Enhanced Pedestrian Detection using Deep Learning based Semantic Image Segmentation

Tianrui Liu and Tania Stathaki
Department of Electrical and Electronic Engineering
Imperial College London, United Kingdom
t.liu15, t.stathaki@imperial.ac.uk

Abstract—Pedestrian detection and semantic segmentation are highly correlated tasks which can be jointly used for better performance. In this paper, we propose a pedestrian detection method making use of semantic labeling to improve pedestrian detection results. A deep learning based semantic segmentation method is used to pixel-wise label images into 11 common classes. Semantic segmentation results which encodes high-level image representation are used as additional feature channels to be integrated with the low-level HOG+LUV features. Some false positives, such as falsely detected pedestrians located on a tree, can be easier eliminated by making use of the semantic cues. Boosted forest is used for training the integrated feature channels in a cascaded manner for hard negatives mining. Experiments on the Caltech-USA pedestrian dataset show improvements on detection accuracy by using the additional semantic cues.

Index Terms—Pedestrian detection, semantic segmentation, pixel-wise image labeling, filtered feature channels.

I. INTRODUCTION

Pedestrian detection is of great research interest in computer vision. It has many significant applications including video surveillance, robotics automotive, and intelligence transportation. A substantial number of methods have been developed in order to improve detection accuracy [1], [2], [3], [4], [5], [6]. HOG [1] based detectors using the multi-scale sliding window mechanism have long been the dominant approaches for pedestrian detection. While no single hand-craft feature has been shown to outperform HOG, the combinations of HOG with other features have made improvements by making use of complementary visual cues. A texture descriptor based on local binary patterns (LBP) [7] was combined with HOG in [8] to cope with partial occlusions. HOG descriptors are used together with LUV color features in the form of image channels features (ICF) in [9]. The ICF detector outperforms HOG at a faster computational speed by computing integral images over feature channels. Aggregated channel features (ACF) [10] is a generalization of ICF which approximates multiscale scale gradients using nearby scales. In such way, the ACF [10] detector can achieve very fast feature pyramidal for real-time multi-scale detection on CPU along.

Pedestrian detection is often challenged by greatly variations in human pose and appearance. It is generally difficult to reduce false positives on hard negative samples such as tree leaves, traffic lights, poles, etc. Some of these hard negatives can be removed by making use of higher level vision cues, i.e. semantic information. For instance, it is very unlikely that a pedestrian is located on a tree. Therefore, it would be easier to eliminate the falsely detected pedestrian located on a tree, given the semantic information that there is a region of tree leaves within and/or around the detection window. This indicates that good pedestrian detectors need to be extended with a better semantic understanding of images. Semantic segmentation methods [11], [12], [13], [14] aiming at classifying image pixels into semantic classes such as sky, building, road, vehicle, etc. Among these classes, backgrounds such as sky, road, buildings provide semantic context that can be used for search space for pedestrian detection, while foreground classes like pedestrian and cyclists can be served as an alternative principle for pedestrian detection.

Object detection and semantic segmentation are two strongly correlated and complementary tasks for better performance [15], [16]. Dai et al. [16] used segments extracted from each object detection hypothesis for better localization. Most recently, Costea et al. [15] proposed to make use of semantic classification cost for pedestrian detection. In their work, decision trees are applied to classify each individual pixel into one of eight semantic classes. However, pixel-wise classification results using decision trees can be noisy and inconsistent (see Fig. 2 (b)). To tackle this problem, they relied on a Conditional Random Field (CRF) inference procedure to improve the segmentation results. Dense CRFs [17] defined over uniform 2-dimensional grids are applied over each pixel for 3 rounds in order to achieve a more consistent semantic segmentation result (see Fig. 2 (c)).

In this paper, we propose a pedestrian detection method making use of semantic context to improve detection results. The semantic channel for pedestrians provides an additional cue for their presence, so that successful detection and segmentation require implicitly agreement of both detection and segmentation predictions. Instead of obtaining semantic segmentation with decision trees and dense CRFs as in [15], we use a deep learning based semantic segmentation method which can achieve superb labeling accuracy in a single round. 11 common semantic channels e.g. sky, building, road, tree, vehicle, pedestrian etc., are used as context information to assist pedestrian detection. As many top performances pedestrian detectors [18], [3], [19] are based on the same 10 LUV+HOG channels used by the aggregated channel features (ACF) [10] detector , we also take the ACF detector as our based line. In addition, we make use of the results from
semantic segmentation to provide higher level representations. The semantic information is integrated together with the LUV+HOG channels and they are all trained jointly using boosted forest to distinguish a pedestrian from its backgrounds. In order to enhance the feature representation capability, we use intermediate filtering layer to filter the feature channels before boosted trained by RealBoost. The semantic feature channel can be regarded as a reinforcement of the filtered channel features for pedestrian detection. Experimental results showed improvements on detection rates using the additional semantic channels.

II. PROPOSED METHOD

Fig. 1 illustrates the motivation for proposing the pedestrian detector assisted by a deep learning based semantic image segmentation method. Given an image, instead of performing pedestrian detection directly as in the upper path, we first apply a deep learning method to semantically label the image into 11 classes. The semantic information is utilized as a semantic feature channel to be integrated together with the LUV+HOG channel features and they are trained altogether using boosted forest.

A. Semantic Image Labeling for Pedestrian Detection

In this part, a deep learning based semantic segmentation method is used to perform semantic image labeling for the proposed detector. Semantic image labeling aims to assign every pixel of an image with an object class label, challengingly combining image segmentation and object recognition in a single process. In semantic labeling problems, there are typically too many categories of objects in an image for traditional machine learning methods to handle at once. Deep learning has recently been investigated for the semantic segmentation task [11], [20], [12] owing to its strong capability of feature representation for multiple classes classification.

SegNet [12] is a practical deep convolutional neural network architecture for semantic pixel-wise image labeling. The architecture of SegNet can be seen as a convolutional encoder-decoder framework followed by a pixel-wise classifier. The encoder aims to generate feature maps and the decoder is used to upsample the feature maps output from the encoder before fed into the softmax classifier for pixel-wise classification. As many other advanced deep architectures for semantic segmentation [11], [20], the encoder network of SegNet is identical to the first 13 convolutional layers of the VGG16 network [21]. The encoder network consists of convolutional layers associated with batch normalization and ReLU non-linearity procedures. Some of the convolutional layers are followed by a max-pooling layer. The encoder network has strong feature representation capability thanks to the deep framework with consequence non-linear processing layers. However, output feature maps throughout 5 max-pooling layers of the encoder network are of very low resolution resulting. Take an input image of size $480 \times 360$, for example, SegNet encoder will shrink the image size by $2^5$ so that the output feature maps are of size $15 \times 6$. Therefore, upsampling is required to map the low resolution feature maps into original image size for pixel-wise semantic labeling. The decoder network of SegNet is topologically axisymmetric to the encoder network. It acts as a “deconvolutioner” in order to upsample the lower resolution feature maps. In SegNet, the decoder network is proposed to use pooling indices (i.e. indices of the pixels retained during max-pooling) of the corresponding encoder to perform upsampling. Since the feature maps obtained from this non-linear upsampling process are sparse, these feature maps are then convolved with trainable filters to produce dense feature maps.

The SegNet architecture is trained using Caffe-SegNet [22], [12] on a large database combining a set of urban traffic images [23] as in [12]. 11 common semantic classes are used for training: building, tree, sky, car, sidewalk, column pole, sign-symbol, road, fence, pedestrian and bicyclist; while pixels of other classes are labeled in black and will be ignored during training. Network weights of the 13 layers encoder are pre-trained on ImageNet [24]. An example of semantic labeling result obtained by SegNet is given in Fig. 2 (d). Compared to (b)-(c) which are semantic labeling results in [15], SegNet provides more clear cut labeling results and removes some of the ambiguities between buildings and cars.
in the example image. The reported performance in terms of average class classification accuracy on the CamVid test is 55.5% [15] versus 71.2% [12]. The detection time for semantic segmentation using SegNet is ~1ms using single GPU. The pixel-wise semantic segmentation results is encoded as a semantic channel containing the class index of every image pixel ranging from 0 to 11 (0 for unlabelled pixels). The semantic index map is using as an additional feature channel for the proposed pedestrian detector.

B. Filtered Channel Features based Pedestrian Detector using Semantic Image Segmentation

In this part, we integrated the semantic image segmentation result with our baseline detectors [10], [3] which make use of filtered channel features. In the ACF method, an input image is transformed into a set of feature channels, i.e. 10 HOG+LUV channels which contain 6 gradient orientations, 1 gradient magnitude, and 3 color channels. The feature channels are aggregated via sum-pooling before fed into a decision forest for feature selection. According to [3], using an intermediate layer to filter the low-level features before the boosting stage can provide better performing pedestrian detectors. We adopted this filtered channel features strategy to improve the feature capacity with an additional filter layer. There are filter bank of other format, such as decorrelated filters [19], Haar-like filters [18], and checkerboards [3] filters which can influence the detection quality by different amount. The filter bank used in [15] consist of only 3 filters for each scale, i.e. a simple pooling filter, a vertical filter and a horizontal filter. In our work, we apply the Checkerboards filters [3] (a set of filters that appear like checkerboard patterns, including uniform squares and horizontal/vertical gradient filters, see Fig. 3(d)), which can achieve top performance.

1) Integrating Semantic Feature for Pedestrian Detection:

The framework of a filtered channel features based detector is illustrated in Fig. 3. Given an image in Fig. 2(a), 10 HOG+LUV feature channels Fig. 2(c) are computed efficiently using integral images. Semantic features in Fig. 2(b) are obtained from the deep learning based semantic labeling method. In the experiment, the semantic channel is encoded as a channel containing the class index of every image pixel (range from 0 to 11), here Fig. 2(b) is shown in color for visualization. The integrated feature channels are filtered with the Checkerboards filter bank Fig. 2(d). Boosted Forest (BF) [25] is used to train the semantic features together with the HOG+LUV feature channels to distinguish pedestrian from background.

2) Boosted Forest for Integrated Feature Channels:

Boosted Forest is an ensemble learning method which can achieve fast and accurate classification by training accumulated hard negatives samples in cascade. Owing to high accuracy and low computation cost, decision forests have been widely used in computer vision tasks such as image classification [26], [27], object recognition [28], [29] and super-resolution [30], [31], [32]. During learning, each decision tree is responsible for finding the binary test parameter for feature selection and responsible for learning the thresholds in the split nodes. Generally, the split nodes in the trees are a simple comparison between a feature value and a learned threshold. A non-leaf node of decision tree will stop splitting and be declared as a leaf node if stop criterion have been met, i.e. typically when the maximum tree depth is reached; or the size of training data is too small. After training, each non-leaf node stores its binary test parameters and each leaf node stores the learned prediction models.

By applying filter banks for feature channel filtering, the features become simple pixel lookup and there is no need for integral images. The boosted forest classifier imposes no constraint on the dimensions of features. Hence, we directly append the obtained semantic channel feature to the 10 feature channels.
channels. The integrated features channels are used to train BF classifiers.

III. EXPERIMENTS

We train our pedestrian detector on the Caltech-USA dataset [33]. Under default settings, the training dataset contains 3880 frames selected from every 30th frame of the videos. Positive pedestrian samples are extracted and resized into windows of size (120, 180) in terms of window (width, height). There are 1586 positive pedestrian samples in the 3880 training images. In order to acquire more training data, each pedestrian window is mirrored along the horizontal direction to generate 3172 positive samples in total.

The work is implemented based on the publicly available toolbox [34]. To train the decision forests, we use 32, 512, 1024, 2048 and 4096 weak classifiers respectively for 5 bootstrapping rounds as in [3]. Initially, the training set consists of all positive examples and 10000 negative samples. In the first training stage, negative training samples are randomly generated, avoiding the regions containing pedestrians. For the other 4 stages, hard negative samples are selected using the detector trained from the previous stage. The number of hard negative samples to be added after each bootstrapping round are limited to 10000. The BF classifier obtained at the final stage is used for testing.

For evaluation, the “reasonable” testing set of Caltech-USA [33] is used. The testing set contains 4024 frames and pedestrians above 50 in height with no occlusion or only partial occlusions are used for evaluation. Following the evaluation method given in [33], measuring the log average miss rate ranging from $10^{-2}$ to $10^{0}$ false positive per image (FFPI). Performance of the proposed pedestrian detector is compared to VJ [35], HOG [1], HogLbp [8], LatSVM-V2 [2], FPDW [36], ICF [9], and ACF [10] in Fig. 4. As we can see, there is a big improvement between the [10] and [3]. This improvement reveals the effective on the additional filtering layer using the Checkerboards filter bank. Overall, the proposed detection method outperforms ACF [10] by about 30%, and outperforms the state-of-the-art work in [3] by 1.4%. The results reveal that it is beneficial to improve the detection accuracy by using additional semantic cues. In order to to see whether the proposed detector is hunger of data using the original Caltech dataset, we also generate the Caltech 10x dataset by saving the image frames every 3th frame from the Caltech videos. While the original Caltech-USA training dataset contains 3172 positive samples, the number of positive samples Caltech10x is increased to 32752. Detection result on the proposed pedestrian detector trained using Caltech10x set is given in Fig. 4(b). By comparing the performance in Fig. 4 (a), it is clearly demonstrated that the proposed detector trained using Caltech10x has been improved by around 4% by using more positive samples.

IV. CONCLUSIONS

In this paper, we proposed a pedestrian detector which makes use of semantic image segmentation results in order to improve the filtered channel feature based detector. We use a deep learning method to pixel-wise label images into 11 semantic classes. The semantic image segmentation results are used jointly with the HOG+LUV features. The proposed approach relies on the learning mechanism to assess the semantic content cues for a more powerful pedestrian detector. Experiments on the Caltech-USA dataset indicate that the proposed detector make improvement on the baseline detector by enforcing the consistency between detection and segmentation. In the future, we consider a fast implementation of the proposed detector which make use of GPUs.

ACKNOWLEDGMENT

This work was supported by the EU H2020 TERPSICHORE project “Transforming Intangible Folkloric Performing Arts into Tangible Choreographic Digital Objects” under the grant agreement 691218.
REFERENCES


