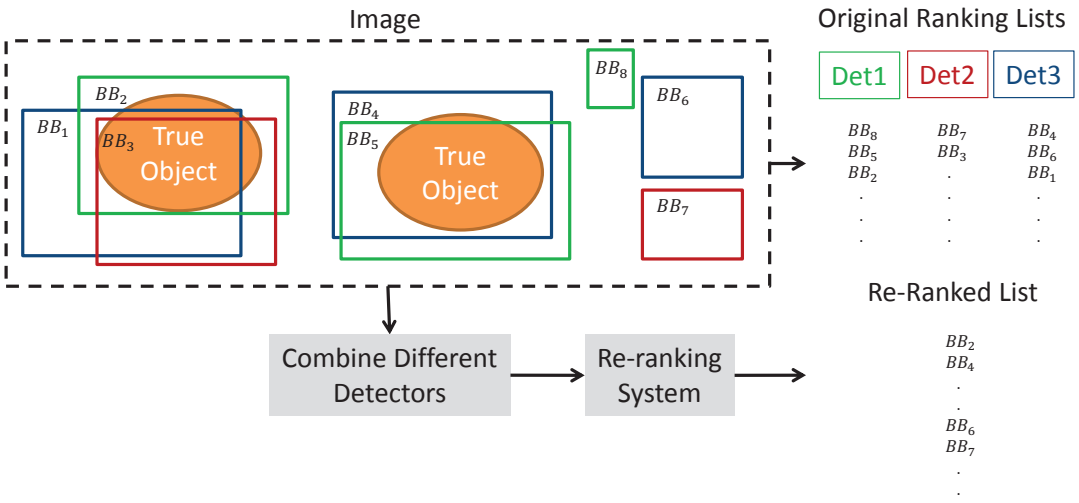


Fig. 1. Flow of the proposed method (Best viewed in color). Initial detections from different detectors namely, Det1(green), Det2(red) and Det3(Blue) are combined by a learning to rank algorithm. False detections of the individual detectors are learned by detector-detector relations and obtain less confidence when combined, whereas consistency in detectors BB_1, BB_2 and BB_3 are rewarded by the re-ranking system.

universal. With the large variety of available methods, the question is how to combine these object detectors to preserve their strengths while reducing their limitations and assumptions. In this paper, we consider a rank learning approach to combine object detection methods. The proposed framework combines detections (detector outputs which consist of a classifier score and bounding box locations) of different well-known object detectors including DPM [1], CN [3] and EES [4]. Furthermore, the method extracts high-level context features such as detector-detector consistency, detector-class preference, object-saliency of a detection, and object-object relations. These features are used in a learning to rank framework to yield a combined detection list. The flow of the proposed method is summarized in Fig 1.

The proposed approach offers the following advantages over single object detectors:

- Missed detections (false negatives) of single detectors are compensated by combining detections of different detectors.
- Detections are re-ranked by using information gathered by other detectors. True detections (true positives) of each detector are rewarded and false detections (false positives) of each detector are penalized within the learning to rank framework.
- The combined list maintains the strengths of the detectors. Therefore, it is more robust than each individual detector for varying imaging conditions.

To the best of our knowledge, we are the first to propose using re-ranking approaches to combine object detectors. Experiments on VOC07 and VOC10 show that the proposed method significantly outperforms single detectors. The proposed method (including code and the detector outputs) will be made publicly available. This allows other researchers to add new detectors.

Our contributions are the following:

- Detector combination: We provide a new perspective on how to approach the object detection problem. As there is no universal object detector, we propose to combine the state-of-the-art object detectors rather than creating a new one.
- Formulating detector combination: We formulate the problem of combining detectors in a learning to rank framework which has not been considered before in object detection.
- Detector contextual integration: We propose high-level context features (e.g. detector-detector relations and object-saliency cues) to combine detections in a learning to rank framework.
- Detector consistency: We show that the state-of-the-art detectors have many detections in common. These common detections are proven to be very informative to re-rank detection scores.
- Detector complementarity: We show that existing state-of-the-art object detectors also have complementary detections. These complementary detections reduce missed detections of single detectors in a combined list.

II. Related Work

A. Object Detection

In general, papers on object detection aim to design a single detector, descriptor or classifier [1], [2], [5], [6], [8], [14], [15]. Felzenszwalb et al. [1] propose a part-based object detection method using HOG features and a latent SVM. This algorithm outperforms the state-of-the-art methods for standard object appearances. The use of template-based models limits a detector's ability to detect deformable objects [9]. Moreover, template-based models (using HOG features) are designed to accommodate for shape information and are less suited to differentiate visually similar categories (e.g cats and dogs). In contrast to part-based detection methods, Vedaldi et al. [2] propose the use of a bag-of-words model for object detection. Multiple features are used within a multiple kernel learning framework which is able to distinguish between

visually similar object categories. However, Hoiem et al. [9] show that this approach is sensitive to object size due to the bag-of-words model. Khan et al. [3] propose to use additional color information for object detection. The color information contains expressive power for object classes in which the intra-class color variations are low (e.g. potted-plants or sheep). However, color may have a negative influence on the detection of classes in which the intra-class color variations are high (e.g. bottles or buses) [3].

Malisiewicz et al. [4] propose to learn a linear classifier per exemplar in the training set. The algorithm benefits from a large collection of simpler exemplar classifiers. In this way, the method is tuned to the appearance of the exemplar. While the detections of this detector cover the objects in the dataset (high recall), the detector usually provides low average precision. This is due to the large number of false detections introduced by each of the exemplar specific classifiers. Currently, remarkable results for object detection are obtained by convolutional neural networks [5], [6]. Girshick et al. [5] employ the CNN of [16] to a set of candidate windows obtained by selective search [7]. Recently, Hosang et al. [17] used various object proposals (BING [18], OBJ [19], CORE [20] etc.) to generate candidate windows and evaluate their performance for object detection using RCNN detector. The authors also report the best performance using candidate windows generated by selective search.

B. Contextual Information for Object Detection

Contextual information for object detection has been exploited over the past few years. Contextual information includes the relation between objects [21], [22], scene layout [23] or characteristics [24], [25], surrounding pixels [21], [26], [27] and background segments [28]. [25] shows that real-world scene structures can be modeled by inference rules. Therefore, in addition to the appearance of objects, contextual information provides useful information for object detection [29], [30]. For example, Choi et al. [24] model the object spatial relationships and co-occurrences by employing a tree-structured graphical model. Desai et al. [23] model the spatial arrangements between objects to detect objects in a structured prediction framework. Cinbis and Sclaroff [31] formulate the object and scene context in terms of relative spatial locations and relative scores between pairs of detections as sets of unordered items. Felzenszwalb et al. [1] re-score their DPM detections by exploiting contextual information as a post processing. Their re-scoring scheme relies on object co-occurrences as well as the location and size of the objects. The above methods show that contextual information is important for object detection. However, these methods have certain limitations. For example, the above methods rely on object-object co-occurrences and spatial relationships and hence are suited for images consisting of (many) different objects. Further, the context-based methods aim at re-scoring detections. They do not introduce new detections and hence are not able to recover from missed detections of single detectors.

C. Score Aggregation

The approach of aggregating the responses of classifiers and learning a second level SVM to re-score them for different tasks such as action recognition [32], image retrieval [33] and object recognition [34], [35] has been exploited in the literature. The organizers of Pascal VOC12 use seven methods submitted to the classification challenge. The scores of each submission are concatenated to form a single vector to train another linear classifier. Substantial increase for average precision is reported for classes such as potted plants and bottles. However, the problem of aggregating scores of different object detectors is not straightforward as other problems mentioned. More precisely, for these problems each instance in the dataset has a response from each classifier. By contrast, the object detectors do not generate candidate regions (exactly) at the same locations. Therefore, each candidate region does not necessarily contain a response from other detectors. Recently, Xu et al. [36] propose combining different pedestrian detectors through score calibration and detection clustering steps. The authors reduce false and missed detections of pedestrian detectors per image. However, they do not aim to perform a global ranking of detections over the entire dataset for different object classes. In a different work, Ladicky et al. [37] jointly estimates object location and segmentation by minimizing a global energy function on a Conditional Random Field (CRF) model. [37] combines results from detectors (single detector trained for different object classes), pairwise relationships between superpixels, and other low-level cues to perform better segmentation.

III. Object Detectors

In this section, the detectors used in this paper are outlined. We focus on publicly available detectors. Note that there are no constraints on the type of detector since the proposed method only requires detections (bounding box locations with classifier scores) of a detector.

A. DPM

Felzenszwalb et al. [1] propose an object detector in which each object category consists of a global template and deformable parts. The global template and deformable parts are represented by HOG features extracted at different scales. Training of the object models is done in a latent SVM framework. Each detection $\{x_1, x_2, ..., x_n\}$ in the training set is given a corresponding label, y_i, which is either $+1$ or -1. Each detection x is scored as

$$f_\beta(x) = \max_{z \in Z(x)} \beta.\Phi(x, z). \quad (1)$$

The set $Z(x)$ defines all possible latent values for detection x. β and $\Phi(x, z)$ is a vector of model parameters and a feature vector, respectively. β is trained by minimizing the following objective function:

$$L(\beta) = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n} max(0, 1 - y_i f_\beta(x_i)), \quad (2)$$

where $max(0, 1 - y_i f_\beta(x_i))$ is the hinge loss and constant C is the regularization parameter.

B. CN

Khan et al. [3] propose an object detector which uses color attributes as an additional feature alongside DPM based HOG features. The color attributes are combined with HOG features in a late fusion manner. The proposed color attributes are compact and efficient. They are proven to be effective for the object classes in which intra-class color variations are low such as potted-plants and sheep. Beside extending HOG features with color attributes, training is done exactly the same as in DPM.

C. EES

Malisiewicz et al. [4] propose an object detector which is trained by a parametric SVM for each positive exemplar in the training set. Consequently, a large collection of simpler exemplar specific detectors, which are highly tuned to the appearance of the exemplars, are obtained. Each exemplar is represented using a rigid HOG template [38] to train a linear SVM. Then, each Exemplar-SVM, (β_E,b_E), is used as a learned instance-specific HOG weight β_E vector to score. β_E is learned by optimizing the following convex objective function:

$$\Omega_E(\beta, b) = ||\beta_E||^2 + C_1 h(\beta^T x_E + b) + C_2 \sum_{x \in N_E} h(-\beta^T x - b), \tag{3}$$

where $h(x) = max(0, 1 - x)$ is the hinge loss and C_1 and C_2 are regularization parameters. Training each detector allows detectors to be tuned based on variations on the exemplar's appearance (viewpoint and object geometry). As a result, high recall is obtained for object detection.

IV. Combining Detectors by Learning to Rank

To combine detections from different detectors, learning to rank (L2R) is used. L2R aims to rank groups of items according to their relevance to a given task. Fig. 2 illustrates a common L2R flow. In our framework, the training set consists of detections $X = \{x_i\}_{i=1}^m$ (m is the number of the items in training set) and the ground truth label (y). Feature vector Φ and y are used in training data to learn a ranking model (g). To re-score detections, g is described as follows:

$$g(x) = w\Phi(x). \tag{4}$$

Using varied loss functions ξ (see section V), the weight (w) is optimized by minimizing the following objective function:

$$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i. \tag{5}$$

To learn a ranking algorithm that performs re-ranking, the proposed method starts with the feature extraction step using detections x from different detectors.

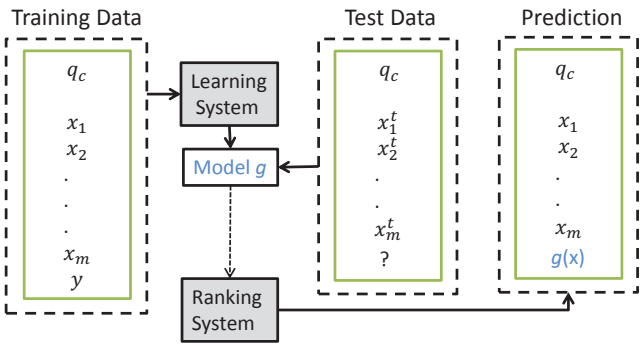


Fig. 2. Learning to rank framework for detection re-ranking.

A. Context Features

The proposed method starts with high-level context feature extraction to learn how to combine the ranking of detections from different detectors into a single detection list. We aim to extract generic features which exploit the correlation and consistency between detectors.

1) Detector-Detector Context: We introduce a notion of detector consistency which measures whether different detectors are generating object detections within the same image region. Agreement of all object detectors for a certain location increases the probability of a correct object detection. However, different detectors may generate detections at different locations even for the same image. As a result, it is hard to obtain an exact bounding box location where all detectors provide a detection. Therefore, a relative detector score is defined. To obtain a relative score for each detection, a correspondence term is computed by considering the overlapping ratios between all other detections. In this way, an image is represented as a collection of detections obtained by different object detectors j, where $j = \{1, 2 \dots n\}$ and n is the number of the detectors used. For the i^{th} detection in the image, the maximum overlapping detection with each detector is given by:

$$A_{i,j} = \frac{Area(BB_i \cap BB_j)}{Area(BB_i \cup BB_j)} \quad , \tag{6}$$

$$[\Gamma_i(j), \varphi_i(j)] = \max\left(A_{i,j}\right) \quad , \tag{7}$$

where Γ is the overlap ratio and φ is the index of the maximum overlapping detection for detector type j. Then, the corresponding relative score R of a detector j to the i^{th} detection is $R_{i,j} = \Gamma_i(j) \times S(\varphi_i(j))$, where S is the initial classification score of the detector. Note that if a detection has no overlap with other detectors ($\Gamma_i(j) = 0$), its relative scores will be zero. In this way, higher relative scores correspond to more reliable detections because more detectors agree on a particular location (see Fig. 3). If a detection has high relative score from each single detector it corresponds to a high probability of being a true detection. Whereas a low relative score corresponds to a false detection. Moreover, a mid-level consistency in relative score can be considered as a good indication of poor localization error.

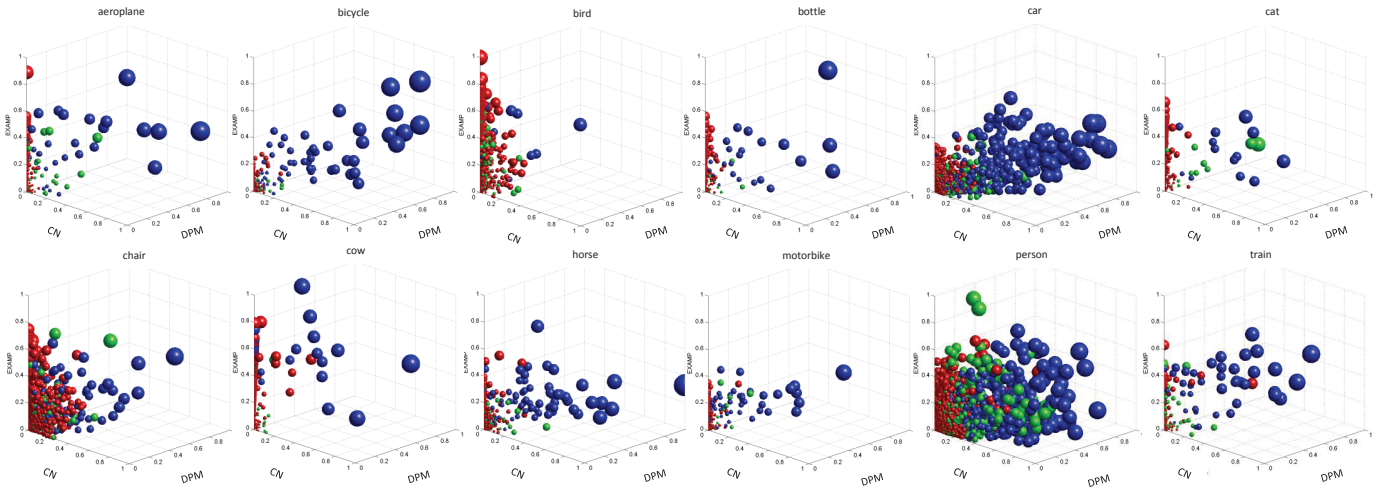


Fig. 3. The figure illustrates relative score R for each detection in VOC 2007 trainval set. Each sphere represents a detection in the trainval set whereas each axis represents relative score from detectors namely, DPM, CN and EES. The color blue, green and red holds for true detection, poor localization and false detection, respectively. Best viewed in color.

Relative score of a detection does not include the information of which detector it belongs to. However, some detectors perform better than others for some classes, hence their detections should get higher scores than detection of lower-performing detectors (to emphasize the strength of detectors on tasks for which they are successful). Therefore, a detector indicator term is specified. The aim is to provide information to the learning system for identifying detector preferences for particular classes. To give an indication of which detector the detection belongs to, a binary vector I_D of three dimensions (i.e. three detectors in our case) is used. The value of the dimension is assigned to be one in case of a detection by the corresponding detector otherwise the value is set to zero. This feature vector is at the detector level. Therefore, all detections of the same detector have the same binary coding I_D.

The final corresponding score feature Rs, for the i^{th} detection is denoted by $Rs_i = \{I_{D,i},\ R_{i,1},\ R_{i,2}, ...,\ R_{i,n}, R_{i,1} + R_{i,2}, R_{i,1} + R_{i,3}, R_{i,2} + R_{i,3}, ..., R_{i,n-1} + R_{i,n}, R_{i,1} + R_{i,2} + R_{i,3} + ..., + R_{i,n}\}$. The dimension of Rs is limited to the number of the detectors.

2) Object-Saliency: A feature vector O_s is proposed to represent how likely it is that a detection contains an object. EES [4], OBJ [19] and CORE [20] are used to measure the object-saliency of a detection. OBJ and CORE are category independent region proposal methods. They are mostly used by the current object detection algorithms to avoid an exhaustive sliding window search. These methods provide region candidates/proposals (bounding box) which are likely to contain objects. Both methods result in approximately 1000 candidate regions per image. In addition to these category independent region proposal methods, EES [4] is also used to provide region candidates. The overlap ratios between these different region proposals and object detections are calculated according to eq. 6. Then, the feature vector O_s for the i^{th} detection is given by:

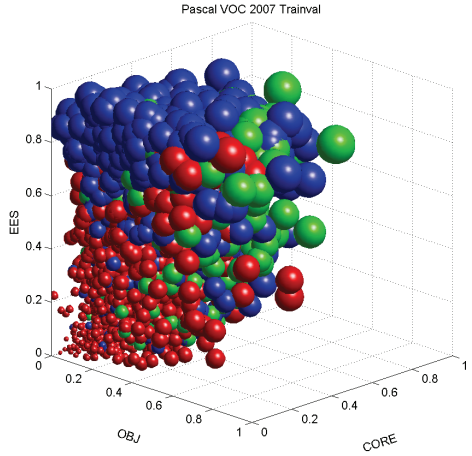


Fig. 4. The figure illustrates object likelihood score O_s for each detection in VOC 2007 trainval set. Each sphere represents a detection (randomly sub-sampled over all classes) in the trainval set whereas each axis represents object likelihood score from object indicators namely, OBJ, CORE and EES. The color blue, green and red holds for true detection, poor localization and false detection, respectively. Best viewed in color.

$$\Psi_{i,j} = sort(A_{i,j}) \quad , \tag{8}$$

$$O_s(i,j) = \frac{1}{n}\sum_{k=1}^{n} \Psi_{i,j}(k), \tag{9}$$

where n is the number of neighbors to measure object-saliency, Ψ is the sorted list of overlaps and j is the indicator of different regions proposals, namely OBJ, CORE and EES. Additionally, we use the confidence scores of the maximum overlapping neighbors of detections by EES [4] in eq. 9 since these regions proposals are class specific. A detection with a high object-saliency value is considered to be a good indicator for a correct detection. These features may be useful for

Feature	Notation	Dimension
Detector Relative Score	R_s	10
Object Likelihood Measure	O_s	4
Object-Object Context	S_o	20
Total		34

TABLE I
CONTEXTUAL FEATURES USED IN THE PROPOSED LEARNING TO RANK FRAMEWORK.

assigning lower confidence scores to false detections. Fig. 4 illustrates that true or false detections are highly correlated with the object likelihood scores.

3) Object-Object Relation: The likelihood of an object being present is inferred by using other object class likelihoods. Let $S_{c,j}$ be the detection with maximum confidence for object class c ($c = \{1, 2, \ldots, m\}$) by detector j (j=$\{1,2,3\}$) in an image, where m denotes the number of object classes. Then, the object-object context S_o is given by

$$S_o(c) = \sum_{j=1}^{3} S_{c,j}. \tag{10}$$

This feature exploits the object-object relations. For instance, when three detectors locate a cow with high confidence, it is less likely to have a sofa or tv in the same image.

The compactness of the proposed contextual features used in this paper is shown in Table I. We normalize each feature dimension by subtracting its mean and dividing by its standard deviation.

B. Learning

L2R methods are used to learn the ranking models. L2R methods used in this paper can be categorized in two groups [39]. The first type of algorithms is called pointwise techniques. Pointwise approaches represent the problem of ranking as a regression or classification problem. These techniques are straightforward approaches to learn the ranking model. Pointwise algorithms are preferred because of their efficiency and effectiveness. These methods have been optimized to work on large scale data.

The second type of L2R algorithms are pairwise techniques. These methods consider the problem of ranking as a pairwise classification problem. The aim is to learn a binary classifier to determine which instance is most relevant from a given pair of instances. The goal of these algorithms is to minimize the average number of misorders in ranking rather than the traditional misclassification in the ordinary pointwise approach.

C. Non-maximum Suppression

Duplicate removal for the same instance is a known problem for single detectors. Obviously, by combining multiple detectors, the proposed method increases the number of duplicates. To this end, we propose to suppress these multiple detections by non-maximum suppression (nms). The common application of nms considers all bounding boxes (over a certain overlap threshold) for suppression. We use only correspondences (overlaps between detections of other detectors) obtained for each detection in eq. 7 for suppression. After applying the re-ranking system, the corresponding detections are sorted and the highest among the others remains constant while detections which are at least 40% covered by the highest detection are suppressed.

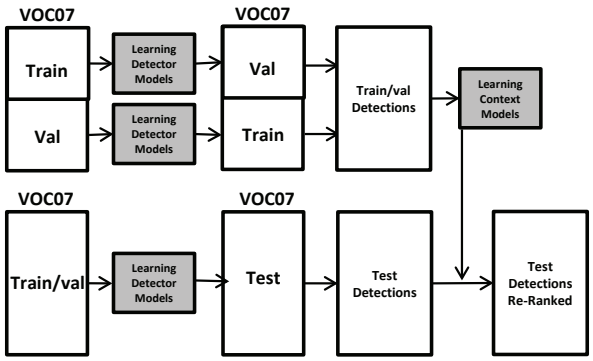


Fig. 5. Each training is used for learning detector models and context models. To avoid overfitting, the object detectors for context models are trained on the train set to generate detections on validation. Further, they are trained on validation to provide detections on train.

V. EXPERIMENTS

Experiments are conducted on the Pascal VOC07 and VOC10 datasets. VOC07 dataset consists of 9963 images of 20 different object classes (24640 annotated objects) with 5011 training images and 4952 test images. The VOC10 train/val dataset contains 10103 images of 20 different categories (23374 annotated objects). Object detections for the $train$ set are obtained via models trained on 2007val and detections for the val set are trained on the 2007$train$ set to learn detector-detector context. Detections for the $test$ set are obtained by models trained on the 2007$trainval$ set for both dataset evaluations. This process is summarized in Fig. 5.

A. Detector Bounds

In this experiment, we evaluate the maximal mAP that can be achieved by the detections of the baseline detectors and their combinations. The maximal mAP of a detector is calculated when all true detections are ranked at the top of the detection list (precision-recall curve of the maximal AP: precision is always at 1 and the cut-off is at the maximum recall). Since AP corresponds to the area under the precision-recall curve, AP for the maximal AP is $(1 \times max(Recall))$. Consequently, Table II corresponds to a recall table. Table II shows that re-ranking DPM, CN and EES detections results in a substantial performance improvement, 17.5%, 16.2% and 33.1%, respectively. This result shows the positive effect of re-ranking detection scores of object detectors.

Table II shows that DPM and CN have similar maximal mAPs of 45.6% and 46.2%, respectively. However, their

	aero	bike	bird	boa	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pers	plnt	shp	sofa	tra	tv	mAP
DPM [1]	26.7	56.9	2.6	12.8	21.9	46.0	55.3	13.7	19.0	19.4	12.6	2.2	58.1	47.3	40.9	6.8	15.0	26.9	43.4	38.8	28.3
CN [3]	28.7	55.9	6.3	11.6	18.2	44.3	55.5	17.7	18.3	20.5	14.9	4.9	57.3	48.9	41.5	15.0	21.8	28.1	44.1	45.7	30.0
EES [4]	17.9	47.2	2.8	10.6	9.1	39.3	40.3	1.6	6.2	15.3	7.0	1.7	44.0	38.1	13.2	4.6	20.0	11.6	35.9	27.6	19.7
M-DPM	39.3	66.5	29.2	25.5	36.2	58.2	73.4	36.3	53.8	33.6	19.9	22.5	74.7	65.5	62.5	35.0	28.5	37.2	66.0	51.6	45.8
M-CN	43.5	61.7	26.1	20.5	34.5	56.8	72.2	39.4	46.3	33.2	22.8	22.5	73.3	63.1	65.0	38.3	38.8	43.9	62.1	60.1	46.2
M-EES	47.7	72.4	38.3	37.3	46.1	64.3	64.1	45.0	44.4	50.8	44.7	43.1	69.5	63.4	54.9	35.6	47.9	50.2	62.8	73.7	**52.8**
M-(DPM + EES)	60.7	80.4	48.6	46.0	54.6	73.7	80.2	59.5	67.6	58.2	51.9	51.3	82.2	73.8	73.1	49.0	51.7	60.3	76.2	76.0	63.7
M-(DPM + CN)	48.8	68.0	36.8	27.4	40.7	62.9	77.4	49.4	61.5	39.8	31.6	33.7	78.7	70.8	71.2	48.5	42.1	49.4	70.6	61.7	53.5
M-(EES + CN)	59.3	79.5	46.6	43.7	54.6	72.8	78.9	62.0	63.2	55.7	51.5	52.1	82.5	71.7	74.7	50.0	55.8	66.1	73.4	76.3	63.5
M-All	62.5	81.3	52.3	47.5	56.7	76.1	82.3	65.9	71.2	59.0	55.3	56.9	84.2	75.7	77.5	56.0	57.0	67.4	78.4	76.3	**67.0**

TABLE II
MAP VALUES FOR BASELINE DETECTORS DPM, CN AND EES. CLASS SPECIFIC AND OVERALL MAXIMAL MAP VALUES OF BASELINE DETECTORS M-DPM, M-CN AND M-EES, AND THEIR COMBINATIONS M-(DPM+CN), M-(CN+EES), M-(DPM+EES) AND M-(ALL) ON PASCAL VOC07.

combination has a significantly higher maximal mAP (53.5%) than both of them individually. This shows that although these two detectors are very similar in nature, they have complementary detections. Furthermore, when the detectors have intrinsically different designs (e.g. DPM and EES or CN and EES), they produce more complementary detections. This can be derived by the performance gain obtained by combining DPM+EES and CN+EES in Table II, 10.9% and 10.7%, respectively. Consequently, the proposed method would benefit from more detectors.

Another observation that can be derived from Table II is that aside from detectors having complementary detections to each other, they also have detections in common. While these shared detections are useful to learn consistency in their output, complementary detections compensate for missed detections from each individual detector.

Table II shows that the performance of detectors is limited by their correct detections. Therefore, detector combinations always show higher mAP values than individual detectors. The proposed method highly benefits from this, whereas other context based re-ranking methods lead to a limited performance improvement (limited to correct detections of a single detector).

B. Direct Combination of Detections

In this experiment, several ways of combining (without learning) detector outputs are investigated. Because the detectors are trained independently, detector scores are not necessarily compatible. A calibration process [40] is applied before merging different detector outputs. Given a detection x and the learned sigmoid parameter (α, β), the calibrated detection score is calculated as

$$f(x|\alpha,\beta) = \frac{1}{1+exp(x\alpha+\beta)}, \tag{11}$$

where α and β for each detector are learned on the $trainval$ set. After the scores are calibrated, we evaluate three different approaches for combining detections:

- **NaiveI**, after scores are calibrated, detections are merged into a single list.
- **NaiveII**, after scores are calibrated, detections are sorted in a descending score order for each single detector. Then, detections are combined by taking one by one from the top of each sorted detector outputs.
- **NaiveIII**, the detectors are combined based on their training set performance. The output of the best performing detector is first added to the list followed by the others based on their performance.

After the detections are combined in a single list, nms (see section IV-C) is applied. It can be derived from Table III that naively combining detector outputs outperforms baseline scores. The improvements are due to the increase in recall of the combined detection list.

The minimum performance improvement is obtained by $NaiveII$. $NaiveII$ gives equal importance to each single detector. This means that although EES detections are not precise, they become as important as DPM and CN. Therefore, more false positives are introduced at the top of detection list which negatively affects the detection performance. This result shows the importance of properly weighting the detections.

$NaiveIII$ is expected to perform better than other naive methods since it incorporates the training performances of the baseline detectors. However, the $trainval$ performance of the baseline detectors explains the lower performance of $NaiveIII$. To obtain $trainval$ performance detector models are: a) trained on $train$ to test on val and b) trained on val to test on $train$ (see Fig. 5). Since the detectors are trained with fewer samples for $trainval$ detections, baseline performances do not necessarily correspond to their $test$ performances. Training with fewer examples has also an influence on our context models.

C. Learning to Rank Detectors

In this experiment, four different L2R algorithms are evaluated. The pointwise methods we use are the $L2$-regularized support vector classifier ($PoW1$), the logistic regressor ($PoW2$) and the support vector regressor ($PoW3$). The pairwise method is $RankSVM$ [41] ($PaW1$), since it is commonly used as a pairwise L2R method. Pointwise approaches represent the problem of ranking as a regression ($PoW3$) or classification ($PoW1, PoW2$). It takes as input the feature vectors for individual samples and learns a mapping

	aero	bike	bird	boa	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pers	plnt	shp	sofa	tra	tv	mAP
DPM [1]	26.7	56.9	2.6	12.8	21.9	46.0	55.3	13.7	19.0	19.4	12.6	2.2	58.1	47.3	40.9	6.8	15.0	26.9	43.4	38.8	28.3
CN [3]	28.7	55.9	6.3	11.6	18.2	44.3	55.5	17.7	18.3	20.5	14.9	4.9	57.3	48.9	41.5	15.0	21.8	28.1	44.1	45.7	30.0
EES [4]	17.9	47.2	2.8	10.6	9.1	39.3	40.3	1.6	6.2	15.3	7.0	1.7	44.0	38.1	13.2	4.6	20.0	11.6	35.9	27.6	19.7
NaiveI	31.0	61.6	6.1	13.7	22.7	48.9	58.4	19.6	20.5	22.3	19.3	3.9	63.2	52.1	44.3	14.5	22.7	31.5	47.8	47.4	32.6
NaiveII	30.8	57.7	6.1	14.2	20.2	47.6	55.2	13.3	16.7	22.4	20.0	4.4	61.4	50.2	33.4	11.8	23.4	28.0	46.4	41.6	30.2
NaiveIII	28.3	61.3	2.8	13.3	22.8	48.1	58.7	18.5	19.5	15.3	19.0	1.8	61.7	52.6	41.9	14.8	20.0	29.3	48.9	48.3	31.4
PoW1	36.8	62.7	10.0	18.1	24.3	51.6	59.5	21.2	22.5	25.4	22.4	7.8	64.2	57.3	44.9	18.7	26.7	34.1	54.1	47.8	35.5
PoW2	36.7	62.8	13.3	18.4	27.0	52.3	59.9	24.7	21.9	24.8	25.8	10.6	65.4	55.9	44.7	19.2	21.2	37.5	54.0	46.5	36.2
PoW3	35.6	63.1	9.7	17.0	25.0	51.2	60.0	21.3	22.5	25.1	21.5	8.1	65.0	56.4	43.8	18.2	27.0	33.9	53.5	48.2	35.3
PaW1	34.5	59.4	10.2	16.2	19.8	49.5	54.4	24.6	20.7	19.7	24.0	8.0	61.0	51.5	40.9	16.7	25.9	31.1	48.3	41.5	32.9
Imp	8.1	7.1	6.2	5.6	5.1	6.3	4.5	7.0	3.5	4.9	10.9	5.7	7.4	8.4	3.4	4.2	5.2	9.4	9.9	2.4	6.3

TABLE III
THE RESULTS USING LEARNING TO RANK ALGORITHMS. NAIVE: DIRECT MERGING METHODS WITHOUT LEARNING. IMP: THE IMPROVEMENT OVER MAXIMUM BASELINE DETECTOR BY MAXIMUM LEARNING ALGORITHM.

to the ground truth labels whereas pairwise approach takes as input pairs of feature vectors and maps them into binary labels indicating whether two samples are presented in correct order or not.

L2R algorithms differ mainly by their loss functions ($\xi(w; x_i, y_i)$) in eq. 5. ξ for ($PoW1$), ($PoW2$), ($PoW3$) and ($PaW1$) are $\max(0, 1 - y_i w^T x_i)$, $\log(1 + e^{-y_i w^T x_i})$, $(\max(0, |y_i - w^T x_i| - \epsilon))^2$ and $\max(0, 1 + w^T x_i - w^T x_j)$ respectively. w represents weights, x instances, y corresponding labels and ϵ parameter to specify the sensitiveness of the loss.

$Liblinear$ [42] implementations for pointwise approaches and rankSVM implementation by Joachims [41] are used with default parameter settings. Ground-truth overlap ratios are taken as training labels. Pascal VOC (> 0.5) overlap criteria is used to assign positive and negative labels for $PoW1$ and $PoW2$, while overlap ratios are directly used as training labels for $PoW3$ and $PaW1$. $PaW1$ requires pairwise preferences between samples, and these preferences are created based on their ground-truth overlaps. Since there is no preference between samples for which the ground-truth overlap equals 0, we do not generate preferences between those samples.

Table III shows that the proposed learning to rank approach outperforms the baseline detectors for all classes, DPM(7.8%), CN(6.2%) and EES(16.5%). While learning based methods always perform better, logistic regression ($PoW2$) based learning method performs slightly better than other L2R algorithms. Slightly better performance of classification- over regression-based pointwise methods can be explained by the fact that regressor methods try to predict continuous values and do not pay attention to the strict 0.5 overlap boundary of VOC evaluation. Therefore, errors within this range harm regressor results. However, classifier-based methods attempt to minimize these errors. The performance of $RankSVM$ is slightly lower than other L2R methods. This might be due to unbalanced data. The number of negative samples is significantly larger than positive samples. $RankSVM$ treats all the samples equally, therefore some pairs might be overly emphasized within the model.

Considering the low dimensionality of the proposed feature vector, the feature space may not be linearly separable. Therefore, other non-linear kernel options for the classifier could be tested. However, we avoid learning a non-linear SVM due to its long learning time and the need for costly parameter validation. Therefore, we use a feature mapping method proposed by Vedaldi and Zisserman [43]. A 34 dimensional feature vector is mapped to a higher dimensional feature space. The best performing linear classifier in Table III($PoW2$) is applied to this new feature space. Through use of this feature space, the $PoW2$ classifier obtains a 0.6% mAP improvement(36.8%). Increasing the dimensionality results in support vectors which are better able to separate the feature space. Increasing the feature vector dimension with additional context features may further improve the results.

The improvement by the proposed learning scheme over direct merging methods in Table III indicates that the performance gain is not only due to the increased recall but also the effectiveness of the contextual information and the chosen learning scheme.

D. Detection Error Analysis

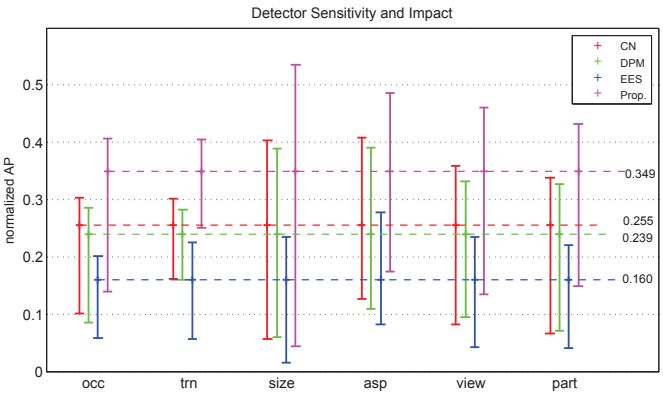


Fig. 6. Average (over classes) AP_N for the highest and lowest performing subsets within each different object characteristics such as occlusion, truncation, bounding box area, aspect ratio, viewpoint and part visibility.

To provide more insight into the performance obtained by combining the baseline detectors, we follow the procedure

introduced by Hoiem et al. [9]. Our first analysis regards detector sensitivities. The detector sensitivity is calculated based on the difference between max and min normalized AP for each characteristic (occlusion, truncation, bounding box area, aspect ratio, viewpoint, part visibility). Each colored plot in Fig. 6 shows the mean (over all classes) normalized AP for specified detectors. The results show that the proposed method does not reduce the sensitivity. However, it improves both the highest and lowest performing subsets for nearly all object characteristics. This indicates that the proposed method improves robustness for all object characteristics. The sensitivity is not reduced with the proposed method. This is due to commonly missed detections (hard detections cannot be detected easily even for human observers). While some of these hard detections are covered by one of the baseline detectors, they mainly remained undiscovered. That is why the minimum normalized APs for each characteristic increase but not as much as the maximum normalized APs. Consequently, the difference between max and min normalized AP increases.

Hoiem et al. [9] show the problem of small objects. Since small sized objects are mainly missed by all detectors, we observe that the min normalized AP for category "size" is not improved even if three baseline detectors are combined.

Fig. 7 shows the changes in the percentage of each false positive (FP) types with an increasing total number of FP. FPs are divided into four categories as follows:

- Poor localization (Loc) occurs when the label of detection is correct but misaligned with the ground-truth detection ($0.1 \leq$ overlap ≤ 0.5 or a duplicate detection.
- Confusion with similar classes (Sim) occurs when a false detection has an overlap with an instance of a similar class.
- Confusion with dissimilar object categories (Oth) occurs when a false detection is obtained for dissimilar classes.
- Confusion with background (BG) occurs when a false detection has no overlap with an instance of similar or dissimilar classes.

The errors originate from poor localization rather than other errors. This shows the effectiveness of relative score features. For instance, consider an image region where all detectors generate a detection. All detections belonging to this region have high classifier scores because of the high relative score. Consequently, these detections are ranked at the top of the detection list. However, the proposed method creates preferences for certain detectors when dealing with particular classes. Consider a detection by a detector preferred for a particular class that has a localization error within the region. The corresponding detections of the other detectors are suppressed by nms. The suppressed detectors may be true detections. This explains why top ranked false positives of the proposed method are mostly the result of poor localization.

Fig. 7 illustrates that the confusion with background error is significantly reduced. This shows the effectiveness of the proposed object likelihood features. Such strong object-saliency cues positively affect the proposed method to detect false detections.

Another observation shown in Fig. 7 is that the proposed features could not reduce the confusion caused by similar object categories. However, they are effective on limiting the confusion between dissimilar object categories.

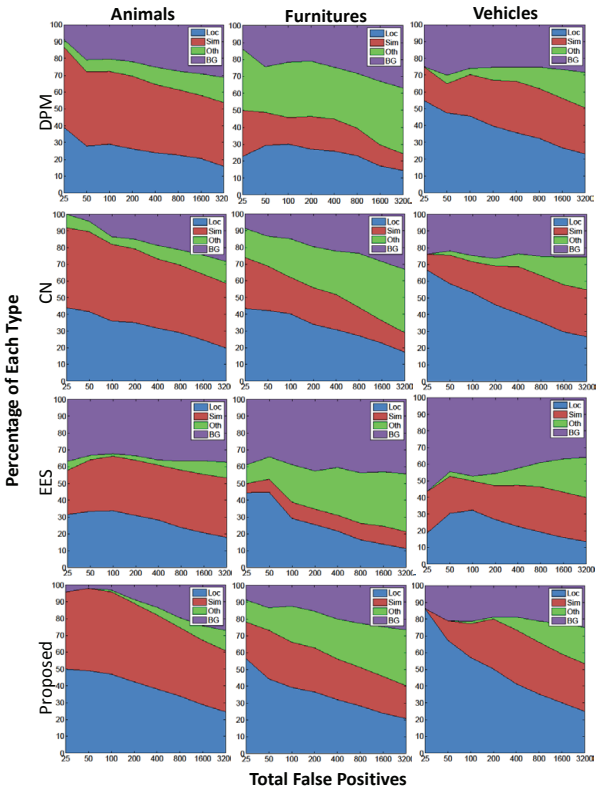


Fig. 7. Figure shows the fraction of false positives of each type (animal, furniture and vehicle) evolving as the total number of false positives increase.

E. Feature Importance

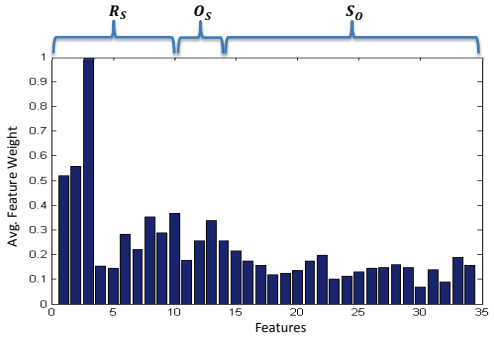


Fig. 8. Classifier weights are averaged over different classes to see the importance of features individually.

In this experiment, we study the influence of each individual feature. The weights are obtained by averaging the absolute classifier weights over the classes. The importance of proposed detector-detector context features (R_s) is highlighted in Fig. 8. Moreover, feature weights also emphasize the importance of proposed object-saliency features (O_s). As stated earlier, the proposed R_s and O_s features are more generic and independent of the number of object categories. However, object-object relationships exploited by other state-of-the-art context

	aero	bike	bird	boa	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pers	plnt	shp	sofa	tra	tv	mAP
R_s	33.7	62.4	9.5	15.2	22.9	50.4	59.6	18.8	22.0	22.8	20.3	6.2	62.7	53.7	44.6	16.5	24.4	34.4	50.9	46.7	33.9
R_s + O_s	35.6	62.1	10.6	17.4	24.6	50.8	59.3	25.1	21.3	23.2	23.9	10.5	63.1	51.0	45.5	14.7	26.3	37.8	50.6	47.5	35.1
R_s + S_o	35.4	63.7	10.6	18.2	26.5	51.7	60.3	18.7	22.7	24.1	21.5	6.6	63.8	57.3	43.7	18.5	24.3	34.5	53.0	45.3	35.0
All	36.8	64.2	12.3	20.3	27.3	53.0	60.3	27.0	22.0	25.3	27.1	11.1	63.7	56.6	45.4	19.3	24.0	38.0	54.5	46.8	36.8

TABLE IV
THE INFLUENCE OF SELECTED FEATURES FOR THE FINAL DETECTION PERFORMANCE.

	aero	bike	bird	boa	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pers	plnt	shp	sofa	tra	tv	mAP	Imp.
DPM + Context	29.8	57.5	9.9	16.4	24.1	46.3	58.0	21.1	19.6	20.4	15.1	7.5	58.3	50.4	42.0	14.3	18.2	28.0	49.0	39.6	31.3	3.0
CN + Context	33.3	55.1	11.4	13.4	22.7	44.9	57.0	22.6	18.6	19.4	17.5	8.5	56.0	50.9	42.1	17.4	20.9	31.3	48.5	45.5	31.9	1.9
EES + Context	31.4	57.2	10.6	16.9	21.0	46.6	51.5	13.3	15.5	20.6	15.2	8.1	57.3	51.5	32.9	14.1	18.0	20.1	46.9	44.5	29.7	10.0

TABLE V
THE RESULTS OF THE RE-RANKED SINGLE BASELINE DETECTOR OUTPUTS USING CONTEXTUAL FEATURES. THE RESULTS OF SINGLE DETECTORS ARE IMPROVED USING CONTEXT.

	aero	bike	bird	boa	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pers	plnt	shp	sofa	tra	tv	mAP
BL1	28.6	55.1	0.6	14.5	26.5	39.7	50.1	16.5	16.5	16.8	24.6	5.0	45.2	38.3	35.8	9.0	17.4	22.7	34.0	38.3	26.8
[23]	1.7	0	0.1	1.4	0	-3.5	1.3	0.5	-2.8	1.2	-0.7	0.2	0.5	1.1	-2.8	-1.1	-2.3	-0.7	0.5	0	-0.3
[24]	2.4	-4.2	2.3	0.8	-1.1	-0.2	-0.4	3.9	1.6	0.9	2.3	6.9	5.6	2.2	0	4.7	3.8	2.8	4.7	-0.1	1.9
[31]	5.6	2.7	9.2	0.8	3.2	1.9	3.4	5.0	0	0.7	1.4	7.9	5.9	4.6	3.5	4.2	3.1	4.9	4.9	0.3	3.6
BL2	27.8	55.9	1.4	14.6	25.7	38.1	47.0	15.1	16.3	16.7	22.8	11.1	43.8	37.3	35.2	14.0	16.9	19.3	31.9	37.3	26.4
[44]	2.4	1.9	0.5	0.2	3.2	2.6	2.9	-0.9	0.9	1.9	0.2	5.3	1.3	3.3	3.6	3.0	3.2	3.7	2.9	-0.5	2.0
BL3	26.7	56.9	2.6	12.8	21.9	46.0	55.3	13.7	19.0	19.4	12.6	2.2	58.1	47.3	40.9	6.8	15.0	26.9	43.4	38.8	28.3
Proposed	10.1	7.3	9.7	7.5	5.4	7.0	5.0	13.3	3.0	5.9	14.5	8.9	5.6	9.3	4.5	12.5	8.9	11.0	11.1	8.1	8.4

TABLE VI
COMPARISON OF THE STATE-OF-THE ART CONTEXT BASED OBJECT DETECTION METHODS ON PASCAL VOC07 DATASET. THE RESULTS OF REFERRED WORKS [23], [24], [31] AND DPM BASELINE SCORES ($BL1$) ARE REPORTED IN [31] WHEREAS [44] AND DPM BASELINE SCORES ($BL2$) ARE REPORTED IN [44]. $BL3$ IS DPM BASELINE SCORE OBTAINED IN THIS PAPER. THE RESULTS REPRESENTED AS PROPOSED ARE THE IMPROVEMENTS OVER DPM BASELINE IN THIS PAPER.

based object detection methods [24], [23], [31] is dependent on the image characteristics. Therefore, the accuracy gain is limited to the image characteristics for these methods.

We now investigate the influence of each feature on the final mAP score. The detector scores are essential for ranking the detection list. Therefore, it is not possible to evaluate O_s and S_o individually. We evaluate mAP using only the R_s feature. For the rest of the features, R_s is also included. It is shown in Table IV that using only R_s improves the baseline detectors significantly. An object likelihood measure also improves the accuracy (e.g. for animal classes such as cat, dog or sheep). Significant improvement for these classes is due to the poor representation capacity of template-based detectors for non-rigid objects. Deformable part based object detectors are well suited for detecting rigid parts of the objects (see top ranked visual results of category cat in Fig. 10.) Due to the homogeneous appearances of cats, dogs, and sheep, most object proposals contain the full object shape. Therefore, detections of the entire object receive higher confidence than detections for object parts. The object size plays role for other animals, such as horse and cow. Object proposal methods used in this paper tend to have better performance when detecting small sized objects. Moreover, it is less likely to happen that the object proposal methods generate many large bounding boxes for a specific image region. Therefore, the average overlap of a detection with these windows becomes lower. Adding object-object context (S_o) slightly improves most of the object classes. However, its contribution to the average precision increases when it is combined with the object-saliency. Furthermore, S_o clearly improves the accuracy for class "bottle" in which samples usually occur within a context (usually on a table or in the hand of a person).

F. Re-ranking Detections from a Single Detector

In this experiment, we exploit the effectiveness of context features without combining detectors into a single list. The proposed context features are only used to re-rank individual detectors. It is shown in Table V that the proposed method is still effective and improves the baseline detectors. However, the accuracy gain is relatively smaller than using the combined detector outputs in Table III. These results underline the importance of combining different detector outputs to recover from missed detections to improve the overall object detection performance.

Note that a detector with a high recall and low precision such as EES can be as powerful as other, more precise detectors (DPM, CN) using the proposed context features.

G. Comparison to Other Context Methods:

In this experiment, we compare the proposed method against the state-of-the-art context based object detection re-ranking

	aero	bike	bird	boa	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pers	plnt	shp	sofa	tra	tv	mAP
DPM [1]	26.7	56.9	2.6	12.8	21.9	46.0	55.3	13.7	19.0	19.4	12.6	2.2	58.1	47.3	40.9	6.8	15.0	26.9	43.4	38.8	28.3
CN [3]	28.7	55.9	6.3	11.6	18.2	44.3	55.5	17.7	18.3	20.5	14.9	4.9	57.3	48.9	41.5	15.0	21.8	28.1	44.1	45.7	30.0
EES [4]	17.9	47.2	2.8	10.6	9.1	39.3	40.3	1.6	6.2	15.3	7.0	1.7	44.0	38.1	13.2	4.6	20.0	11.6	35.9	27.6	19.7
RCNN [5]	62.4	70.9	46.5	37.3	31.8	63.3	72.1	62.3	28.3	64.1	49.2	56.2	66.2	65.2	53.2	28.4	53.1	49.9	57.2	62.2	54.0
Proposed	63.5	74.3	47.1	39.1	38.5	67.1	74.5	62.9	30.7	64.4	50.5	56.3	71.3	68.6	56.4	29.2	53.5	54.2	61.5	63.4	56.4

TABLE VII
THE RESULTS FOR BASELINE DETECTORS CN, DPM, EES, RCNN AND PROPOSED DETECTOR MERGING SCHEME ON VOC07 TEST SET. THE PROPOSED METHOD OUTPERFORMS ALL BASELINE DETECTORS OVER ALL CLASSES.

	aero	bike	bird	boa	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pers	plnt	shp	sofa	tra	tv	mAP
BL1	46.3	49.5	4.8	6.4	22.6	53.5	38.7	24.8	14.2	10.5	10.9	12.9	36.4	38.7	42.6	3.6	26.9	22.7	34.2	31.2	26.6
DPM-Context[1]	0.1	1.3	2.7	1.8	-0.6	1.8	2.9	-4.8	0.5	1.3	0.7	1.0	1.5	1.5	2.5	0.6	-2.8	4.9	6.6	2.7	1.2
[45]	6.5	-0.7	7.2	4.4	6.5	1.7	6.9	7.2	0.0	2.1	2.8	3.7	3.4	5.5	2.5	4.6	8.4	3.3	7.9	3.1	4.2
BL2	37.4	51.8	5.1	3.9	20.3	51.4	39.2	13.3	15.2	9.5	7.2	4.8	40.1	43.4	41.5	9.8	13.2	16.4	31.9	26.5	24.1
Proposed	7.3	2.9	8.5	6.6	2.5	7.1	5.1	15.0	2.8	3.5	5.8	7.1	3.6	7.7	3.5	8.0	8.4	3.4	10.6	11.8	6.6

TABLE VIII
COMPARISON OF THE STATE-OF-THE ART CONTEXT BASED OBJECT DETECTION METHODS ON PASCAL VOC10val. THE RESULTS OF REFERRED WORKS [45] AND DPM BASELINE SCORES ($BL1$) ARE REPORTED IN [45]. $BL2$ IS DPM BASELINE SCORE OBTAINED IN THIS PAPER. THE RESULTS REPRESENTED AS PROPOSED ARE THE IMPROVEMENTS OVER DPM BASELINE IN THIS PAPER.

	aero	bike	bird	boa	bot	bus	car	cat	chr	cow	tab	dog	hor	mbik	pers	plnt	shp	sofa	tra	tv	mAP
[1]	36.8	50.1	4.3	10.6	14.3	50.0	40.4	13.9	15.9	14.2	9.4	4.7	41.8	43.0	40.9	5.9	11.6	15.3	33.4	31.4	24.4
[3]	34.5	48.8	5.3	10.4	11.4	52.1	40.9	18.7	14.9	15.7	7.1	5.9	41.3	45.5	42.2	10.1	14.0	18.1	36.2	35.8	25.4
[4]	22.6	34.9	3.2	9.4	4.5	45.9	25.0	2.1	7.2	10.7	4.3	2.0	21.7	31.7	10.0	2.1	11.6	8.1	21.3	23.6	15.1
PoW2	44.8	53.3	14.3	14.6	14.2	56.3	44.7	27.2	18.9	19.6	14.5	15.0	44.1	50.0	45.4	13.2	17.6	22.5	42.0	39.1	30.6
[1]	37.4	51.8	5.1	3.9	20.3	51.4	39.2	13.3	15.2	9.5	7.2	4.8	40.1	43.4	41.5	9.8	13.2	16.4	31.9	26.5	24.1
[3]	36.6	45.0	6.0	4.7	17.9	52.5	40.2	18.8	15.3	10.6	6.5	5.2	39.7	44.4	44.0	15.5	16.4	13.0	35.6	33.8	25.1
[4]	19.9	36.8	1.8	3.3	7.2	46.2	23.5	2.0	4.2	6.4	2.1	1.3	20.6	30.4	9.5	2.8	14.5	7.0	24.0	24.7	14.4
PoW2	44.7	54.7	13.6	10.5	22.8	58.5	44.3	28.3	18.0	12.9	13.0	11.9	43.7	51.0	45.0	17.8	21.6	19.8	42.5	38.2	30.7

TABLE IX
THE RESULTS FOR BASELINES (DPM, CN AND EES) AND PROPOSED DETECTOR MERGING SCHEME USING POW2 ON VOC10 (UPPER: $train$ SET AND LOWER: val SET).

methods. Table VI shows the baseline scores of DPM and improvements reported by the papers [31], [44] on VOC07. The gain in performance by our method indicates the importance of high level contextual features and L2R based detector merging.

Moreover, the proposed method is compared to the recent work by Mottaghi et al. [45] on VOC10 dataset (See Table VIII). The authors also report on the context re-ranking method of DPM (See [1] for details) discussed in Section II.

The contextual features proposed by other methods in Table VI and Table VIII are from different sources. Hence, they can be complementary to the proposed features. Combination of these features may further improve the results.

H. Increasing Number of Detectors:

We performed another experiment to gain more insight into detector correlations and performance improvement. In this experiment, we focus on the state-of-the-art object detector of [5] (RCNN) in addition to three baseline detectors. The state-of-the-art detector of [5] uses the selective search paradigm [7] to generate object candidates which are classified by convolutional neural networks. [5] obtains the highest detection rate (in the literature) for the Pascal VOC 2007 dataset. The results are summarized in Table VII. Table VII shows that the proposed method improves the performance for all classes and outperforms [5]. This indicates that the proposed method is still effective when there is one strong detector (RCNN) which is implemented using substantially different methods than other detectors (CN, DPM and EES). Moreover, Table VII shows that RCNN significantly outperforms other detectors for classes "bird", "table" and "dog". However, combining weak detectors (CN, DPM and EES) still provides an improvement in the performance of RCNN for those classes. The maximum gain over RCNN is obtained for classes "bottle" (6.68%) and "horse" (5.12%). Additionally, 16% recall improvement over the single RCNN is obtained. This indicates that CN, DPM and EES still have complementary detections to RCNN.

I. Tests on VOC10

We also evaluate our method on the PASCAL VOC10 dataset. The VOC10 annotations of the test samples are not publicly available. Therefore, we use only the "$train/val$" dataset. All the training is done on the VOC07 $trainval$ set, including object detection models and detector-detector

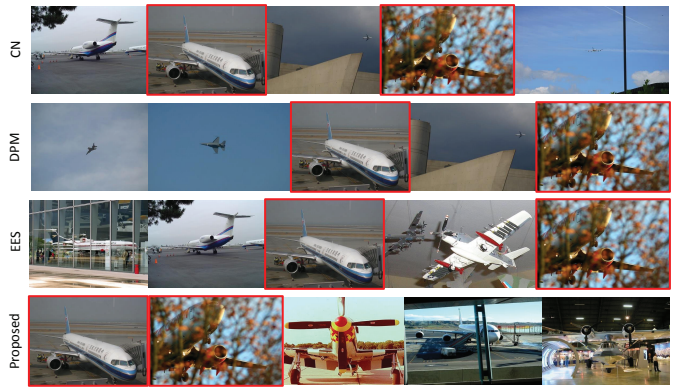


Fig. 9. The first 5 missed objects by single detectors and proposed method for class "aeroplane". The missed detections are listed based on ascending order of dataset image numbers. The first missed object by "CN" is also missed by "EES", however it is detected by "DPM". Therefore, the proposed method recovers this object. All three baseline detectors missed the objects outlined by red lines. Therefore, the proposed method misses these objects. All missed detections of single detectors which are not outlined by a red line are recovered within the proposed method.

relation models. Table IX shows the results. Table IX indicates that the proposed method outperforms the baseline detectors for all classes also on the cross dataset evaluation. The results show that the learned detector-detector context is generic and it is not dataset dependent.

VI. Discussion

A. Recall and Precision Improvement:

Missed detections of individual detectors are recovered when detections of different detectors are combined in a single list (increased recall). Table II shows that combining multiple detectors will lead to an increase in true object detections. This indicates that missed detections of individual detectors are recovered by the combined list. Fig. 9 also shows that the proposed method only misses objects which are missed by all three baseline detectors.

In [9], it has been shown that detectors can detect objects which contain consistent appearances. They experimentally derive that object detectors have common detections (Section V-A and Table II). Most detector outputs will overlap for true detections because their aim is to detect the same objects. Therefore, the overlapping information indicates the consistency between detectors and can be used to give more confidence to those detections which overlap with other detections (increase precision). The overlap information is useless for "orthogonal detections". The question is how to derive more confidence to those "orthogonal detections" to increase their precision. Therefore, the proposed approach makes use of other features such as "I_D-detector indicator", "object-saliency (O_s)" and "object-object relations (S_o)". These features are generic and independent of detector orthogonality.

B. Detector Correlation and Diversity:

In this paper, diversity, and thus potential complementary detections of CN, DPM and EES exist mainly due to three reasons. First, DPM and CN represent all positive samples of a given category as a whole (learn models per-category). However, EES proposes to train a separate linear SVM classifier for each positive sample in the training set (learns models per-sample). Accordingly, DPM and CN are more generic and EES is more discriminative. Second, not only the type of the feature but also how the features are used is crucial for object detection. DPM and CN represents objects using HOG features extracted from object parts and the whole object whereas EES represents objects using HOG features extracted only from the whole object. Moreover, CN uses color information as an additional feature. This results in complementary detections due to photometric invariance and discriminative power enabled by the color attributes. The discriminative power and photometric invariance do not always guarantee an improved object detection performance. Therefore, CN and DPM detections are complementary to each other. Third, the objective functions are different. DPM and CN minimize the inconsistencies between object parts using latent SVM whereas EES defines per-exemplar distance (more like nearest neighbor search) using linear SVM. These differences have a substantial influence on their final outputs. Table II, Table III and Fig. 12 represent the differences in the final outputs. Although these detectors have detections in common, they have also complementary detections. While common detections are useful to learn consistency in their output to increase precision, complementary detections resolve missed detections for each individual detector.

We also show that when detectors are implemented using substantially different methods, the proposed method still outperforms each individual detector. The experiment conducted in Section V-H indicates that increasing the number of the detectors will further increase the performance of the proposed method.

C. Computational Time vs Detection Performance:

There is a tradeoff between computational time and performance improvement. The computational time increases linearly with the number of detectors (assuming the same detection time per detector). The computational time can be reduced either by parallel processing or removing redundant operations e.g. the computation of HOG features, candidate regions etc. With a linear increase in time, the improvement in true object detection (recall) starts to slow down eventually. This deceleration depends on the complementarity of the detectors (See Table II). For instance, when CN and DPM are combined, the gain in recall is 7%. When combining EES with CN + DPM, the gain is 13.5%. However, detection performance is not only dependent on the recall. Increasing the number of detectors will yield a higher precision because of an enhanced consistency between detectors. Detections which are supported (overlap) by more detectors are good candidates to be true detections. Hence, agreement between more detectors increases the precision of detections (See Table V). Detections of single detectors become more precise with the help of other detectors. Eventually, the improvement in precision will also slow down (stop) when near optimal detections are provided.

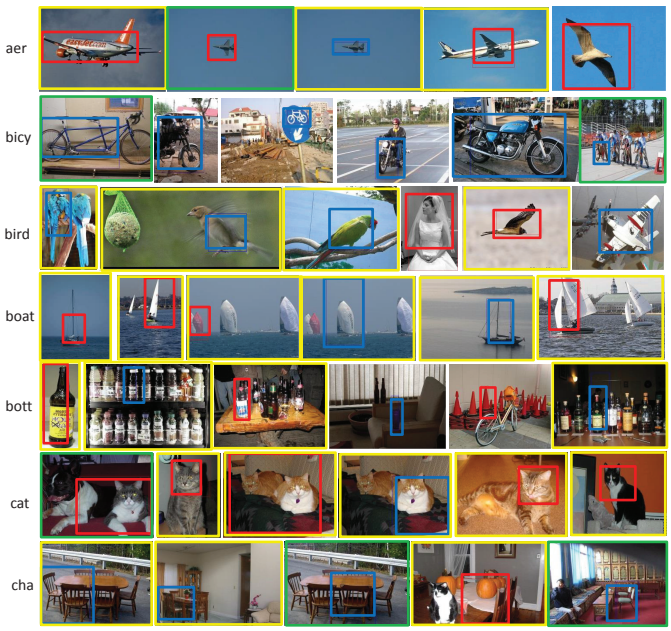


Fig. 10. Top ranked false positives of the proposed method for specified classes. Blue and red colors indicate the detector type, DPM and CN, respectively. Yellow and green colors correspond to poor localization and multiple detections, respectively. The image frames without color information indicates no overlap between ground truth object, either due to miss classification or background clutter.

Therefore, it can be argued that after a certain detection performance is obtained, the method may stop including more detectors.

D. The Choice of Learning to Rank Algorithm:

L2R methods are categorized into three groups: pointwise, pairwise and listwise. In general, these three methods differ by their objective functions. All three methods can be used within the proposed framework. However, there are some advantages/disadvantages for each method.

The selection of an algorithm is performed based on the following criteria: scalability, computational complexity and performance. Pointwise techniques are the most straightforward to learn the ranking model. These methods are optimized to deal with large scale data. Hence, they are fast in training and testing. Their main drawback is that the order between samples cannot be considered in the training step. This is because the algorithm optimizes loss functions based on individual sample errors.

A major advantage of the pairwise approach over the pointwise approach is that it focuses on the relative order of samples. Further, like the pointwise approach, the pairwise approach does not consider the position of all the samples in the ranked list. It does not require a quantitative label for each single sample but it requires pairwise preferences between samples. However, in our approach, each detection has a real value (ground truth overlap). Moreover, a pairwise approach is computationally more expensive than the pointwise approach due to the number of pairwise preference constraints formed by the pairs.

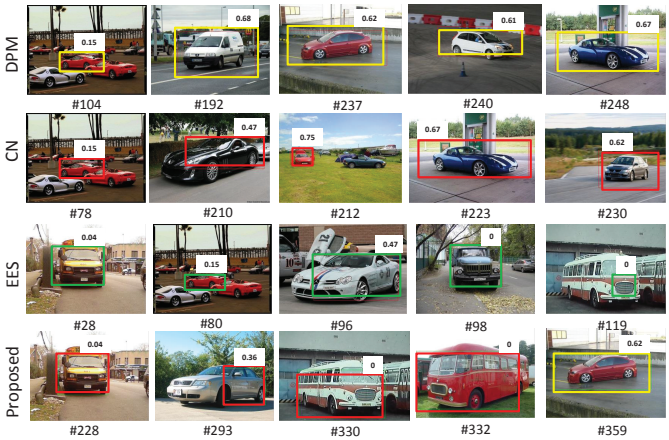


Fig. 11. Top five ranked false positives of baseline detectors and proposed method with their rankings below (object class car). The color of the detection yellow, green and red indicates the type of detector DPM, CN and EES respectively. The false positives of individual detectors are pushed down in the proposed method.

The listwise approach considers the position of all the samples in the ranked list when the loss function is optimized. Consequently, the listwise approach has an exponential number of ranking permutations yielding an increase in complexity and computational time. Incorporating the position information in the loss function may further improve the final results. However, the tradeoff between complexity and accuracy should be taken into account.

In fact, accuracy, training speed and memory requirements are important factors to select the number of detectors, object classes and number of training samples. For a (moderate size) dataset such as Pascal VOC (consisting of 20 object classes, 5K samples for training, and three detectors yielding in total 100K detections per class), pairwise solution is still tractable (since most of the detections do not have ground-truth overlap and there is no preference generated between those detections). While it is good to have more training data, it is challenging for pairwise algorithms to handle big data. For a dataset such as ImageNET (consisting of 200 object classes and 500K samples for training) the pairwise approaches need to be further optimized [46] to make the problem tractable. Pairwise approaches are still an area of active research and further improvements are possible by employing recent techniques such as active learning to rank proposed in [47]. Pointwise approaches are already capable of handling such large scale datasets and their performance are proven to be good for this task on Pascal VOC dataset. Therefore, for such large scale experiments pointwise approaches should be used.

E. Possible Improvements:

To avoid overfitting, the object detectors are trained on $train$ to test on val. Subsequently, they are trained on val to test on $train$, in which case the detectors are trained with fewer examples. This has an impact on the performance of detectors on the $train/val$ set in which we learn the relationship between detectors. It is observed that for some classes the performance of object detectors on the $train/val$

set are not inline with the $test$ set. Therefore, learning the models for detectors on a larger dataset may further improve the proposed learning to rank scheme.

The non-maximum suppression technique is a widely used ad-hoc method in object detection literature. However, learning to detect multiple detections from different detectors may be more appropriate for the proposed method.

The proposed method does not provide new bounding boxes. Therefore, it cannot recover from poor localization errors. Error resulting from poor localization becomes problematic for some cases (See Fig. 10 and Fig. 11 for top ranked false positives). This problem can be resolved by proposing new bounding boxes using object proposals or using a method similar to [5].

With the help of the proposed method, future object detectors can focus on more specific solutions to harder detection problems. Their results will be combined with other detection methods to carry object detection algorithms a step further. The contribution of a new method can be compared against the combination of the-state-of-the-art methods.

VII. Conclusion

No detection algorithm can be considered universal. As a consequence, we have proposed an approach to combine different object detectors. The proposed approach uses (single) object detectors to exploit their correlation by learning a re-ranking scheme.

The proposed method uses the agreement among the detections of different detectors to award a detection based on detector correlation and consistency. Furthermore, the proposed method exploits complementary detections of detectors to help recover missed detections of individual detectors.

Experiments on the PASCAL VOC07 and VOC10 datasets show that the proposed method significantly outperforms individual object detectors (DPM (8.4%), CN (6.8%) and EES (17.0%) on VOC07 and DPM (6.5%), CN (5.5%) and EES (16.2%) on VOC10.)

We show that there are no constraints on the type of the detector. The proposed method outperforms (2.4%) a state-of-the-art object detector (RCNN) on VOC07 when the RCNN is combined with other detectors used in this paper.

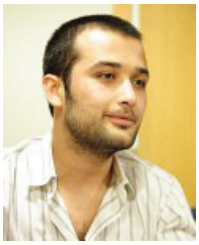
Acknowledgment

This publication was supported by the Dutch national program COMMIT.

References

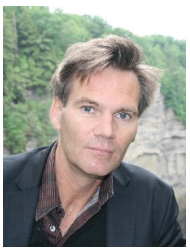
[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *TPAMI*, 2010.
[2] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *ICCV*, 2009.
[3] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, M. Vanrell, and A. M. López, "Color attributes for object detection," in *CVPR*, 2012.
[4] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *ICCV*, 2011.
[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
[6] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *ICLR*, 2014.
[7] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *IJCV*, 2013.
[8] R. G. Cinbis, J. Verbeek, and C. Schmid, "Segmentation Driven Object Detection with Fisher Vectors," in *ICCV*, 2013.
[9] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *ECCV*, 2012.
[10] J. H. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?" in *BMVC*, 2014.
[11] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *TPAMI*, 2010.
[12] T. Malisiewicz and A. A. Efros, "Recognition by association via learning per-exemplar distances," in *CVPR*, June 2008.
[13] C. Gu, P. Arbelaez, Y. Lin, K. Yu, and J. Malik, "Multi-component models for object detection," in *ECCV*, 2012.
[14] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *IJCV*, 2013.
[15] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *CVPR*, 2014.
[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
[17] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *arXiv:1502.05082*, 2015.
[18] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014.
[19] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *TPAMI*, 2012.
[20] E. Rahtu, J. Kannala, and M. B. Blaschko, "Learning a category independent object detection cascade," in *ICCV*, 2011.
[21] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *ECCV*, 2008.
[22] A. Farhadi and M. A. Sadeghi, "Phrasal recognition," *TPAMI*, 2013.
[23] C. Desai, D. Ramanan, and C. C. Fowlkes, "Discriminative models for multi-class object layout," *IJCV*, 2011.
[24] M. J. Choi, A. Torralba, and A. S. Willsky, "A tree-based context model for object recognition," *TPAMI*, 2012.
[25] A. Torralba, "Contextual priming for object detection," *IJCV*, 2003.
[26] G. Carolina, M. Brian, B. Serge, and R. G. L. Gert, "Multi-class object localization by combining local contextual interactions," in *CVPR*, 2010.
[27] M. Fink and P. Perona, "Mutual boosting for contextual inference," in *NIPS*. MIT Press, 2004.
[28] C. Li, D. Parikh, and T. Chen, "Extracting adaptive contextual cues from unlabeled regions." in *ICCV*, 2011.
[29] G. Carolina and B. Serge, "Context based object categorization: A critical survey," *CVIU*, 2010.
[30] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *CVPR*, 2009.
[31] R. G. Cinbis and S. Sclaroff, "Contextual object detection using set-based classification," in *ECCV*, 2012.
[32] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei, "Action recognition by learning bases of action attributes and parts," in *ICCV*, 2011.
[33] M. Douze, A. Ramisa, and C. Schmid, "Combining attributes and fisher vectors for efficient image retrieval," in *CVPR*, 2011.
[34] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *ECCV*, 2010.
[35] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *CVPR*, 2011.
[36] P. Xu, F. Davoine, and T. Denoeux, "Evidential combination of pedestrian detectors," in *BMVC*, 2014.
[37] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr, "What, where and how many? combining object detectors and crfs," in *ECCV*, 2010.
[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.
[39] T.-Y. Liu, *Learning to Rank for Information Retrieval*. Springer, 2011.
[40] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *ADVANCES IN LARGE MARGIN CLASSIFIERS*, 1999.
[41] T. Joachims, "Optimizing search engines using clickthrough data," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.
[42] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *JMLR*, 2008.

[43] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *TPAMI*, 2012.
[44] Y. Zhu, J. Zhu, and R. Zhang, "Discovering spatial context prototypes for object detection," in *ICME*, 2013.
[45] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.
[46] D. Sculley and G. Inc, "Large scale learning to rank," in *Workshop on Advances in Ranking in NIPS*, 2009.
[47] B. Qian, X. Wang, N. Cao, Y. Jiang, and I. Davidson, "Learning multiple relative attributes with humans in the loop," *TIP*, vol. 23, no. 12, 2014.



Sezer Karaoglu is a Ph.D. candidate at Computer Vision Group, Informatics Institute, University of Amsterdam. He was selected as tuition fee scholar for a European Master degree from Color in Informatics and Media Technology (CIMET) program. He holds double master degree. He graduated from optics, image and vision master degree at University Jean Monnet in France and media technology master degree at Gjovik University College in Norway. His research interests are Computer Vision, Pattern Recognition and Intelligent Human Machine Interactions.

Yang Liu received the B.S. degree and Ph.D degree in Dept. of Computer Science of Shandong University in 2008 and 2013, respectively. His research interests are in the areas of Scene Understanding, Color Constancy and 3D Scene Reconstruction.



Theo Gevers Theo Gevers is a Full Professor of Computer Vision with the University of Amsterdam (UvA), Amsterdam, The Netherlands. His main research interests are in the fundamentals of image understanding, 3-D object recognition, and color in computer vision. He is the Founder of Sightcorp and 3DUniversum, spin-offs of the Informatics Institute of the UvA.

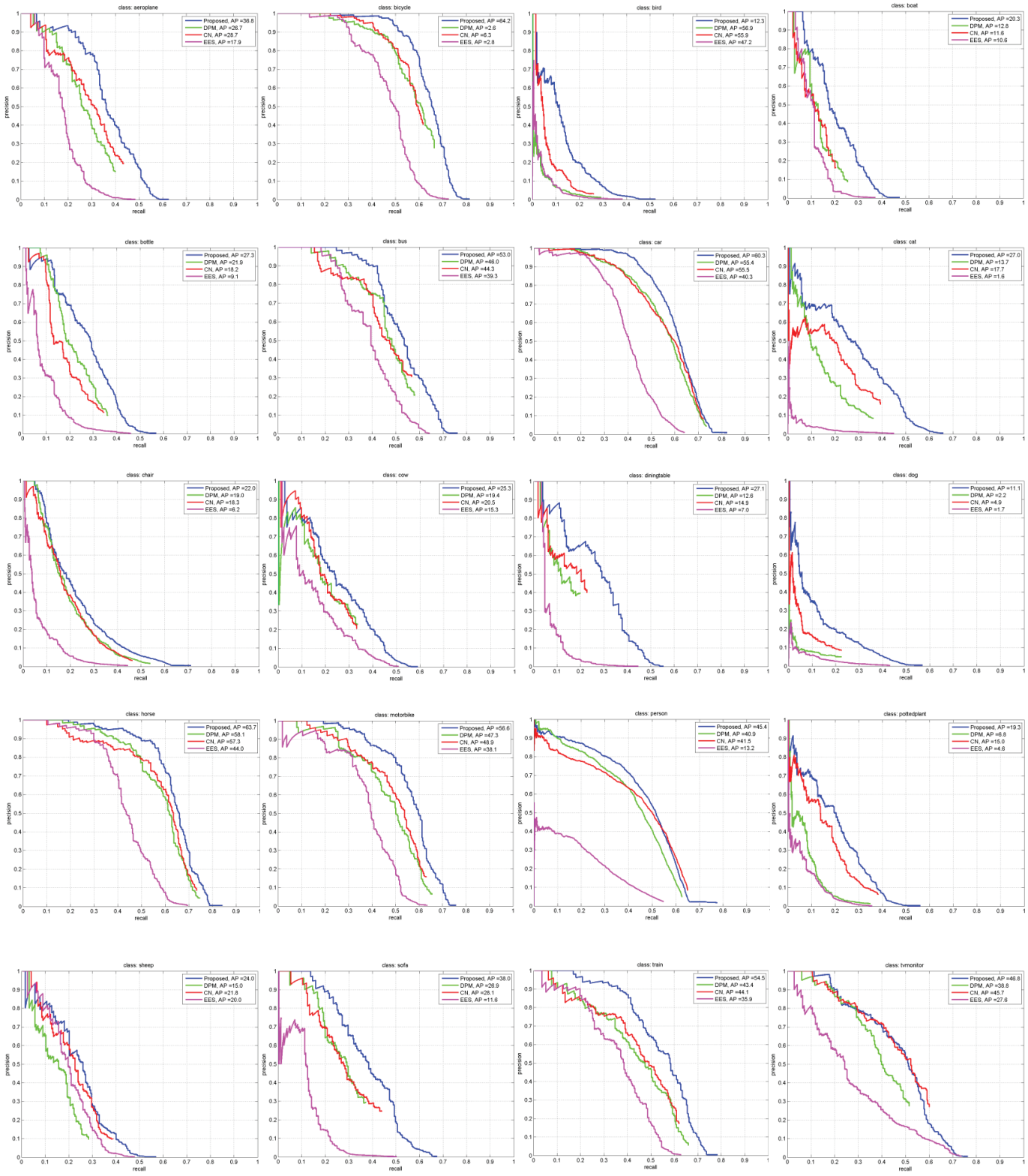


Fig. 12. Precision-recall curves on PASCAL VOC 2007. The proposed method significantly outperforms all single detectors. Furthermore, it is shown that detections of baseline detectors have remarkable differences.