LETTER

A Part-Based Gaussian Reweighted Approach for Occluded Vehicle Detection

Yu HUANG†, Nonmember, Zhiheng ZHOU†(a), Member, Tianlei WANG†(b), Qian CAO†, Junchu HUANG†, Nonmembers and Zirong CHEN†, Nonmembers

SUMMARY Vehicle detection is challenging in natural traffic scenes because there exist a lot of occlusion. Because of occlusion, detector’s training strategy may lead to mismatch between features and labels. As a result, some predicted bounding boxes may shift to surrounding vehicles and lead to lower confidences. These bounding boxes will lead to lower AP value. In this letter, we propose a new approach to address this problem. We calculate the center of visible part of current vehicle based on road information. Then a variable-radius Gaussian weight based method is applied to reweight each anchor box in loss function based on the center of visible part in training time of SSD. The reweighted method has ability to predict higher confidences and more accurate bounding boxes. Besides, the model also has high speed and can be trained end-to-end. Experimental results show that our proposed method outperforms some competitive methods in terms of speed and accuracy.

key words: vehicle detection, occlusion, reweight

1. Introduction

Vehicle detection is important for road safety. However, vehicle detection is challenging due to complex traffic environment especially occlusion. The goal of vehicle detection is to obtain locations and confidences of vehicles. Recently, deep learning based detectors have attracted more and more researchers because of their high accuracies. Methods base on deep learning are mainly divided into two modes: region proposal based two stage methods [1]–[4] and regression based single stage methods [5], [6]. R-CNN [1] is one of two stage methods. R-CNN crops and scales the image into multiple images and then using convolutional neural network extracts features for Support Vector Machine (SVM) classifier. R-CNN finally regresses the offsets of bounding boxes for better localization. Fast R-CNN [3] straightly adds bounding box regression to the training of CNN and adds adaptive pooling layer according to SPP-Net [2]. Faster R-CNN [4] proposes a new region proposal method called Region Proposal Network (RPN) to reduce the time that generate region proposals. Another methods are single stage methods. To improve mean average precision (mAP), Single Shot Detector (SSD) [5] inherits anchor box mechanism of Faster R-CNN and applies regression to multilayer. This mechanism improves mAP of detection, also has high processing speed. DeepParts [9] model constructs a set of part prototypes and trains a classifier for each part prototype for occluded pedestrian detection.

However, above mentioned detectors except DeepParts all have mismatch between features of proposals and labels for occluded vehicle detection. The mismatch will lead to low confidences and less accurate bounding box. DeepParts can tackle this problem. However, multiple components in DeepParts should be separately learned. Besides, DeepParts can’t meet the real-time demand of vehicle detection. In order to improve detection results of occluded vehicles, we propose a new method by reweighting the weights of anchor boxes. Our observation is that original matching strategy of SSD may have mismatch between receptive field of anchor boxes and features of current vehicles. As a result, predicted bounding boxes might shift to surrounding vehicles when the training features are closer to surrounding vehicles. This type of detection results will be counted as false positives. Our main goal is to predict more accurate bounding boxes and give them higher confidences by reweighting the anchors which are closer to current vehicles with higher weights. So the anchors which have closer receptive field to the features of matched ground truth will have more contributions to loss. This letter proposes a method for occluded vehicle detection. We calculate the center of visible part based on prior occlusion information in traffic scenes and then reweight each matched anchor box according to the unoccluded part and a variable-radius Gaussian function. Our main contribution is that: First, for occluded situation, the proposed method has ability to predict more accurate bounding boxes and higher confidences. Second, the model can be trained end-to-end. Besides, compared to other methods, the proposed method has higher speed and real-time inference. Finally, to verify the the effectiveness and efficiency of our method, the model is tested on KITTI [7] dataset. Also, the approach is extended to CITYSCAPES dataset [8]. Experiments show the robustness of our method.

2. Reweighted Approach According to Occlusion

The proposed method consists of two parts: calculation of center of non-occluded region and Gaussian weight based reweighted loss function. The first part computes the cen-
2.1 Center of Non-Occluded Region Calculation

A bounding box \( b = (b^x, b^y, b^w, b^h) \) contains center \((b^x, b^y)\), width \(b^w\) and height \(b^h\) of this bounding box. For each non-occluded ground truth, we directly use its center as center of non-occluded region. For each occluded ground truth, we get its non-occluded region based on the ground truths which have overlap with current ground truth. For example, if one ground truth bounding box \( g_j \) is occluded by other ground truths \( \{g_k\}_{k \neq j} \), we firstly get all ground truths which satisfy \( \text{IoU}(g_k, g_j) > 0 \) and \( g_k^b > g_j^b \). This set is denoted as \( G_j = \{g_k|g_k^b > g_j^b, \text{IoU}(g_k, g_j) > 0, k \neq j\} \). Given a ground truth \( g_j \) and an occluded set \( G_j \), the bounding box \( o_j \) is used to denote occluded region. The \( o_j \) can be computed by Eq. (1):

\[
\begin{align*}
\hat{o}_j^l &= \min\{g_k^x - \frac{1}{2}g_k^w, g_k \in G_j\} \\
\hat{o}_j^r &= \max\{g_k^x + \frac{1}{2}g_k^w, g_k \in G_j\} \\
\hat{o}_j^u &= \min\{g_k^y - \frac{1}{2}g_k^h, g_k \in G_j\} \\
\hat{o}_j^d &= \max\{g_k^y + \frac{1}{2}g_k^h, g_k \in G_j\} \\
o_j &= \left(\frac{1}{2}(o_j^l + o_j^r), \frac{1}{2}(o_j^u + o_j^d), o_j^l - o_j^r, o_j^u - o_j^d\right)
\end{align*}
\]

Where \( o_j^l, o_j^r, o_j^u, o_j^d \) represent the coordinate of left, right, top and bottom boundary of occluded region respectively. Occluded boundary can be obtained by all the other ground truths that have overlapped region with ground truth \( g_j \).

Non-occluded region is the region in ground truth \( g_j \) excluding the overlap region between occluded region \( o_j \) and ground truth \( g_j \). The non-occluded region can be divided into multiple rectangles. Each rectangle is surrounded by the three straight lines which coincide with the edges of the ground truth bounding box and a straight line of the edge of the occluded area. For the situation that only has one rectangle, we use its center as center of non-occluded region. For the situation that has two rectangles, the centers of two rectangles are \( r_1^q \) and \( r_2^q \), where \( q \in \{x, y\} \). Following equation is used to get the center of non-occluded region:

\[
\begin{align*}
q^* &= \arg\min_q\{|r_q^q|q \in \{1, 2\}\} \\
\rho^q &= \min\{|r_1^q, r_2^q\} + (1 - \frac{a_q}{a_1 + a_2})|r_1^q - r_2^q|)
\end{align*}
\]

For brief expression, we omit subscript \( j \) for \( r^q, a \). We denote area as \( a \) and denote new center for non-occluded region as \( \rho^q \). Equation (2) means that the center is on the straight line between \( r_1^q \) and \( r_2^q \). The location of the center is closer to the center of rectangle region which has larger area. For the situations that have more than two rectangles, we first calculate a new point between two rectangles using Eq. (2) and a new area is defined as summation of area of two rectangles. Then we get the center between previous calculated point and the center of next rectangle using Eq. (2) recurrently. Finally, we get the non-occluded bounding box \( v_j = (v_x^x, v_y^y, v_w^w, v_h^h) \), where \( v_x^x = \hat{r}^x, v_y^y = \hat{r}^y, \) width \( v_w^w \) and height \( v_h^h \) are defined as width and height of a rectangle which has maximal area.

After above mentioned calculation, the center of non-occluded region is obtained. Prior information about location of vehicles help us to get the occlusion relations among vehicles, and occlusion information help us to find the center of non-occluded region for each ground truth. Figure 1(b) illustrates the calculation. Based on the calculated \( v_j \), a reweighted approach is proposed in next subsection.

2.2 Reweighted Approach Based on Occlusion

Anchor boxes which have \( \text{IoU} > 0.5 \) with ground truths are considered as positive samples (Pos) and the rest anchor boxes are negative samples (Neg) in Single Shot Detector. All the weights of positive samples are set to same value so that all positive samples have same importance. However, this approach may lead to the mismatch of receptive fields, so this approach may generate false samples. Because of the mismatch, receptive fields of some anchor boxes which are considered as positive samples are not on current object. Figure 1(a) illustrates an example for this phenomenon.

For this reason, this letter proposes an approach that can relieve mismatch between receptive fields and ground truths. The basic concept of our approach is that we focus training on anchor boxes near non-occluded region. Therefore, the anchor boxes closer to center of non-occluded region, the larger weights the anchor boxes ought to have in training time. Through this improvement, Single Shot Detector will concentrate on learning non-occluded region instead of union region of non-occluded region and surrounding object. In other words, detector can predict bounding box of whole vehicle via part. Our approach achieves higher AP value than original Single Shot Detector. The results of
our experiment support our viewpoint in next section. In order to achieve this goal, a Gaussian function is used to reweight matched anchor boxes because Gaussian function is symmetric and the function values around mean value are bigger than the values far away from it. Based on the calculated center of non-occluded region according to ground truth, the reweighted approach is formulated by:

\[ f_{ij}(d, v) = \alpha \cdot \exp\left(-\frac{1}{2}(d_i^e - v_j^e)^T A_j \sum_i^{-1} (d_i^e - v_j^e)\right) \]

\[ \sum_i^{-1} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}, \quad A_j = \begin{pmatrix} 1/\sigma_j^w & 0 \\ 0 & 1/\sigma_j^h \end{pmatrix} \]  
(3)

\[ d_i^e = \begin{pmatrix} d_i^{ex} \\ d_i^{ey} \end{pmatrix}, \quad v_j^e = \begin{pmatrix} v_j^{ex} \\ v_j^{ey} \end{pmatrix} \]

In Eq. (3), elements \( \sigma_j^w, \sigma_j^h \) in matrix \( A_j \) denote height and width of a non-occluded rectangular which has maximal area respectively. Parameter \( \alpha \) is initial weight of anchor boxes. Vectors \( d_i^e, v_j^e \) are the center of anchor box \( d_i \) and non-occluded region \( v_j^e \) respectively. Because different scales of objects have different non-occluded region, Gaussian function should have different radius for each occluded object. Parameters \( \sigma_1, \sigma_2 \) in Matrix \( \sum_i^{-1} \) control the radius of Gaussian function. Matrix \( A_j \) is introduced to adaptively control radius of Gaussian function according to the width \( (v_j^w) \) and height \( (v_j^h) \) of non-occluded region. The larger width and height will lead to larger radius. Finally, a reweighted function \( f_{ij}(d, v) \) is used to adjust the weight of anchor boxes according to non-occluded region. This process can be demonstrated by Fig. 2. After above reweighting, the model uses loss function as following:

\[ L(x, c, l, g, v) = \frac{1}{N} (L_{conf}(x, c, v) + L_{loc}(x, l, g, v)) \]  
(4)

\[ L_{loc}(x, l, g, v) = \sum_{c \in \text{Pos}} \sum_{i \in \text{Pred}} x_{ij}^l \text{SmoothL1}(l_{ij} - \tilde{g}_j^l) f_{ij}(d, v) \]

\[ \tilde{g}_j^l = (g_j^{lx} - d_i^{lx}) / d_i^w \]

\[ \tilde{g}_j^h = (g_j^{hy} - d_i^{hy}) / d_i^h \]  
(5)

\[ L_{conf}(x, c, v) = -\sum_{c \in \text{Pos}} x_{ij}^p \log(\tilde{c}_i^p) f_{ij}(d, v) - \sum_{c \in \text{Neg}} \log(\tilde{c}_i^p) \text{ where } \tilde{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)} \]  
(6)

The loss function (4) is a sum of confidence loss (conf) and localization loss (loc) similar to SSD. We use \( x_{ij}^p = \{0, 1\} \) to denote an indicator for matching the \( i \)-th anchor box \( (d_i) \) and \( j \)-th ground truth box \( (g_j) \) of category \( p \). Localization loss is a reweighted Smooth L1 loss [3] between predicted box \( \hat{l} \) and ground truth box \( \hat{g} \), where elements in \( l \) represent predicted offsets to center of anchor \( i \) and log scales of height and width of anchor \( i \). Confidence loss is a reweighted softmax loss over predicted vehicle class confidences \( \exp(p, p \neq 0) \) and background class confidences \( \exp(0) \).

### 3. Experiment

The vehicle detection benchmarks KITTI dataset and CITYSCAPES dataset is used to evaluate the proposed approach, and we compare our method to other recent competing methods. KITTI dataset contains 7481 training images and 7518 testing images. Vehicles in these images are various on different perspectives, scales, brightness and occlusion. There are three levels: “Easy”, “Moderate”, “Hard” according to occlusion, truncation, and minimum pixels of object. KITTI considers a bounding box which has 0.7 IoU with ground truth as a “detection”. Original KITTI dataset only contains training set and testing set. The backbone network VGG16 is same as SSD. Learning rate is 0.0005 and weight decay is 0.0005. Parameter \( \alpha \) in formulation (3) is set to 1. SGD and Momentum of 0.9 are used for our experiment. Our approach is implemented in Caffe toolbox. A NVIDIA Titan X GPU was used for training and evaluation. Parameters \( \sigma_1, \sigma_2 \) in Eq. (3) determine the radius of Gaussian function. Due to the large number of different \( (\sigma_1, \sigma_2) \) pairs, we set \( \sigma_1 = \sigma_2 = \sigma \). Radius of Gaussian function decreases as \( \sigma \) increases. When \( \sigma \) is very small, radius will be large. The parameter \( \sigma \) is set to 0.25 and IoU is set to 0.4.

We compare our approach to two kinds of approaches: one stage approach SSD and two stage approaches R-CNN and Faster R-CNN. Table 1 shows the results of detection in KITTI car. Our approach improves the detection results comparing to R-CNN approach by 19.78%, 25.26%, 25.91% in three difficulty levels “Easy”, “Moderate”, “Hard” respectively. Comparing to Faster R-CNN, our approach outperforms it by 1.27%, 1.48% corresponding to “Easy”, “Hard”. However, in difficulty level “Moderate”, AP value of Faster R-CNN is slightly higher than our method, but our model has less detection time. For one stage method, SSD decreases the detection time at the cost of sacrificing detection accuracy. Through the improvement,

| Method          | Easy(%) | Moderate(%) | Hard(%) |
|-----------------|---------|-------------|---------|
| R-CNN[1]        | 68.20   | 55.97       | 46.70   |
| Faster R-CNN[4] | 86.71   | **81.84**   | 71.13   |
| SSD[5]          | 84.19   | 75.52       | 68.32   |
| Our Method      | **87.98**| 81.23       | **72.61**|
Table 2 AP values and recall of SSD and the proposed method on CITYSCAPES car validation set

| Method      | Easy(%) | Moderate(%) | Hard(%) |
|-------------|---------|-------------|---------|
| SSD [5]     | 67.91   | 61.61       | 50.88   |
| recall      | 80.49   | 73.17       | 60.98   |
| Our Method  | 71.52   | 66.04       | 54.75   |
| recall      | 82.93   | 78.05       | 65.85   |

Our approach outperforms SSD by 3.79%, 5.71%, 3.29% AP for “Easy”, “Moderate”, “Hard” level respectively. As shown in the results, the proposed approach can improve AP value of three levels of difficulty. However, improvement of “Moderate” level is larger than “Easy” and “Hard” level. It is probably because our proposed method uses relatively smaller IoU which may increase number of training samples, and the proposed method focuses on the features of visible parts. So the method improves the AP value for all levels of difficulty. However, improvement of AP value of “Hard” level is smaller than the “Moderate” level’s. The main reason is probably that features of non-occluded region for high occluded vehicles are not enough to predict more accurate bounding boxes and higher confidences. Nevertheless, features of non-occluded region for vehicles of “Moderate” level are able to predict better results. However, detection speed is also another important criterion. We will compare the processing time of above detectors. An Intel CPU i7-6950X with 3.00GHz and a TITAN X GPU are used for testing. R-CNN takes about 50s for detection per image and Faster R-CNN takes 0.2s per image. SSD is faster than Faster R-CNN and is near real-time (about 0.045s per image). Our approach is based on SSD and we only improve the training process. So, the model has the same inference time as SSD but higher AP values.

We also conduct our experiments on the CITYSCAPES dataset. This dataset consists of 2975 images for training and 500 images for validation. We pick the smallest box that encapsulates all the pixels of the mask of cars as the bounding box. The model is trained on training set and evaluated on the validation set only for the category car. The evaluation metric is the same as KITTI. The results are shown in Table 2. Our approach outperforms SSD by 3.61%, 4.43%, 3.87% AP on “Easy”, “Moderate”, “Hard” respectively. Also, the proposed approach is better than SSD by 2.44%, 4.88%, 4.87% for recall. These results demonstrate that our method has ability to detect more accurate bounding boxes especially for occluded vehicles.

Comparison of detection with SSD is shown in Fig. 3. The first row is results of SSD and the second row is results of our proposed approach. As shown in the column (a)(c), some predicted bounding boxes of SSD shifted to surrounding vehicles. So SSD detected the junction between two vehicles. This is mainly due to the mismatch of anchor boxes and receptive fields in the training step in SSD method. After improving the training step, this phenomenon can be relieved and the model can generate more accurate detection results. As shown in the second column(b), it is hard for SSD to detect the vehicles for high occlusion level. However, the proposed method can detect more occluded vehicles and product higher confidences for occluded vehicles, as well as higher recall. This is probably because our method concentrates on the non-occluded region of vehicles. Some of the other detection results are shown on Fig. 4. As shown in Fig. 3 and Fig. 4, the proposed method has ability to detect different degrees and various perspectives of occlusion.

4. Conclusion

The mismatch between anchor boxes and features for occluded vehicles may lead to lower confidences and less accurate bounding boxes. Therefore, this letter reweights each anchor box in loss function based on occlusion information and a variable-radius Gaussian function. Experimental results show that our proposed approach achieves better performance comparing to recent competing methods in terms of accuracy and speed.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” Proc. CVPR, pp.580–587, 2014.
[2] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” Proc. ECCV, pp.346–361, 2014.
[3] R. Girshick, “Fast r-cnn,” Proc. ICCV, pp.1440–1448, 2015.
[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Machine Intell., vol.39, no.6, pp.1137–1149, 2017.
[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg, “Ssd: Single shot multibox detector,” Proc. ECCV, pp.21–37, 2016.
[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” Proc. CVPR, pp.779–788, 2016.
[7] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” Proc. CVPR, pp.3354–3361, 2012.

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” Proc. CVPR, pp.3213–3223, 2016.
[9] Y. Tian, P. Luo, X. Wang, and X. Tang, “Deep learning strong parts for pedestrian detection,” Proc. ICCV, pp.1904–1912, 2015.