ENHANCED OBJECT DETECTION VIA FUSION WITH PRIOR BELIEFS FROM IMAGE CLASSIFICATION

Yilun Cao[*†], *Hyungtae Lee*[*‡§], *and Heesung Kwon*[§]

[†]University of Southern California, Los Angeles, California, U.S.A.
[‡]Booz Allen Hamilton Inc., McLean, Virginia, U.S.A.
[§]U.S. Army Research Laboratory, Adelphi, Maryland, U.S.A.
yiluncao@usc.edu, lee_hyungtae@bah.com, heesung.kwon.civ@mail.mil

ABSTRACT

In this paper, we introduce a novel fusion method that can enhance object detection performance by fusing decisions from two different types of computer vision tasks: object detection and image classification. In the proposed work, the class label of an image obtained from the image classification task is viewed as prior knowledge about existence or non-existence of certain objects. The prior knowledge is then fused with the decisions of object detection to improve detection accuracy by mitigating false positives of an object detector that are strongly contradicted with the prior knowledge. A recently introduced novel fusion approach called dynamic belief fusion (DBF) is used to fuse the detector output with the classification prior. Experimental results show that the detection performance of all the detection algorithms used in the proposed work is improved on benchmark datasets via the proposed fusion framework.

***Index Terms*—** dynamic belief fusion, object detection, image classification

Fig. 1. **The proposed fusion concept:** the first row presents detection results from a person detector. In the second row, only the left-most image is determined as a 'person' image after fusing detection and classification results. True positives and false positives are indicated with white and red bounding boxes, respectively. Note that the precision values of the first and second row are 5/14 and 5/7, respectively. Precision can be increased by fusing object detection and image classification.

1. INTRODUCTION

Object detection is a fundamental problem in computer vision where one must localize and identify objects of interest in an image. Over the past decade, many algorithms have been developed to tackle the problem such as HOG-SVM [1], DPM [2], and CNN-based detectors [3, 4, 5, 6, 7]. In addition to efforts in algorithm development, attempts have been made to improve performance through preprocessing (e.g. object proposals [3, 8]), post-processing (e.g. bounding box refinement [9, 10]), and fusing the output of different algorithms (i.e. late fusion [11]). In particular, using late fusion approaches can be quite advantageous when the selected detection algorithms are complementary to each other, resulting in improved fusion performance. In the proposed work, a relatively unconventional approach compared to previous fusion methods is used that combines the outputs of two different types of computer vision tasks: object detection and image classification.

Image classification aims to determine the label of an image by calculating the likelihood of the presence of certain objects of interest inside the image. Image classification methods cannot inherently localize objects of interest in an image, yet they can provide useful information in the form of a degree of confidence about whether the image contains certain objects or not. The degree of confidence from classifiers can basically be viewed as prior knowledge about the existence (or non-existence) of the object classes in the image. This prior knowledge becomes very valuable when fused with the outputs of object detectors as it can possibly remove some false positives of object detectors if the prior knowledge strongly contradicts them. Similarly, the prior knowledge reinforces the findings of true positives if it strongly agrees with current detection results. Therefore, fusing the outputs of ob-

*The first two authors contributed equally to this paper.

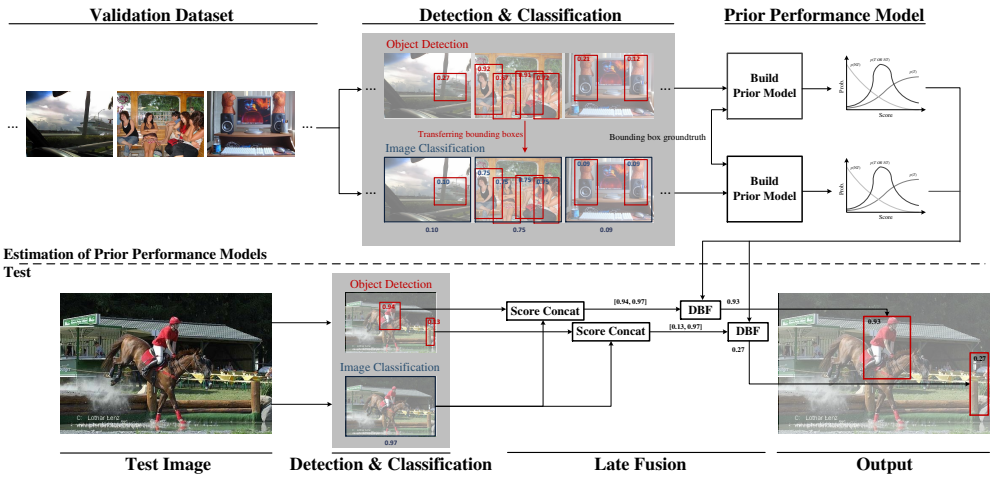

Fig. 2. Illustration of the overall process of the proposed fusion framework.

ject detectors and classifiers can possibly enhance precision, a ratio between the number of true positive and the number of true positives and false positives together, eventually improving average precision (AP). Figure 1 presents the proposed idea to use image classification to enhance object detection.

Optimally integrating decisions from object detectors and image classifiers is also a key to enhancing detection performance. Lee et al. [11] recently introduced a late fusion approach, called dynamic belief fusion (DBF), which can effectively integrate decisions from multiple complementary object detection algorithms providing enhanced fusion performance. DBF basically assigns probabilities to detection-relevant hypotheses, which are *target*, *non-target*, and *target OR non-target*, based on confidence levels in the detection results conditioned on the prior performance of individual algorithms. For object detection, DBF clusters candidate bounding boxes that potentially come from the same object generated by multiple algorithms. The clustered bounding boxes are the ones that are located closely to each other from which a fusion score is calculated. However, since an image classification approach cannot localize objects in an image, a strategy is needed that can convert classification scores to detection scores associated with corresponding bounding boxes. We use a relatively simple strategy that assigns a classification score of an image equally to all the bounding boxes (object candidates) found by object detectors from the same image.

For object detection, three detection algorithms with varying degrees of performance are selected: HOG-SVM [1], DPM [2], and Faster R-CNN [5]. We also use a weakly supervised convolutional neural network (WCNN) [12] to perform image classification. Therefore, in this work, fusion is performed on benchmark datasets using decisions from each of the three detection-classification pairs.

Our contributions are summarized as follows:

1. We introduce a novel fusion framework that can enhance detection performance of object detectors by using prior knowledge about existence or non-existence of certain objects in an image estimated from image classification.

2. To the best of our knowledge, the proposed fusion approach is the first attempt to combine detection and classification tasks to improve detection accuracy of current state-of-the-art detection approaches.

2. THE PROPOSED APPROACH

2.1. Overview

The proposed fusion framework consists of three steps: (i) training an object detection algorithm and an image classification algorithm, (ii) estimating prior performance models for individual algorithms, and (iii) integrating the outputs of individual algorithms by using DBF, a novel fusion algorithm previously developed by two of the authors. The dataset is divided into three non-overlapping subsets (*train* / *validation* / *test*). Note that both the object detection algorithm and the image classification algorithm are trained on *train* dataset, and *validation* and *test* sets are used for prior performance modeling and performance evaluation, respectively. The overall fusion process of the proposed work is illustrated in Figure 2.

Estimation of Prior Performance Models. With DBF, the prior performance models of an individual object detection algorithm and an image classification algorithm are estimated from the validation set. The prior models are estimated in the form of the precision-recall (PR) relationship to represent a level of prior confidence of both the detection and the classification algorithms. To calculate the PR curve for the object detector, all detection windows are labelled as *true* or *false positives* in reference to ground truth bounding boxes. If the intersection-over-overlap between a detection window and the corresponding ground truth bounding box is over a

certain threshold (e.g. 0.5), the detection is labeled as *true positive*, otherwise *false positive*. It is important to note that to calculate equivalent object detection performance of an image classifier, the output score of the classifier is converted to a detection score by assigning the classification score to all the detection windows equally found by the corresponding object detector for the same image, as shown in Figure 2.

Test. For each detection window of an object detector, the corresponding detection scores from both the detector and the classifier are concatenated to form a score vector, which is used as an input to DBF. DBF then estimates a fused score of the corresponding detection by integrating the score vector, a current observation, with the prior confidence models of both the detection and the classification algorithms.

2.2. Dynamic Belief Fusion (DBF)

To effectively fuse detection scores of each detection window from a detection-classification pair, a novel fusion method proposed by Lee et al. [11] called Dynamic Belief Fusion (DBF) is used to build a probabilistic fusion model.

For a two-class object detection problem, DBF uses a set of hypotheses defined as $\{T,\ NT,\ T\ OR\ NT\}$, where T and NT are a *target* and *non-target* hypothesis, respectively. $T\ OR\ NT$ represents detection ambiguity, which indicates that the subject observation could be either *target* or *non-target*. For each detection, the corresponding probabilities are assigned to the three hypotheses by linking the current detection score to the prior performance model of each detector, as shown in Figure 3. The prior performance model is basically a precision-recall relationship estimated from the validation dataset. Then the probabilities assigned to the three hypotheses are defined as

$$\begin{aligned} p(T) &= prec(s), \\ p(NT) &= rec^n(s), \\ p(T\ OR\ NT) &= 1 - prec(s) - rec^n(s), \end{aligned} \tag{1}$$

where $prec(s)$ and $rec(s)$ are precision and recall of detections whose scores are greater than s in the validation dataset, respectively. s is a detection score.

Once, for each algorithm, detection score is converted to the probabilities of the three hypotheses, a final fused probability is calculated by using Dempster's combination rule, which is defined as

$$p_D \oplus p_C(Z|s_1, s_2) = \frac{1}{L} \sum_{X \cap Y = Z, Z \neq \emptyset} p_D(X|s_1) p_C(Y|s_2), \tag{2}$$

where $L = \sum_{X \cap Y \neq \emptyset} p_D(X|s_1) p_C(Y|s_2)$ is a normalization term. p_D and p_C are the probabilities from detection and classification, respectively. X, Y, Z can be any hypothesis from a set of $\{T,\ NT,\ T\ OR\ NT\}$. $p(T) - p(NT)$ becomes the fusion out score. More details are described in [11].

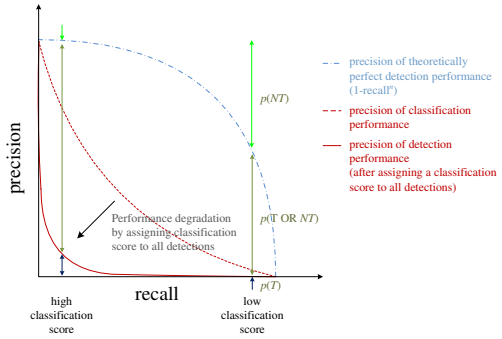

Fig. 3. The prior performance model of an individual algorithm. The plot also shows the probability assignments for *target*, *target or non-target*, and *non-target* hypotheses for an image classification algorithm. The equivalent detection precision of the classification algorithm is estimated by assigning the classification score to all the detections of a detector.

Effect of the Reduction of False Positives in the Background Image. Figure 3 shows that for a classification algorithm, the degradation of detection precision caused by false positives occurs mainly by assigning a classification score directly to all detection windows. Note that the precision degradation lowers $p(T)$ but does not affect $p(NT)$ since $p(NT)$ is only defined by recall that does not depend on false positives. For a test image with a high classification score does not affect detection accuracy because of relatively low $p(T)$ and $p(NT)$. But, for a test image with a relatively low classification score, a strong prior of NT, a large value of $p(NT)$ is assigned, which suppresses all the detections from a detector as false positives.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets. The proposed work is evaluated on the PASCAL VOC 2007 [13] and VOC 2012 [14], which have been widely used for evaluating object detection performance. VOC 2007 contains $\sim$ 2.5k images in `train` set, $\sim$ 2.5k images in `val` set, $\sim$ 5k images in `test` set, and $\sim$ 25k object annotations. VOC 2012 is a similar dataset with approximately twice the number of the images and objects in VOC 2007. [11] use a dataset partition of `train`/`val`/`test` to avoid overfitting in building prior performance models. In addition to this partition, we also evaluate with a common partition of `trainval`/`trainval`/`test` to validate output of individual algorithms by comparing to the performance reported in the original literatures of selected detection algorithms.

Image Classification. We use a recently introduced weakly supervised convolutional neural network (WCNN) [12],

Table 1. **VOC2007 detection performance.** The mean of average precision (mAP) across all object categories is used as an evaluation metric. FUSION indicates fusing a object detector with an image classification, WCNN. FR RCN is Faster R-CNN with ZF net [16].

method	train set	val set	mAP	gain
HOG-SVM	train		.184	
FUSION	train	val	.248	+ .064
DPM	train		.222	
FUSION	train	val	.237	+ .015
FR RCN	train		.585	
FUSION	train	val	.607	+ .022
HOG-SVM	trainval		.228	
FUSION	trainval	trainval	.303	+ .075
DPM	trainval		.312	
FUSION	trainval	trainval	.343	+ .031
FR RCN	trainval		.643	
FUSION	trainval	trainval	.660	+ .017

Table 2. **VOC2012 detection performance.** The mean of average precision (mAP) across all object categories is used as an evaluation metric. FUSION indicates fusing a object detector with an image classification, WCNN. FR RCN is Faster R-CNN with ZF net [16].

method	train set	val set	mAP	gain
HOG-SVM	train		.179	
FUSION	train	val	.250	+ .071
DPM	train		.258	
FUSION	train	val	.312	+ .054
FR RCN	train		.534	
FUSION	train	val	.553	+ .019
HOG-SVM	trainval		.203	
FUSION	trainval	trainval	.277	+ .074
DPM	trainval		.288	
FUSION	trainval	trainval	.349	+ .061
FR RCN	trainval		.569	
FUSION	trainval	trainval	.583	+ .014

which provides significantly enhanced performance in image classification on the PASCAL VOC dataset. The architecture of WCNN consists of 9 convolutional layers, the first five of which are pre-trained on ImageNet [15]. All 9 layers are further fine-tuned to the PASCAL VOC 2007 and 2012 datasets. In WCNN, an input image is first decomposed into multi-scale images to which multi-scale CNN networks are applied. The classification scores from individual multiscale CNN pipelines are averaged together to produce a final output score. Since training the CNN pipelines does not require bounding box labels, the algorithm is called weakly supervised CNN.

Object Detection. For object detection algorithms, three different algorithms with varying degrees of performance are used: (i) support vector machine with histograms of oriented gradient features (HOG-SVM) [1], (ii) deformable part models (DPM) [2], which represents objects as a collection of local parts, and (iii) faster R-CNN [5], which is the current state-of-the-art in object detection and also runs in real time.

3.2. VOC 2007 and 2012 Results

Detection Accuracy. Tables 1 and 2 show the detection performance of the three detection algorithms as well as the fusion performance with WCNN on both VOC 2007 and VOC 2012, respectively. It is shown that all the detection algorithms benefit from fusing with WCNN, via DBF on both datasets. The fusion gain is mainly attributed to the reduction of false positives, objects of non-interest recognized as objects of interest by the detectors. The false positives that are strongly contradicted with the classification prior indicating a high level of likelihood of non-existence of certain objects are basically eliminated through the fusion process. It is observed that the fusion gain for HOG-SVM is much greater than DPM and FR-RCN. This is because HOG-SVM results in more false positives than DPM and FR-RCN, many of which are removed using the prior knowledge from WCNN.

Dataset Partitions. We use two different dataset partitions for evaluation. The first partition, which is `train` / `val` / `test`, avoids overfitting while optimizing both training detectors/classifier and building prior performance models. The second partition, which is `trainval` / `trainval` / `test` allows the overfitting. However, the evidence of performance degradation by overfitting has not been observed.

4. CONCLUSIONS

We have introduced a novel fusion framework that can enhance the detection accuracy of state-of-the-art object detection algorithms by fusing with the prior knowledge about the existence (or non-existence) of certain objects obtained from image classification. In the proposed work, we mainly focus on mitigating false positives from object detection by using the proposed fusion strategy that can eliminate any false positive strongly contradicted with the prior knowledge estimated from image classification. The experimental results in Tables 1 and 2 show that the reduction of false positives via the proposed fusion approach directly leads to enhanced detection accuracy. It is also observed that the proposed fusion with the detection algorithms with more false positives, such as HOG-SVM and DPM, provides greater fusion gain than the fusion with faster R-CNN. This confirms the basic premise of the proposed fusion strategy that the prior knowledge from image classification can effectively reduce false positives.

5. REFERENCES

[1] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[2] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[4] Ross Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[5] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[7] Archith J. Bency, Heesung Kwon, Hyungtae Lee, S Karthikeyan, and B. S. Manjunath, "Weakly supervised localization using deep feature maps," in *European Conference on Computer Vision (ECCV)*, 2016.

[8] Pedro Felzenszwalb, Ross Girshick, and David McAllester, "Cascade object detection with deformable part models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[9] Spyros Gidaris and Nikos Komodakis, "Object detection via a multi-region & semantic segmentation-aware cnn model," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[10] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang, "Iterative locallization refinement in convolutional neural networks for improved object detection," in *IEEE International Conference on Image Processing (ICCV)*, 2016.

[11] Hyungtae Lee, Heesung Kwon, Ryan M. Robinson, William D. Nothwang, and Amar M. Marathe, "Dynamic belief fusion for object detection," in *IEEE Winter conference on Applications of Computer Vision (WACV)*, 2016.

[12] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic, "Is object localization for free? weakly-supervised learning with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[14] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The PASCAL Visual Object Classes challenge: A retrospective," *International Journal of Computer Vision (IJCV)*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, A. Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[16] Matthew Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision (ECCV)*, 2014.