VEHICLE DETECTION AND POSE ESTIMATION BY PROBABILISTIC REPRESENTATION

Yao Xue[1], *Xueming Qian*[2], *IEEE Member*

[1]University of Alberta, [2]Xi'an Jiaotong University
Email: yxue2@ualberta.ca, qianxm@mail.xjtu.edu.cn

ABSTRACT

The ability to recognize a vehicle and estimate its pose is important for to a wide range of practical scenarios. Recently, it has been shown that several vehicle components could be a discriminative visual pattern during recognition and detection. To detect these components, our observation is that it is more reliable to build probabilistic prediction maps rather than simply relying on region proposals. Another key insight here is that the spatial relationship between these vehicle components shows some consistent patterns, and is able to provide strong semantic cues for vehicle pose estimation. In this paper, we propose a supervised learning framework with Fully Convolutional Network (FCN) to build probabilistic prediction maps of vehicle components, then vehicle pose is estimated based on the absence/presence and the spatial relationship (e.g. the relative distance, scales, angles) of these components. The proposed method is evaluated with three state-of-the-art approaches on public vehicle image datasets and achieve superior performance.

Index Terms— Vehicle pose estimation, Vehicle detection, Fully convolutional network, Probabilistic representation.

1. INTRODUCTION

Vehicle detection and pose estimation has been applied to a wide range of practical scenarios like traffic video surveillance, public safety and even behavior analysis. This is primarily the result of an increasing demand for an automated understanding of the pose or viewpoint of vehicles.

Previously, one mainstream direction is to apply the Part-based model [1]. However, it is well known that the problem of detecting and localizing individual object instance is still far from being solved, especially in real world application like vehicle recognition and detection, where the variation in visual appearance caused by the diversity in vehicle type, illumination, pose, and occlusion, etc could be quite huge. Therefore, it is better to build probabilistic prediction maps as the representation of vehicle components.

Following the big success of deep learning in computer vision tasks, R-CNN [2] and its variants (Fast R-CNN [3], Faster R-CNN [4]) obtain remarkable success in object detection. Most of the existing literature treats object detection and pose estimation as two separate problems. However, our observation here is that the information given by object detection could be used to refine pose estimation result.

Object Probabilistic Detection

In recent years, the state-of-the-art object detectors follow the framework of R-CNN, which takes an ImageNet pretrained Convolutional Neural Network (CNN) [5] and fine-tunes the network on PASCAL VOC detection data. At test time, the principle of R-CNN is remarkably simple: it samples image regions using a proposal mechanism such as Selective Search and classifies those regions as foreground or background. After that, Fast R-CNN and Faster R-CNN are proposed as an improvement over R-CNN, especially in terms of speed-up and End-to-End trainability. However, all these R-CNN based methods leave several deficiencies. The first question is how to enable a CNN to preserve sufficient geometric information to localize objects considering the presence of fully connected layers. The second question is whether it is reliable enough to localize each object using region proposal, which outputs a bounding box rather than probabilistic representation. In this work, we choose the Fully Convolutional Network (FCN) [6] and make it as a pixel-wise probabilistic detector.

Originally, FCN is proposed for semantic segmentation, by replacing fully connected layer with convolution or deconvolution layer. This adaptation transforms the FCN into a deep filter that preserves spatial semantic information of input image along its whole architecture from the first layer to the last layer. Another most obvious feature of FCN is its ability to provide a 2-D pixel-to-pixel map. So far, FCN has also shown compelling quality and efficiency for density prediction. For example, [7] develops a FCN-based cell counting approach ,where their FCN is responsible for predicting the cell density map for an input cell microscope image. In general, FCN has been proven a natural fit as pixel-wise machine learning predictor.

Discriminative Components Prediction

In addition, It has been confirmed that several vehicle parts could be a perfectly discriminative visual pattern in vehicle recognition or detection problem. Similar to their finding, we also observe that several vehicle components are more discriminate, for examples: (1) wheel, (2) headlight, (3) side

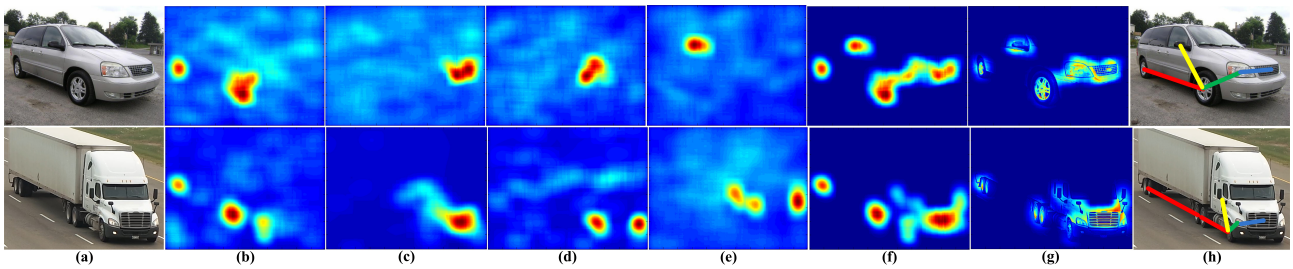

Fig. 1. original image (a) and its predicted component-level probabilistic score maps of "(b) wheel", "(c) headlight", "(d) side rear view mirror", "(e) front air-intake lattice", predicted by our Discriminative Components Fully Convolutional Network (DC-FCN); and (f) mixed score map after normalizing and threshold over the 4 score maps of discriminative components; (g) visualizes the real vehicle components based on (f); (h) captures the spatial relationship of different components, which provides strong semantic cues of vehicle pose.

rear view mirror, (4) front air-intake lattice. These components share some similar overall appearance, even considering the variation in vehicle appearance due to many reasons. Another key insight here is that the spatial relationship between these vehicle components share some consistent patterns, which carry important information of vehicle pose.

In this paper, we propose a vehicle detection and pose estimation method. In its first part, we build a FCN-based Discriminative Components prediction Model (DC-FCN), which is designed to generate a small set of instance-level pixel-wise prediction score maps, each of them measures the likelihood that a vehicle discriminative component exists in the corresponding location as shown in (b,c,d,e) of Fig.1. After that, an unified probability map is obtained after normalization and threshold over the set of score maps, as shown in (f) of Fig.1. And (g) of Fig.1 visualizes what exactly the probability map (f) finds over the real vehicle image. Finally, the spatial relationship (e.g. the relative distance, scales, angles, etc) of these components provides us strong contextual cues for final vehicle pose estimation.

2. THE PROPOSED FRAMEWORK

2.1. System Overview

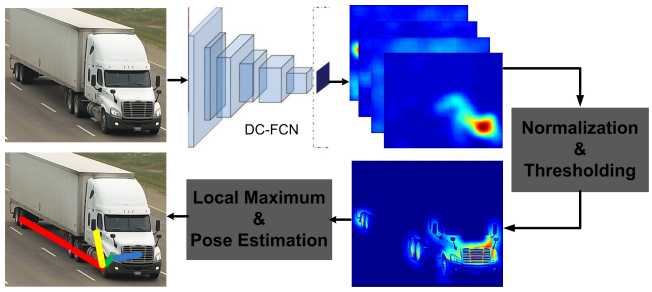

Fig. 2. System overview of the proposed method

Fig.2 summarizes the pipeline of the proposed method. To detect four discriminate vehicle components, a FCN-based Discriminative Components prediction Model (DC-FCN) is built to generate a small set of instance-level pixel-wise prediction score maps. Each of them represents the probability that a vehicle component exists in the corresponding location. Then normalization and thresholding are performed over the set of score maps to get global representation of activation map. After that, local maximum over the activation map is responsible for finding the center location of four discriminative components. Finally, vehicle pose is estimated based on the spatial relationship (e.g. the relative distance, scales, angles) of these component centers.

2.2. DC-FCN Model

Typical recognition nets (including LeNet, AlexNet, ResNet, etc) take fixed-sized inputs and produce non-spatial outputs. As we know, traditional CNN contains the Convolutional (Conv), Max-pooling (M), and Fully-Connected (FC) layers, etc. Both Conv and M layers are translation invariant and can be operated on input of arbitrary size. FC infers the classification scores for an input image with fixed dimensions, but discard spatial coordinates. Furthermore, the introduction of FC layers requires the input with a fixed size, as shown following:

$$h_j^l = \sigma\left(W_j^l h^{l-1} + b_j^l\right) \tag{1}$$

where W_j^l is the weight matrix connecting the neurons h^{l-1} in $(l-1)$-th FC layer and j-th index neuron h_j^l in the l-th FC layer, b_j^l is the bias and $\sigma(\cdot)$ is the element-wise nonlinear activation function. In fact, the fully connected layers are equivalent to convolutional layers with kernel size 1×1:

$$h_j^l = \sigma\left(\sum_{m=1}^{M} W_{jm}^l \otimes h_m^{l-1} + b_j^l\right) \tag{2}$$

where $\otimes$ denotes the 2D spatial convolution, $W_{jm}^l \in R^{1\times 1}$ is the convolution kernel connected to j-th feature map h_j^l and

the m-th feature map in the previous layer h^{l-1}, and M is the total number of feature maps in h^{l-1}. By employing Eq.2, we can convert the fully connected layer into a fully convolutional layer. Once the filters have been learned, they can be applied to the input image of arbitrary size. Thus, each convolution operator works on local input regions, and the whole network FCN naturally operates on an input of any size, and produces an output of corresponding spatial dimensions. The spatial output maps of the fully convolutionalized model make it a natural tool for pixel-wise prediction.

In this paper, we propose the FCN-based Discriminative Components prediction Model (DC-FCN). To produce such a network, we take the AlexNet basic architecture (5 Conv layers + 3 FC lyaers), and we convert the model into the corresponding fully convolutional 32 stride network (FCN-32s) as presented by [6]. First, each input image is paded with 100 pixels before features are extracted. Next, each of the three fully connected layers are converted into convolutional layers, where layer 6 has 4096 convolutions with 6×6 sized kernels, layer 7 has 4096 convolutions with 1×1 sized kernels, and the final score layer has $K + 1$ convolutions with 1×1 sized kernels (where K is the number of categories, plus one for background). Furthermore, additional deconvolution and crop layers are added which bilinearly upsample the score map produced by the 8th layer (bilinear interpolation) and crops the pixel level score map to be the size of the input image. It means the final output of our network is a score per category per pixel. Actually, not only the probabilistic score maps corresponding to 4 discriminative vehicle components are predicted by DC-FCN, it is also observed that the feature maps from the middle layers of DC-FCN are able to provide rich information for recognition, Fig.3 visualizes some feature maps learned during the training process.

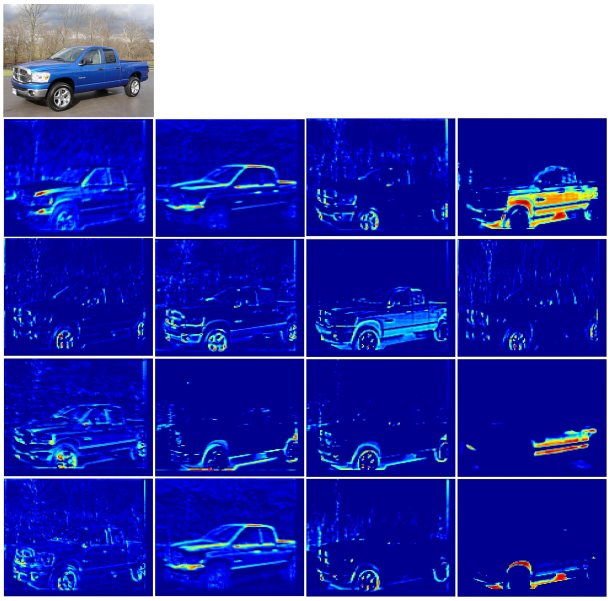

Fig. 3. Evolution of a randomly chosen subset of model features through training. Four colomuns corrrespond to the feature maps learned at training epoch 5, and feature maps from the same layer are shown in the same colomun. These activation maps provide rich information during recognition.

3. EXPERIMENTS AND PERFORMANCE

3.1. Datasets and Evaluation Criteria

In this paper, we use three public vehicle datasets: TU Graz [8], Stanford-cars [9] and EPFL [10], on which the proposed method and other comparison methods are evaluated. For the three datasets, we randomly select images for training and testing. For each training image, a class-sensitive segmentation mask is provided, which respectively labels the image regions of four discriminate vehicle components: (a) wheel, (b) headlight, (c) side rear view mirror, (d) front air-intake lattice, as shown in Fig.4.

3.2. Benchmark Evaluation

We perform two-level evaluations on vehicle detection results and pose estimation results, respectively.

Detection results evaluation

In order to test the detection performance, we use Intersection over Union (IoU) between detection and ground-truth as threshold. A predicted bounding box is considered correct if it overlaps more than an IoU threshold, otherwise overlapping detections are declared as false positives. Please note that precision depends on the threshold of IoU, which allows us to vary the overlap threshold between prediction and ground truth. The precision under different IoU threshold is summarized in Table.1. The major observation here is that the proposed method is able to continuously detect the 4 vehicle components, especially works better on Wheel and Air-intake classes, under the fact that we have tuned the criterion of true detection from loose (IoU=0.1) to harsh (IoU=0.7).

Table 1. The detection precision of our method. Each column in the table corresponds to the average precision for the four vehicle components.

	IoU=0.1	IoU=0.2	IoU=0.3	IoU=0.5	IoU=0.7
Wheel	0.765	0.718	0.694	0.640	0.618
Headlight	0.686	0.671	0.631	0.603	0.593
Mirror	0.653	0.634	0.624	0.602	0.594
Air-intake	0.755	0.737	0.702	0.654	0.604
mAP	0.7148	0.6900	0.6627	0.6247	0.5022

Pose estimation results evaluation

In order to evaluate the results of pose estimation and compare with others, we adopt a widely accepted [11] [12] pose estimation evaluation metric: Percent of Detected Joints (PDJ), which is used to measure the detection rate of com-

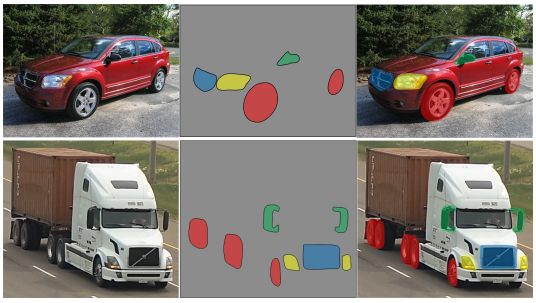

Fig. 4. Each training image has an annotation mask, where four discriminate vehicle components: (a) wheel, (b) headlight, (c) side rear view mirror, (d) front air-intake lattice are labeled respectively.

ponents, where a joint is considered as detected if the distance between the prediction location and the true location is within a certain fraction of the whole object diameter. We carry out experimental performance comparison between our method and three state-of-the-art pose estimation approaches (presented in [11], [12], [13]). Table.2 reports performance of "DeepPose" [11], "Roman et al." [12], "AOG" [13] and "The proposed" in terms of PDJ under diameter-fraction=0.3 for all the four vehicle components and average. Fig.5 visualizes the PDJ value of four methods under variable diameter-fraction.

Table 2. Results of four approaches in terms of PDJ.

PDJ-score	Proposed	DeepPose	Roman et al.	AOG
Wheel	**0.927**	0.896	0.851	0.787
Headlight	**0.883**	0.814	0.776	0.697
Mirror	0.815	**0.847**	**0.841**	0.775
Air-intake	**0.913**	0.871	0.811	0.753
Average	**0.886**	0.857	0.820	0.753

Table.2 shows results on the four discriminative components: Wheel, Headlight, Mirror and Air-intake, as well as the average value of all the four algorithms. The proposed method outperforms other approaches, especially achieving better estimation for Wheel, Headlight and Air-intake. For example, for Wheel we obtain 0.927 compared to 0.896 of the next best performing method. Although DeepPose and Roman et al. don't exhibit strengths across all the components, but they do show strong results for Mirror class.

By varying this fraction, PDJ alleviates the drawback of PCP since the detection criteria for all joints are based on the same distance threshold. To better analyze the performance, Fig.5 presents the results in terms of PDJ under different diameter fractions. Similarly to Table.2, the proposed gets better detection rate. Especially, the proposed method gets remarkable improvenment over others in the fraction's range of [0.1-0.25]. For example, at fraction=0.15 we obtain a competitive detection rate about 0.67, while the next best performing method (DeepPose) gets a detection rate equaling to 0.58. To get a better idea of the proposed method, we visualize a set of vehicle images with their estimated poses in Fig.6, where each line represnts a component pair, lines connecting the same classes of component pair are in the same color (e.g. yellow lines connect tire and mirror). It can be observed that (1) the angle between lines, (2) the length of lines, (3) the absence/presence of several certain lines, etc provide strong information (e.g. orientation, distance) of vehicle pose.

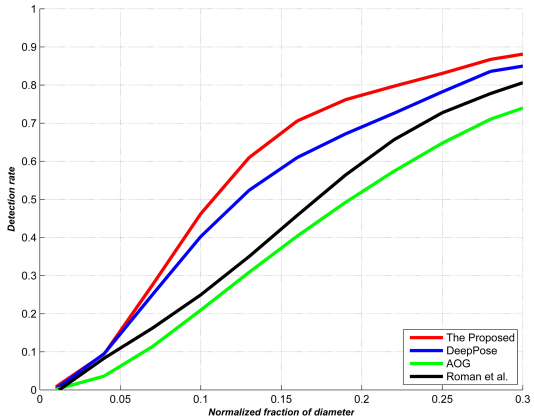

Fig. 5. Average detection rate of 4 methods under varying diameter fraction.

Fig. 6. Vehicle images organized by their estimated poses. Four rows correspond to vehicles with poses from side view, front-right view, front-left view, to front view.

4. CONCLUSION

In this paper, we propose a novel Fully Convolutional Network (FCN) based framwork for vehicle pose estimation. The proposed method is able to detect four discriminative components of a vehicle using probabilistic maps, even vehicles present huge variation in appearance due to many reasons. As the final result, vehicle pose is estimated. And experiments have demonstrated that the proposed approach achieved superior performance compared with three state-of-the-art pose estimation methods.

5. REFERENCES

[1] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2010.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation.," *CVPR*, 2014.

[3] Ross Girshick, "Fast r-cnn.," *CVPR*, 2015.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *NIPS*, 2015.

[5] Geoffrey E. Hinton. Alex Krizhevsky, Ilya Sutskever, "Imagenet classication with deep convolutional neural networks.," *NIPS*, 2012.

[6] Trevor Darrell. Jonathan Long*, Evan Shelhamer*, "Fully convolutional models for semantic segmentation.," *CVPR*, 2015.

[7] Weidi Xie, J. Alison Noble, and Andrew Zisserman, "Microscopy cell counting with fully convolutional regression networks," *Deep Learning Workshop in MICCAI*, 2015.

[8] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," *CVPR*, 2004.

[9] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, "3d object representations for fine-grained categorization.," *4th IEEE Workshop on 3D Representation and Recognition at ICCV*, 2013.

[10] M. Ozuysal, V. Lepetit, and P.Fua, "Pose estimation for category specific multiview object localization.," *CVPR*, 2009.

[11] Alexander Toshev and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks.," *CVPR*, 2015.

[12] Roman Juranek, Adam Herout, Marketa Dubska, and Pavel Zemck, "Real-time pose estimation piggybacked on object detection.," *ICCV*, 2015.

[13] B. Li, T. Wu, and S.-C. Zhu, "Integrating context and occlusion for car detection by hierarchical and-or model.," *ECCV*, 2014.