S-CNN: Subcategory-Aware Convolutional Networks for Object Detection

Tao Chen, Shijian Lu, and Jiayuan Fan

Abstract—The marriage between the deep convolutional neural network (CNN) and region proposals has made breakthroughs for object detection in recent years. While the discriminative object features are learned via a deep CNN for classification, the large intra-class variation and deformation still limit the performance of the CNN based object detection. We propose a subcategory-aware CNN (S-CNN) to solve the object intra-class variation problem. In the proposed technique, the training samples are first grouped into multiple subcategories automatically through a novel instance sharing maximum margin clustering process. A multi-component Aggregated Channel Feature (ACF) detector is then trained to produce more latent training samples, where each ACF component corresponds to one clustered subcategory. The produced latent samples together with their subcategory labels are further fed into a CNN classifier to filter out false proposals for object detection. An iterative learning algorithm is designed for the joint optimization of image subcategorization, multi-component ACF detector, and subcategory-aware CNN classifier. Experiments on INRIA Person dataset, Pascal VOC 2007 dataset and MS COCO dataset show that the proposed technique clearly outperforms the state-of-the-art methods for generic object detection.

Index Terms—Subcategory, object detection, convolutional neural network, ACF detector, subcategory-aware CNN

1 Introduction

Object detection is one fundamental problem in computer vision research. One major challenge in object detection is large intra-class variations in object appearance and deformation, as illustrated in Fig. 1, where the two typical object classes both present large intra-class variations due to different viewpoints, partial occlusions, clutters, etc. Different object detection techniques have been reported including the widely investigated Adaboosting based Aggregated Channel Features (ACF) detector [1], [2] and Histogram of Gradients (HoG) based Deformable Parts Models (DPM) [3], [4], [5]. The ACF detector is fast but often susceptible to object viewpoint changes and deformations. The DPM detector solves the object deformation problem to certain degrees, but often fails to deal with many free-form objects that do not have a well-defined part structure.

Motivated by the powerful convolutional neural network (CNN) in object recognition [6], [7], the object detection using region proposals followed by CNN classification has gained significant performance improvement in recent years [8], [9], [10], [11], [12]. Several representative models such as R-CNN [8], Fast R-CNN [9] and Faster R-CNN [10] have achieved the state-of-the-art results and clearly outperformed the ACF and DPM based methods on several benchmark datasets, including PASCAL VOC [13] and ILSVRC [14]. The success of Fast/Faster R-CNN can be attributed to the supervised deep CNN that learns the discriminative convolutional features, and the region proposal that generates a large number of object candidates, e.g., selective search [15] in R-CNN and Fast R-CNN and region proposal network (RPN) [10] in Faster R-CNN.

More recently, another class of methods which do not rely on object proposals as in R-CNN but directly learn to regress the object bounding boxes from the convolutional feature maps have been reported in [16], [17]. The You Only Look Once (YOLO) method [17] computes a global feature map and uses a fully-connected layer to predict detections in a fixed set of regions. The Single Shot MultiBox Detector (SSD) [17] extends this single shot idea further by adding multiple scales of feature maps and using a convolutional filter at each scale for prediction.

Though the aforementioned methods have made great success in object detection recently, they still suffer from three typical limitations as listed.

First, the current methods using region proposals and single shot boxes train a monolithic CNN model to represent an object category. As object of the same category may experience very large variation due to different camera capturing viewpoints, object deformations and occlusions, a single CNN model may not have sufficient capacity to capture all these object appearance variations. Different non-CNN based methods have been proposed to address the large intra-class variation problem, which typically partition images into a number of more compact and homogeneous subcategories via maximum margin based support vector machines (SVM) [18], [19], graph shift [20], image incongruence [21], spectral clustering [22], etc. Promising results have been achieved in object detection [21], [22], [23] and object recognition [18], [19], [20] which validates the usefulness of the subcategorization approach.

Second, the selective search in R-CNN/Fast R-CNN [9] and default bounding boxes in SSD [17] both produce a large amount of false alarms which increase the computational cost greatly but lead to little performance gain in the ensuing CNN classification and regression. The major reason is that both selective search and default bounding box generation are unsupervised which involve little category-level supervised information. The Faster R-CNN [10] trains a region proposal network to generate object proposals which reduces false alarms and speeds up the detection process greatly. But the RPN under the framework of Faster R-CNN uses only one scale of feature map which may produce miss detections when the objects in images have very different sizes.

Third, current region proposal and single shot based detection methods train the object proposal generator (RPN in Faster R-CNN) and CNN classifier/regressor (R-CNN and SSD) in a single round. On the other hand, better object detection performance can be achieved through an iterative scheme, which continuously refines the determined proposal regions and the learned new CNN classifier/regressor until certain optimization target is met.

We propose a subcategory-aware deep convolutional network i. e. S-CNN to address the large object intra-class variation problem. To the best of our knowledge, this is the first work that integrates image subcategorization with CNN for accurate and robust object detection. Though a multiview CNN model has been developed in [12] which uses the object viewpoint to subcategorize images to train viewpoint-dependent CNNs, the model cannot be generalized to generic object detection where objects often suffer from various occlusions, clutters, deformations, etc. The proposed S-CNN has three major contributions as described below.

First, it designs an instance-sharing maximum margin clustering (MMC) algorithm and a subcategory-aware CNN that relieve the large intra-class variation in object detection effectively. The instance-sharing MMC allows one training sample to be shared by two neighboring subcategories, and accordingly helps to learn more robust and representative subcategory models. The subcategory-aware CNN employs a newly designed softmax objective

- *T. Chen is with Visual Computing Department, Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632. E-mail: ntuchentao@gmail.com.*
- *J. Fan is with the Satellite Department, Agency for Science, Technology and Research, Singapore 138632. E-mail: fanj@i2r.a-star.edu.sg.*
- *S. Lu is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798. E-mail: Shijian.Lu@ntu.edu.sg.*

Manuscript received 17 Aug. 2016; revised 30 May 2017; accepted 23 Sept. 2017. Date of publication 25 Sept. 2017; date of current version 12 Sept. 2018.
(Corresponding author: Tao Chen.)
Recommended for acceptance by T. Darrell.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2756936

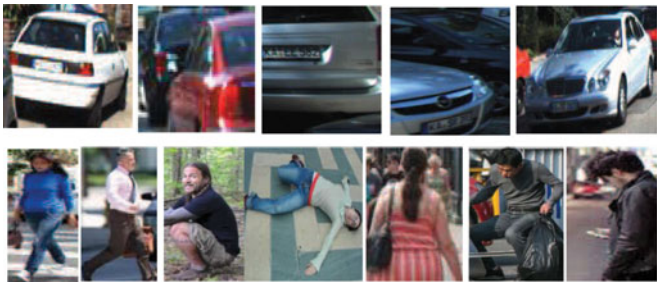

Fig. 1. Illustration of the large intra-class variation due to different viewpoints, object deformation, occlusion, clutters, etc.

function that treats all positive subcategories equally important and differentiates them from the negative category using different weights. The trained S-CNN has demonstrated better discriminative capability with superior object detection performance as evaluated over a number of public datasets.

Second, a multi-component ACF detector is designed to address the false alarm and inaccurate object localization issues. In particular, the multi-component ACF detector employs multiple component detectors each of which is trained using images within one clustered subcategory. It can therefore produce proposals with better localization accuracy and fewer false alarms, leading to better S-CNN classification and object detection performance. Compared with the RPN that uses one single scale of feature map, the multi-component ACF detector works on multi-scale feature pyramids and can detect objects with different sizes. Additionally, it can produce latent samples (The term of latent is used as the samples are produced by detector instead of human labelling) with different overlapping scores with the ground truth. These latent samples can be fused with other training data which greatly help for better subcategory clustering, ACF detector learning and CNN training.

Third, an iterative learning scheme is designed to optimize the learned multi-component ACF detector and the S-CNN continuously for better object detection performance. The more diversified latent samples as produced by the ACF detector enrich the training data greatly, which are feed-forwarded to the instance-sharing MMC clustering for generating new subcategories and further training more discriminative ACF detector and S-CNN iteratively. The learning iteration automatically terminates when the S-CNN training score converges.

The remaining of the paper is organized as follows. Section 2 describes the proposed S-CNN technique including the instance sharing clustering, the multi-component ACF detector, the subcategory-aware CNN training, and the iterative joint training of ACF and S-CNN. Section 3 presents experiments on three public datasets including the INRIA Person dataset, the Pascal VOC 2007 dataset, and the MS COCO dataset. Some concluding remarks are finally drawn in Section 4.

2 The Proposed Method

Fig. 2 shows an overview of the proposed S-CNN system. During the training stage, the training samples are first clustered into multiple subcategories by using the proposed instance sharing clustering algorithm. A multi-component ACF detector is then learned and further applied on the original training images in each subcategory to produce latent samples. A S-CNN model is then trained by combining the clustered training samples and the generated latent samples which also produces a detection score according to the ground-truth object boxes. The generated latent samples are further fed back to the instance-sharing MMC clustering for another round of ACF and S-CNN training until the S-CNN detection score converges. During the testing stage, a test image is first forwarded to the multi-component ACF detector to generate a preliminary set of proposals. The objects of interest are then detected by filtering out false alarms using the trained S-CNN.

2.1 Instance Sharing Maximum Margin Clustering

MMC [18], an extension of the supervised large margin theory (e.g., SVM) to the unsupervised scenario, aims to cluster the visually similar samples into the same category as much as possible. It optimizes the linear models learned for each category and simultaneously classifies each sample into its corresponding category, often leading to more compact clusters than other graph or probability based methods [20], [22].

The standard MMC clusters each sample image into a single cluster, which often introduces two typical problems while training an object detector by using each generated image cluster. The first is related to impaired robustness and representation capability of the trained detectors. In particular, detectors trained using images from neighboring clusters may all produce flat detection scores for ambiguous samples that lie around the boundary of the neighboring clusters. This will lead to miss detections when the maximum of the produced flat detection scores is lower than a predefined threshold. The second is related to clustering imbalance where some generated cluster may have a very small number of samples. Training a detector using such ultra-small cluster can easily lead to overfitting and further miss detections while applying it to new test images.

We propose an instance sharing MMC technique to address the two constraints. By allowing sample sharing among neighboring clusters, the trained detectors are more robust and representative which have better chance (as compared with detectors trained using non-sharing clusters) to produce high detection scores for those ambiguous samples lying around the boundary of the neighboring clusters. Meanwhile, the sample sharing addresses the cluster imbalance effectively where ultra-small clusters will get more training samples from the neighboring clusters. This relieves the overfitting of the trained detectors and leads to better detections when applying the trained model to new test images. Note similar idea has been explored in [24] where features of neighboring object viewpoints are shared for robust object pose estimation.

Suppose $\{\mathbf{x}_i\}_{i=1}^{M}$ denote a set of training samples from an object category, where M is the sample number. The instance sharing MMC tries to solve the following objective,

$$\min_{W,Y,\xi\geq 0} \quad \left\{\frac{1}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|^2 + \frac{C}{K}\sum_{i=1}^{M}\sum_{j=1}^{K}\xi_{ij}\right\} \tag{1}$$

$$\begin{aligned} s.t. \quad & \textstyle\sum_{k=1}^{K} y_{ik}\mathbf{w}_k^T\mathbf{x}_i - \mathbf{w}_j^T\mathbf{x}_i \geq 1 - y_{ij} - \xi_{ij}, && \forall i,j \\ & y_{ik}\in\{0,1\}, && \forall i,k \\ & 1 \leq \textstyle\sum_{k=1}^{K} y_{ik} \leq 2, && \forall i \\ & L \leq \textstyle\sum_{i=1}^{M} y_{ik} \leq U, && \forall k \end{aligned} \tag{2}$$

where $W=\{\mathbf{w}_k\}_{k=1}^{K}$ denote the optimal linear models of the K subcategories, $Y=\{y_{ik}\}, i=1,...,M, k=1,...,K$ denote the subcategory assignment of the M training samples, where $y_{ik}=1$ indicates that the i-th training sample is clustered into the k-th subcategory. The $\xi=\{\xi_{ij}\}, i=1,...,M, j=1,...,K$ denote the slack variables to allow soft margin, and C is the trade-off parameter. The first constraint in Eq. (2) thus enforces a large margin between subcategories by requiring the response score of $\mathbf{x}_i$ to the assigned subcategory to be sufficiently larger than that of $\mathbf{x}_i$ to other subcategories.

The second last constraint enforces each positive training sample to be assigned to either one or two subcategories. We tested different numbers of sharing subcategories, and found that the optimal performance is usually obtained when this number is set at two. The last constraint ensures that the clustered subcategory is

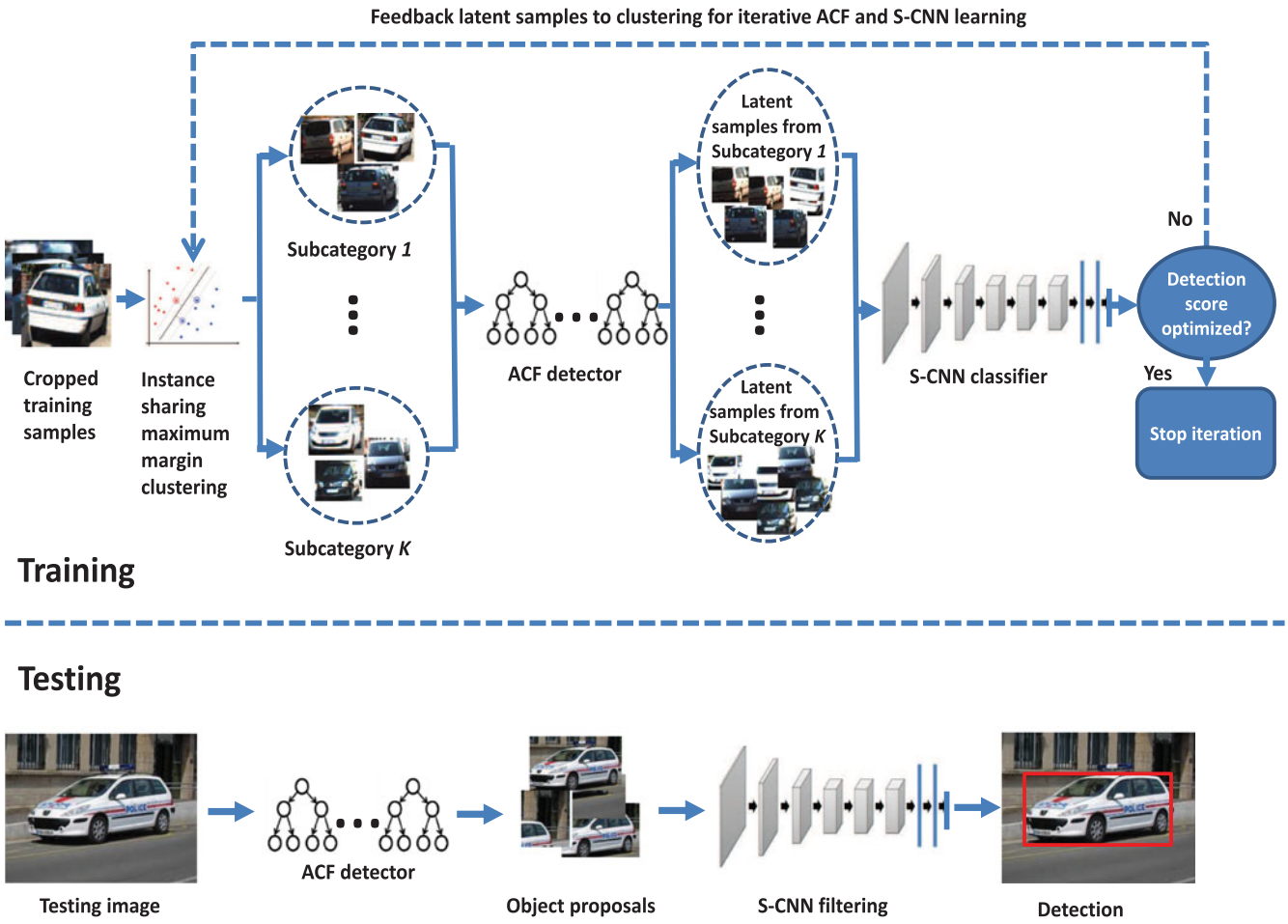

Fig. 2. Overview of the proposed subcategory-aware CNN learning for object detection.

balanced with a moderate sample size. The lower bound L and upper bound U of each subcategory size are set at $0.9\frac{M}{K}$ and $1.1\frac{M}{K}$ respectively, based on the experiments.

When new images are collected for clustering, the pre-trained linear models W can be directly applied to determine their subcategory assignments as follows,

$$k = \arg\max_{k=1,\ldots,K} \mathbf{w}_k^T \mathbf{x} \tag{3}$$

where $\mathbf{x}$ denotes the feature representation of a newly collected image.

2.2 Multi-Component ACF Detector Learning

The original ACF detector [1] resizes training images to a number of aspect ratios (from 0.3 to 3) and trains one ACF component detector by using the training images of one specific aspect ratio. In addition, the sample images of one specific aspect ratio are resized to different scales (0.5 to 2 of the original scale) to train one component detector. Further, 32 pyramid levels are implemented for each training image to learn the feature pyramid [25]. The training is a 4-round boosting process, where each round contains 32, 128, 512 and 2048 decision trees respectively. The boosting improves the discriminative capability of the learned decision trees greatly.

One major constraint of the original ACF detector is that it trains each component detector by resizing all training images to one specified aspect ratio. On the other hand, the resizing often introduces severe object distortion that affects the proposal performance [26]. We propose a multi-component ACF detector that first clusters all training images into a number of subcategories and then trains a component detector by using images within one clustered subcategory. Compared with the training images of one specific aspect ratio (as employed in the original ACF), the subcategory as produced by the proposed instance sharing clustering is more compact and representative which often leads to more discriminative and accurate component detector. Experiments show that the proposed multi-component ACF detector produces less false alarms and can localize object proposals more accurately as compared with the original ACF as well as those state-of-the-art proposal methods such as edge box and selective search (to be discussed in Section 3.3).

Three data sources are combined to enrich the training samples for the training of the multi-component ACF detector and S-CNN as illustrated in Fig. 3. The first is the ground truth boxes. Considering that some subcategories may contain a limited number of training samples, data augmentation by image translation, rotation and mirroring [27] is applied on the ground truth training samples to generate more training images. The second source is ACF detected latent

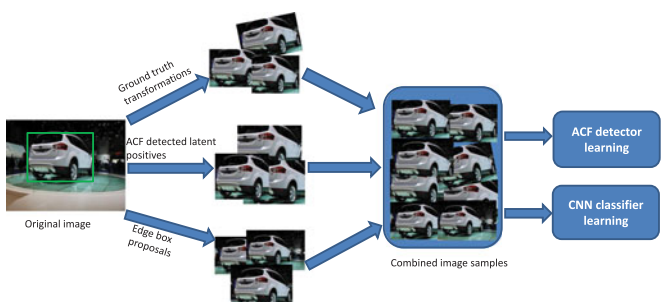

Fig. 3. Image sources for both ACF detector and CNN learning.

samples which are detected throughout the iterative training process. The latent training samples are selected according to the Intersection over Union (IoU) score [3] between the detected rectangles and the ground truth rectangles. In particular, a positive sample is determined if the IoU is above 0.8 and an negative sample is determined if the IoU is below 0.3. The last data source is edge box proposals which remain the same throughout the iterative training process. The edge box is used due to its high detection recall. Similar to the second data source, a positive sample is determined if the IoU is above 0.8 and a negative sample is determined if the IoU is below 0.3.

2.3 Subcategory-Aware CNN

The S-CNN is learned for each object category to determine whether a region proposal contains the specified object or not. The three data sources as described in last section constitute the training samples within each subcategory for the S-CNN learning. The randomly sampled negatives from the original images and negatives as detected by ACF detector and edge box constitute the negative category. It is therefore a typical $K+1$-way classification problem with K positive subcategories and 1 negative category. Compared with the 2-way binary CNN model that is trained by treating all positive samples as one single category, the $K+1$-way S-CNN model trained by clustering the positive samples into K multiple subcategories is more representative for the visually diversified positive samples and has better discriminative capability between the positive and negative classes.

We adopt the VGG-16 layer network structure [28] trained on the ImageNet dataset as the base network. The base network is fine tuned by using the subcategory images to learn the S-CNN, where the output of the last fully-connected layer is changed from 1000-way to a $K+1$-way softmax. All the positive and negative traning samples are resized to 224×224 pixels, mean-subtracted and divided by their standard deviation, and further propagated through VGG network layers to produce a probability distribution over $K+1$ classes. The class that produces the highest probability score is treated as the training sample's belonging (sub)category. The "Dropout" [29] strategy which sets the output of each hidden neuron to zero with probability of 0.5 is used in the first two fully-connected layers.

As this work focuses on object detection instead of object subcategorization, we propose a new softmax function to discriminate the K positive subcategories from the negative category. The idea is to assign equal weight to the K positive subcategories while different weight to the negative category as follows:

$$J = -\frac{1}{N}\left[\sum_{i=1}^{N}\Big(\mathbf{I}(y^{(i)}=1,2,...,K)\log p^{(i)} + \mathbf{I}(y^{(i)}=0)\log\left(1-(p^{(i)})\right)\Big)\right] \tag{4}$$

where J is the softmax loss, $\mathbf{I}()$ is the indicator function, $y^{(i)}=1,2,...,K$ indicates that the test sample $x_i, i=1,...,N$ is classified into one of the K positive subcategories, $y^{(i)}=0$ indicates that the test sample x_i is classified into the negative category, $\log p^{(i)}$ is the weight assigned to positive subcategories, and $p^{(i)}$ is the probability score of a test sample belonging to the defined object category which is defined as,

$$p^{(i)} = \max_{k=1,...,K} p_k^{(i)} \tag{5}$$

where

$$p_k^{(i)} = \frac{\exp(a_k^{(i)})}{\sum_{k=0}^{K}\exp(a_k^{(i)})} \tag{6}$$

where $a_k^{(i)}$ is the output of unit k in the CNN's last fully connected layer for the sample x_i. The decision score of test sample x_i is therefore determined by the maximum of the output of the K subcategory units in the CNN's last fully connected layer. The loss function in Eq. (4) is minimized by the stochastic gradient descent (SGD) [30] with a learning rate of 0.001, a batch size of 256 samples and a momentum of 0.9.

2.4 Joint Learning of Multi-Component ACF Detector and S-CNN

Motivated by the bootstrapping in ACF learning, we train the multi-component ACF detector and S-CNN in an iterative manner for better object detection performance. During each iteration, a number of image subcategories are first produced through the instance-sharing MMC clustering of the current training samples. A multi-component ACF detector is then trained based on the produced image subcategories, which is further applied on the original training images (where the training samples are cropped) to detect new training samples. These samples are finally incorporated to train a better S-CNN based on the proposed softmax function.

The iterative learning can therefore be viewed as a data augmentation approach which detects more training samples to train more representative and discriminative ACF detector and S-CNN model iteratively. In particular, the ACF detected training samples in each iteration are filtered based on their overlaps with the ground truth boxes (as described in Section 2.2). As the learning iteration moves on and more latent training samples are incorporated, the learned multi-component ACF detector becomes more accurate in object localization and false alarm suppression, and the trained S-CNN becomes more discriminative in object candidate classification. The learning iteration terminates automatically according to the convergence of the detection score of the trained S-CNN, e.g., the average precision [31] for the VOC2007 dataset or the FPPI [32] for the INRIA Person dataset.

During the testing stage, a test image is first forwarded to the multi-component ACF detector to produce a number of object candidates. The produced object candidates are then fed to the trained S-CNN to filter out false alarms. Non maximum suppression (NMS) is finally applied to remove those repeated detections with lower CNN scores ($p^{(i)}$), and keep only detections with the highest S-CNN scores.

3 Experiments

3.1 Datasets and Evaluation Criteria

INRIA Person Dataset. The INRIA Person dataset [33] is one most popular person dataset containing 1805 64×128 images of humans cropped from a varied set of personal photos. The persons appear in any orientation with partial occlusions and a wide range of variations in pose, appearance, clothing, illumination and background. It therefore has very large intra-class variations and is suitable to evaluate the proposed S-CNN method.

Pascal Visual Object Classes Challenge 2007. The famous Pascal Visual Object Classes Challenge 2007 (VOC2007) dataset consists of 9,963 images with 24,640 annotated objects from 20 object classes [31]. We follow the approach in [8] to split it into a training set and a testing set, which are used to train and evaluate classification models, respectively..

Microsoft COCO Detection Dataset. The Microsoft COCO object detection dataset [34] contains 80 object categories. We follow [10] to use 80k images for training, 40k images for validation, and 20k images for testing.

Evaluation Criteria. For the INRIA dataset, we follow the evaluation criteria in [32] and use the log-average miss rate (MR) between 10^2 and 10^0 false positives per image (FPPI) for evaluation. For the VOC2007 dataset, we follow the [8] and use the mean average precision (mAP) as the evaluation criterion. For both datasets, a detection is treated as a true positive only if the IoU between the detection box and the groundtruth box is greater than 0.5. For the

TABLE 1
Comparison of Different Techniques on the INRIA Person Dataset

Methods	Log-average miss rate (%)
Fast ACF [1]	17
VeryFast [35]	16
Subcategory + ACF [22]	14
R-CNN [8]	8
S-CNN_sm	6
S-CNN	4

MS COCO dataset, the mAP averaged for IoU $\in [0.5 : 0.05 : 0.95]$ and denoted as mAP@[0.5, 0.95], and mAP@0.5 (PASCAL VOC's metric) are reported.

The system run on a workstation with Intel core i7-5960X CPU 3.00 GHz, NVIDIA GTX-Titan GPU, and 64 GB RAM.

3.2 Experimental Results

3.2.1 Results on the INRIA Person Dataset

For the INRIA person dataset, we compare the proposed S-CNN with several state-of-the-arts including the original fast ACF detector [1], the very fast person detector using geometric context [35], the subcategory combined with ACF detector method in [22] and the typical R-CNN in [8]. In addition, we compare the new softmax function as defined in Eq. (4) with the conventional softmax function (S-CNN_sm) [7] under the same S-CNN framework. The number of subcategories K is set as 6 and the iteration number (for the learning of subcategories, multi-component ACF detector, and S-CNN) is set at 400 based on experiments.

Table 1 shows experimental results. It can be seen that the proposed S-CNN achieves the lowest miss rate among all compared methods. In particular, it outperforms the subcategory-aware ACF detector and the standard R-CNN by 10% and 4%, respectively. This shows the superiority of the proposed S-CNN that jointly learns subcategories, multi-component ACF detector and S-CNN classifier in an end-to-end and iterative manner. Further, it can be seen that the S-CNN using the new softmax function achieves lower miss rate as compared with that using the conventional softmax function. This demonstrates the better discriminative capability of the proposed softmax function which treats all the positive subcategories as equally important and differentiates them from the negative category with different weights as defined in Eq. (4).

3.2.2 Results on the VoC2007 Dataset

For the VOC2007 dataset, the S-CNN is compared with several state-of-the-art methods including the standard DPM [3], selective search proposal with bag-of-words classifier [15], RegionLet [36], R-CNN [8], Fast R-CNN [9], Faster R-CNN [10], single shot multibox detection with 300×300 input image size (SSD300) and 512×512 input size (SSD512) [16]. The YOLO method [17] is not compared here as its performance is below that of Fast/Faster R-CNN and SSD on the VOC2007 datset. Similar to the INRIA dataset, the new softmax function is compared with the conventional softmax function (S-CNN_sm in Table 2) under the same S-CNN framework. The subcategory number is set at 6.

Table 2 shows experimental results. It can be seen that the proposed S-CNN outperforms the state-of-the-art methods including Fast/Faster R-CNN and SSD and achieves the best mAP score of 72.4 percent. In addition, the S-CNN using the new softmax function outperforms that using the conventional softmax function, and this validates the effectiveness of the newly designed softmax function. Further, it can also be seen that the S-CNN achieves better performance improvement for those object categories with larger intra-class variation such as car, cat, cow, dog, person and sheep. This further validates the effectiveness of the proposed S-CNN in dealing with large intra-class variations.

TABLE 3
Object Detection Results (%) on the MS COCO Dataset

Methods	mAP@0.5	mAP@[0.5, 0.95]
Fast R-CNN [9]	35.9	19.7
Faster R-CNN [10]	42.7	21.9
SSD512 [16]	46.5	26.8
S-CNN	49.5	29.6

3.2.3 Results on the MS-COCO Dataset

For the MS-COCO dataset, we compare the S-CNN with state-of-the-art methods including Fast R-CNN, Faster R-CNN and SSD512 [16]. The S-CNN similarly uses 6 subcategories as used for the VOC2007 dataset. Table 3 shows experimental results. It can be seen that the proposed S-CNN achieves a mAP@0.5 of 49.5 percent and a mAP@[0.5,0.95] of 29.6 percent, which are 3 and 2.8 percent higher than SSD512, respectively. The 3 percent performance improvement on the MS COCO dataset is much larger than the 0.8% improvement on the VOC2007 dataset. This can be explained by the fact that objects in the MS-COCO dataset have smaller sizes, and SSD may miss small objects due to the lower resolution of its top layer. The proposed S-CNN can relieve this problem by using the multi-scale feature pyramids and multi-scale scanning windows in the ACF detector.

3.3 Discussion

We investigate the contribution of the three S-CNN components including the instance-sharing subcategory generation, the multi-component ACF detector, and the iterative S-CNN training. The INRIA and PASCAL VOC datasets are selected for experiments.

3.3.1 Subcategory Generation

We compare the proposed instance sharing MMC with several subcategory generation methods including the original MMC without instance sharing [18], the image incongruence [21], graph shift [20],

TABLE 2
Average Precision (%) of Different Object Detection Methods on the VoC2007 Dataset

	plane	bicycle	bird	board	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mean
SS-BoW [15]	43.5	46.5	10.4	12.0	9.3	49.4	53.7	39.4	12.5	36.9	42.2	26.4	47.0	52.4	23.5	12.1	29.9	36.3	42.2	48.8	33.8
DPM_v5 [3]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
RegionLet [36]	54.2	52.0	20.3	24.0	20.1	55.5	68.7	42.6	19.2	44.2	49.1	26.6	57.0	54.5	43.4	16.4	36.6	37.7	59.4	52.3	41.7
R-CNN [8]	73.4	77.0	63.4	45.4	44.6	75.1	78.1	79.8	40.5	73.7	62.2	79.4	78.1	73.1	64.2	35.6	66.8	67.2	70.4	71.1	66.0
Fast [9]	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7	67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8	66.9
Faster [10]	70.0	80.6	70.1	57.3	49.9	78.2	80.4	82.0	52.2	75.3	67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	81.1	67.6	69.9
SSD300 [16]	73.4	77.5	64.1	59.0	38.9	75.2	80.8	78.5	46.0	67.8	69.2	76.6	82.1	77.0	72.5	41.2	64.2	69.1	78.0	68.5	68.0
SSD512 [16]	75.1	81.4	69.8	60.8	46.3	82.6	84.7	84.1	48.5	75.0	67.4	82.3	83.9	79.4	76.6	44.9	69.9	69.1	78.1	71.8	71.6
S-CNN_sm	76.1	72.5	64.1	50.2	43.8	76.6	84.9	85.2	42.8	80.6	63.2	83.6	82.6	81.2	79.6	37.1	77.2	64.9	73.2	75.6	69.8
S-CNN	78.4	75.5	71.2	54.3	45.2	80.1	88.2	87.2	47.6	83.6	66.1	84.3	83.4	83.2	81.2	38.2	80.2	67.3	75.1	77.6	72.4

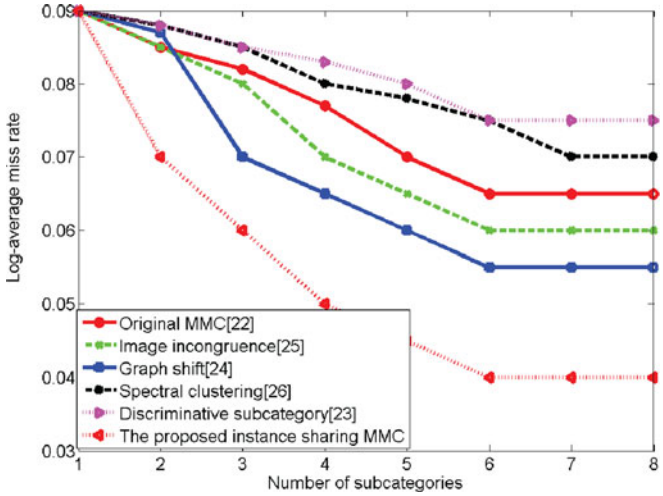

Fig. 4. Experimental results of different clustering methods under different subcategory numbers on the INRIA dataset.

spectral clutering [22] and discriminative subcategory [19] and evaluate them using different numbers of subcategories.

Figs. 4 and 5 show experimental results on the INRIA dataset and the VOC2007 dataset, respectively. As the two figures show, the proposed instance-sharing MMC consistently achieves the best performance on both datasets when different numbers of subcategories are implemented. In addition, all methods perform better when more than one subcategory is used, and the best improvement is up to 5 percent as compared with no subcategorization (i.e., when $K = 1$ as shown in Figs. 4 and 5). This demonstrates the effectiveness of subcategorization which relieves the intra-class variation and improves the object detection performance consistently. Further, all compared methods converge when the subcategory number goes beyond 6 and we therefore set it at 6 as described in Section 3.2. Note that the optimal subcategory number varies with the intra-class variation of different object classes and a larger number of subcategories is typically needed when the object class has larger intra-class variation.

3.3.2 Multi-Component ACF Detector Training

We evaluate the contribution of different object proposal methods for the multi-component ACF detector training. Different combinations of data sources are studied including the ground truth boxes (GT), the ground truth boxes plus their transformation (GTT), the edge box proposal (Edge_BP), and the latent samples as detected by the ACF detector (ACF_DP).

TABLE 4
Investigation of Using Different Data Sources for Multi-Component ACF Detector Learning. GT: Ground Truth Rectangles; GTT: Ground Truth Transformations; ACF_DP: Latent Positives as Detected by ACF Detectors; Edge_BP: Edge Box Detected Object Proposals

Methods	Log-average miss rate (%) on INRIA dataset	mAP on Pascal VoC 2007 dataset
GT + Edge_BP	9.0	65.2
GTT + Edge_BP	7.5	66.3
GT + ACF_DP	7.0	67.6
GTT + ACF_DP	6.5	69.1
GT + ACF_DP + Edge_BP	4.5	70.6
GTT + ACF_DP + Edge_BP	4.0	72.4

Table 4 shows experimental results. It can be seen that using ACF proposals consistently outperforms using edge box proposals and this demonstrates the power of using ACF detector for object proposal. In addition, the including of edge box proposals helps to improve the object detection performance clearly, largely due to its high proposal recall. Further, the model achieves the best performance when all proposals including GTT, Edge_BP, and ACF_DP are used. This shows that these data sources are complementary and their combination produces more representative object proposals which further enhance the representation and discrimination capability of the trained S-CNN.

3.3.3 S-CNN Learning

We also study the the joint iterative learning of ACF detector and S-CNN. Fig. 6 shows the miss rates under different numbers of learning iterations with/without the dropout layer on the INRIA dataset. It can be seen that the S-CNN keeps performing better with the increase of learning iterations with or without the dropout. This clearly shows that the proposed iterative learning approach can improve both the generated subcategories and the trained S-CNN. In addition, it can be seen that using dropout helps to improve the object detection performance greatly, and it also requires a larger number of learning iterations to converge. This validates the effectiveness of dropout for relieving the S-CNN overfitting while having limited training data. Furthermore, it is found that the learning iteration convergence number is different for different object categories and it typically ranges from 100 to 300 iterations for the Pascal VoC dataset.

3.3.4 Timing Analysis

For the INRIA dataset, it is found that the multi-component ACF detector takes around 0.1 seconds on average to extract 50-100

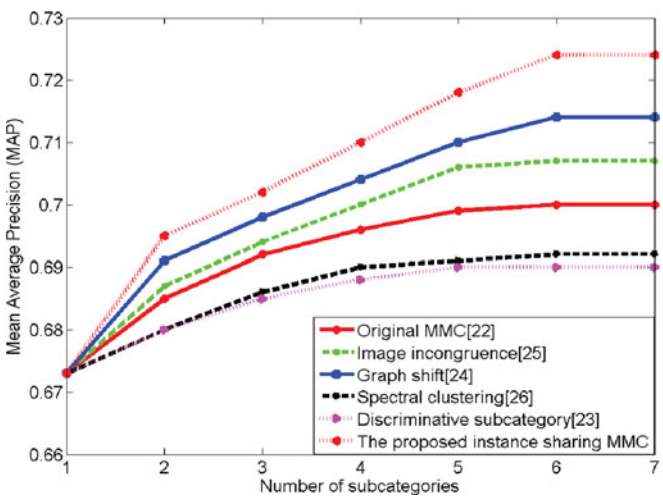

Fig. 5. Experimental results of different clustering methods under different subcategory numbers on the Pascal VoC 2007 dataset.

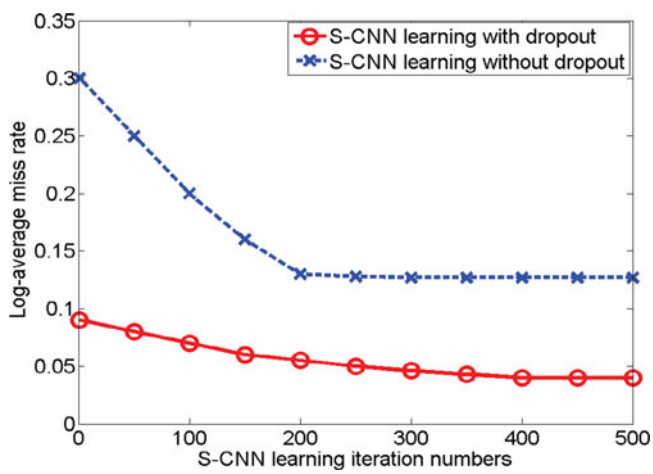

Fig. 6. S-CNN person detection performance for the INRIA person dataset under different numbers of learning iterations

proposals from each image. The S-CNN takes another 0.5 to 1 seconds to extract the convolutional features from the detected proposals. The S-CNN therefore needs around 0.6-1.1 seconds for processing of one image. For the VOC2007 dataset, the computation time is around 0.5 to 1 seconds on average which is similar to that for the INRIA dataset. For the MS COCO dataset, the S-CNN needs around 1 seconds on average.

4 Conclusions

This paper presents a subcategory-aware CNN technique for object detection. A novel instance-sharing MMC algorithm is designed which clusters sample images into a number of subcategories to relieve the large intra-class variation issue. It allows the neighboring subcategories to share samples and accordingly improves the robustness and representation capability of the multi-component ACF detector. In addition, a S-CNN training method is designed which employs a new loss function to capture the multiple subcategories information and helps to improves the object detection performance. Furthermore, an iterative learning scheme is developed which repeats the instance-sharing MMC, the multi-component ACF detector learning and the subcategory-aware CNN training until the object detection score converges. The iteration improves the multi-component ACF detector and CNN iteratively by including more useful latent training samples that are detected in each training iteration. The proposed S-CNN has been evaluated over three public datasets including the INRIA person dataset, the VOC2007 dataset, and the MS COCO dataset. Experiments demonstrate superior object detection performance of the S-CNN as compared with state-of-the-art techniques such as Fast/Faster R-CNN and SSD.

References

[1] P. Dollr, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[2] T. Chen and S. Lu, "Accurate and efficient traffic sign detection using discriminative AdaBoost and support vector regression," *IEEE Trans. Veh. Technol.*, vol. 65, no. 6, pp. 4006–4015, 2016.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[4] L. Zhu, Y. Chen, A. Yuille, and W. Freeman, "Latent hierarchical structural learning for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1062–1069, 2010.

[5] T. Chen and S. Lu, "Robust vehicle detection and viewpoint estimation with soft discriminative mixture model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 2, pp. 394–403, Feb. 2017.

[6] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.

[7] J. Yan, Y. Yu, X. Zhu, Z. Lei, and S. Z. Li, "Object detection by labeling superpixels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5107–5116.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[11] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1134–1142.

[12] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis,.*, 2015, pp. 945–953.

[13] M. Everingham, L. Van Gool , C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[14] O. Russakovsky, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[15] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[16] W. Liu, et al., "SSD: Single shot multibox detector," in *Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[18] G.-T. Zhou, T. Lan, A. Vahdat, and G. Mori, "Latent maximum margin clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 28–36.

[19] M. Hoai and A. Zisserman, "Discriminative sub-categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1666–1673.

[20] J. Dong, Q. Chen, J. Feng, K. Jia, Z. Huang, and S. Yan, "Looking inside category: Subcategory-aware object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1322–1334, Aug. 2015.

[21] D. Coppi, T. de Campos, F. Yan, J. Kittler, and R. Cucchiara, "On detection of novel categories and subcategories of images using incongruence," in *Proc. Int. Conf. Multimedia Retrieval*, 2014, pp. 337–344.

[22] E. Ohn-Bar and M. M. Trivedi, "Learning to detect vehicles by clustering appearance patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2511–2521, Oct. 2015.

[23] T. Lan, M. Raptis, L. Sigal, and G. Mori, "From subcategories to visual composites: A multi-level framework for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 369–376.

[24] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879–2886.

[25] T. Chen and S. Lu, "Accurate and efficient traffic sign detection using discriminative AdaBoost and support vector regression," *IEEE Trans. Veh. Technol.*, vol. 65, no. 6, pp. 4006–4015, Jun. 2016.

[26] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin, "A shape-preserving approach to image resizing," *Comput. Graph. Forum*, vol. 28, no. 7, pp. 1897–1906, 2009.

[27] B. Sapp, A. Saxena, and A. Y. Ng, "A fast data collection and augmentation procedure for object recognition," in *Proc. 23rd National Conf. Artif. Intell.*, 2008, pp. 1402–1408.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[29] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, pp. 1–18, 2012.

[30] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1395–1403.

[31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[32] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893. [Online]. Available: http://lear.inrialpes.fr/pubs/2005/DT05

[34] T.-Y. Lin, et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[35] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool, "Pedestrian detection at 100 frames per second," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, Jun. 2012, pp. 2903–2910.

[36] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2071–2084, Oct. 2015.