Pattern Recognition 64 (2017) 417–424

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning

Liujuan Cao, Feng Luo, Li Chen, Yihan Sheng, Haibin Wang, Cheng Wang, Rongrong Ji*

School of Information Science and Engineering, Xiamen University, 361005, China

ARTICLE INFO

Keywords:
Multiple instance learning
Density estimation
Multiple instance SVM
Vehicle detection

ABSTRACT

Vehicle detection in satellite images has attracted extensive research interest with widespreading application potentials. The main challenge lies in the difficulty of labeling sufficient training instances (vehicle rectangles) across all resolutions and imaging conditions of satellite images, which degenerates the performance of vehicle detectors trained correspondingly. To tackle this challenge, in this paper we propose an intelligent and labor-light scheme for large-scale training of vehicle detectors. Our scheme only requires region-level *group* annotation, *i.e.* whether this region contains vehicle(s) or not, without explicitly labeling the bounding boxes of vehicles. To this end, a novel weakly supervised, multi-instance learning algorithm is designed to learn instance-wise vehicle detectors from such "weak labels". In particular, a density estimator is firstly adopted to estimate the density map of vehicle instances from the positive regions. Then, a multi-instance SVM is trained to classify and locate vehicle instances from this map. We have carried out extensive experiments on a large-scale satellite image collection that contains various resolutions and imaging conditions. We have demonstrated that the proposed scheme has achieved superior performance by comparing to a set of state-of-the-art and alternative approaches.

1. Introduction

Coming with the advances of high-resolution satellite imaging techniques, nowadays traffic conditions can be directly monitored by using satellite images. Comparing to the traditional traffic monitoring methods that rely on surveillance or aerial cameras, satellite image merits in its high convenience, low cost, as well as excellent privacy. As a fundamental driving force, the resolution of satellite images has been significantly improved in the past decade, which nowadays can provide 0.5 1 m resolutions, such as IKONOS (1 m), QuickBird(0.61 m), WordView(0.5 m), *etc.* Although such resolution is still not comparable to that of surveillance or aerial image, it has been already sufficient for traffic monitoring, which has attracted extensive research interest in recent years.

Among various tasks of traffic monitoring, vehicle detection is by any means one of the core steps. Detecting vehicles from satellite images has lots of real-world applications, such as battlefield directing, intelligent transportation, and trajectory analysis, *etc.* To this end, most existing works rely on training detectors by using instance-wise labels [1–3] (*e.g.*, vehicle bounding boxes), which involves the detection of either moving [1,2] or parking vehicles [3]. To sum up, the existing approaches of vehicle detection in aerial or satellite images can be categorized according to the features used in the methods, *i.e.*:

- Explicit model [3–5], which clusters similar pixels into potential vehicle regions. Such method usually adopts a top-down matching scheme to find the best-matched candidate in the satellite image to the vehicle. For instance, Hinz et al. [4] proposed a 3D car model based on significant edges for vehicle detection. Holt et al. [3] and Lenhart et al. [5] adopt clustering based object detection schemes to train car detectors.
- Implicit model [6,7], which models vehicles by using intensity or texture features surrounding individual pixels, for instance surrounding contour [6] or histograms of oriented gradients [7]. In the implicit model, a vehicle is usually described by using wire-frame representation. And the detection is performed by checking the posterior probabilities of the target regions to the vehicle model.

To train a robust vehicle detector with high accuracy, all the existing setting requires a sufficient amount of fully annotated, instance-wise labels of vehicles. Specifically, such labels are quite sensitive across different resolutions and among different imaging conditions/devices, *i.e.*:

* Corresponding author.
E-mail addresses: caoliujuan@xmu.edu.cn (L. Cao), rrji@xmu.edu.cn (R. Ji).

http://dx.doi.org/10.1016/j.patcog.2016.10.033
Received 1 June 2016; Received in revised form 17 September 2016; Accepted 30 October 2016
Available online 03 December 2016
0031-3203/

- To ensure high detection accuracy, the detectors are trained with respect to the visual appearances at different spatial resolutions. As quantitatively shown in the previous work [1], the performance of cross-resolution vehicle detection is typically not acceptable.
- Different imaging conditions and imaging devices typically produce images with diverse quality, even such images are with the same spatial resolution. Even with the same spatial resolution (*a.k.a.*, 0.5 1 m panchromatic images like IKONOS (1 m), QuickBird(0.61 m), WordView(0.5 m)), images produced by different satellites at different imaging devices/conditions are still variable, which is due to issues like spectral bands, aspect ratios, and exposure durations.

To that effect, it is an extremely labor-intensive procedure to collect a sufficient amount of training/labor data to train detectors at different resolutions, or for different imaging conditions/devices.

Is it possible to design a more user-friendly scheme to label large-scale training data to train vehicle detectors for different resolutions (or different imaging conditions/devices)? In this paper, we propose to tackle the drawback of data labeling from a novel weakly supervised, multi-instance learning perspective. Rather than the previous works that label vehicles at the instance level, our scheme only requires region-level annotation, *i.e.* whether this region contains vehicle(s) or not, without explicitly identifying the specific bounding boxes of vehicles.[1]

Our key innovation lies in predicting locations of vehicles by combining multi-instance learning with density estimation. More specially, the proposed scheme predicts locations of vehicles by the multi-instance model trained from region-wise labels, in which the density estimation provides constraints of the vehicle number in the corresponding regions. To that effect, instance-wise vehicle detectors are learned from such labor-light, region-level annotations. We achieve this goal by combining cutting-edge techniques in object counting and multi-instance learning. In particular, we adopt an integer programming based scheme to coarsely estimate the density map of the target object for a given region. Then, a multiple-instance Support Vector Machine is adopted to learn instance-level classifier. Under such a design, the users are only required to label at the region level, *i.e.*, identifying the existence/non-existence of vehicle objects. Such labeling is extremely efficient and user-friendly to collect large-scale annotations. Then, we adopt the proposed scheme to iteratively learn from these weak labels, which produces the final instance-level object detectors. Fig. 1 illustrates the proposed framework.

In particular, we propose a Multi-Instance Learning with Density Estimation scheme, termed MIL-DE, which works via performing the following steps: in the offline training, features such as dense SIFT [8] or random forest [9] are firstly extracted from the training images. Such features are quantized into a codebook via standard schemes, *e.g.*, K-Means clustering, which represents each image by a bag-of-words feature descriptor. In order to reduce the computation cost of high-dimensional features, we then adopt the Principle Component Analysis (PCA) algorithm to reduce the feature dimension. Then, superpixel segmentation is conducted on the entire image, upon which a density map is estimated to measure the potential existence of vehicles, as inspired by [9]. This map aims to minimize the error between the ground-truth object location and the estimated density distribution. Over this map, an integer programming based scheme is further proposed to localize vehicle objects, which is extremely efficient that adopts 2D integer programming to recover the locations of objects in the estimated density map. Such coarse counting is subsequently sent to the multi-instance SVM based classifier [10] to learn a fine guess of the actual object location.

In sum, the main contributions of the proposed MIL-DE scheme are three folds, *i.e.*:

- A user-friendly, labor-light scheme is proposed to efficiently collect large-scale vehicle annotations with respect to varying resolutions and imaging devices.
- A density estimation algorithm with integer programming based scheme is proposed to learn from such weak labels to estimate the initial vehicle location.
- A large-margin classification scheme termed MIL-SVM is further adopted to learn refined vehicle detector, which is discriminative by given the above density map based estimations.

We have quantitatively evaluated the proposed MIL-DE scheme on a large-scale vehicle dataset collected and labeled from satellite imagery. Quantitative results have demonstrated that the proposed scheme has outperformed several state-of-the-art and alternative methods.

The rest of this paper is organized as: Section 2 reviews related work. Section 3 presents the proposed scheme in estimating the object density. Section 4 introduced the overall framework for multi-instance learning based weakly supervised vehicle detection. We present the detailed evaluations in Section 5, with extensive analysis and comparisons to the state-of-the-art and alternative approaches. Finally, this paper concludes in Section 6.

2. Related work

The analysis and classification of satellite images retain in the core of remote sensing research. For example, to analyze high-resolution remote sensing images, Hu et al. [11] presented an improved unsupervised feature learning algorithm, which adopted spectral clustering to solve scene classification. The proposed scheme has been tested on classifying on the UFL-SC benchmark, which adaptively learned local feature representations and discovered intrinsic structures of local image patches. Yang et al. [12] proposed a semi-supervised classification scheme, which solved the classification task of satellite images by learning a high-level feature termed semi-supervised ensemble projection.

Multi-Instance Learning (MIL) was initially proposed by [13] for solving the prediction of the drug molecular activity. The key assumption of MIL lies in that if there is at least one positive example in the bag, the bag is positive, otherwise it is a negative one. Along with decade-long research and development, a wide range of MIL algorithms have been developed, with encouraging progress made in various applications. For instance, Zhang et al. [14] proposed to combine MIL with the expectation maximization and density diverse to predict molecular activity, which is insensitive to the relevant attributes of the data. In this approach, the multiple instance problem can be converted to a single instance problem, and bag's label is predicted by using expectation maximization. In the field of web mining and recommendation, Zhou et al. adopted multiple instance learning to recommend the best suited page index [15]. In this work, an index page is viewed as a bag, and all indexes (web links) in this page are considered as examples. Then, Fretcit-kNN algorithm is adopted to predict bag labels with Hausdorff distance, which measures the similarity between frequent item sets.

In the fields of computer vision and remote sensing, MIL has also been widely used. For example, Chen et al. [16] proposed a new learning method termed DD-SVM, which models the image as a bag, where the target objects and areas are modeled as instances in a given image. A non-linear mapping is defined to map each bag into a point in the feature space of the bag, where support vector machine is adopted to train a bag-level classifier. In recent years, training object detector with weak supervision via MIL has become an emerging research topic. A multi-fold, multiple-instance learning scheme is introduced in [17] to iteratively train object detector and infer object location in the image with object label. Wang et al. [18] adopted a segmentation based region

[1] In particular, the spatial range of labeled regions is much larger than the vehicle bounding boxes.

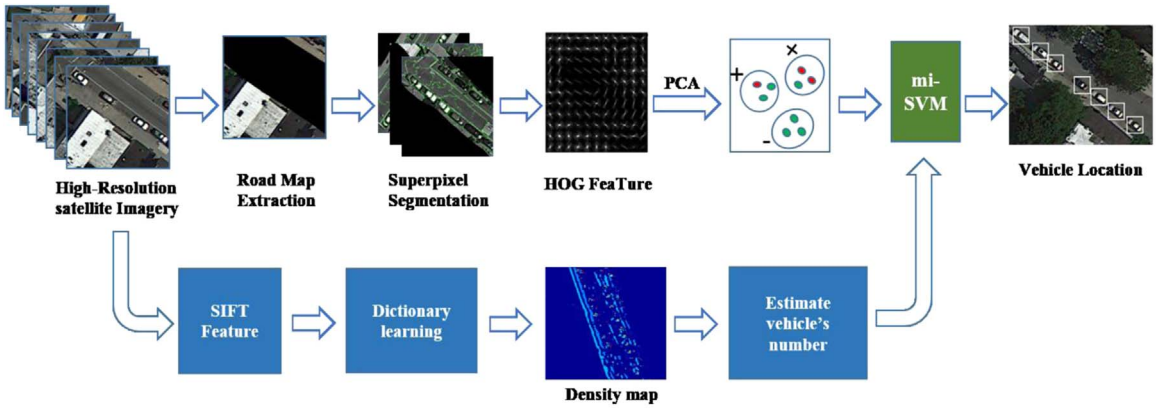

Fig. 1. The proposed weakly supervised vehicle detection scheme for vehicle detection in satellite images via multi-instance learning with density estimation.

proposal to generate candidate object regions. Each region is represented by the Convolutional Neural Network (CNN) features trained on ILSVRC 2011 [19], which are used to train a probabilistic Latent Semantic Analysis (pLSA) for detection. Recent advances in deep learning, for instance the weakly supervised deep detection network (WSDDN [20]), have also been integrated into the field of weakly supervised object detection. In WSDNN, Bilen et al. [21] presented a weakly supervised detection algorithm that encourages similarity between objects to avoid local minima in training, which allows a joint learning of detection and clustering. Wang et al. [22] proposed an MIL strategy for dictionary learning. In the subsequent work [23], Wang et al. proposed to treat the positiveness of instance as a continuous variable, and then adopts Noisy-OR model to enforce the MIL constraints, which jointly optimizes the bag label and instance label in a unified framework. This optimization problem can be efficiently solved using stochastic gradient decent. Very recently, Shen et al. [24] proposed an efficient dividing strategy under MIL setting, named partial random projection tree (PRPT), which represents each sample (bag) by the symmetry features computed at the combinations of specific scales and orientations (instances).

3. Learning to count

In this section, we present the detailed and mathematical formulation of the learning based density estimation, *a.k.a.*, learning to count, which serves as the first component of the proposed scheme. Generally speaking, we first estimate the number of vehicles in a given image, denoted as n. Then, n is fed into an MIL-SVM based large-margin classifier, which restricts the prediction amount of vehicles (instances) and outputs the largest n locations as the positive instances (vehicles). Without loss of generality, we assume that each image is a bag. We extract road regions and segment the regions into superpixels. Then, sliding windows are conducted on these segmentations to get instances, where the instance size and sliding step are set as 90×90 pixels and 10 pixels, respectively. To estimate the density score for each pixel or superpixel, we adopt the object density map proposed by Lempitsky et al., which learns a mapping from the extracted image feature. To learn this mapping, we first collect a set of ground-truth density maps containing vehicle locations. In annotation, the ground-truth objects are annotated as a dot in the center of the object.

In particular, taking pixel-level density for instance, it is assumed that the density of a pixel p around the dot P is in concordance with a Gaussian distribution, corresponding to the sum of the density of all pixels around the dot P. A codebook with K words (in our case K is set as 256) is learned on a set of instance-level training images (in our case set as 20). Then, the local features are quantized by such K words to a bag-of-words descriptor, denoted as e_j where j is the index of the pixel. We further denote the dataset of both labeled and unlabeled bags as $B = (B_1, L_1), (B_2, L_2), \ldots, (B_t, L_t), B_{t+1}, \ldots, B_{t+s}$, where the previous t bags are labeled and the rest are unlabeled. Each bag of B_i is a set of instances in which labels are denoted as $L_i \in \{1, -1\}$, *a.k.a.*, 1 for positive and -1 for negative. We denote $X = x_1, x_2, \ldots, x_n$ as all n instances, where $x_i \in R_k$ represents the feature of the i-th instance. $x_i \in B_j$ means that the i-th instance belongs to the j-th bag. The ground-truth density for training image is:

$$D_i(p) = \sum_{P \in \mathbf{P}_i} N(p; P, \sigma^2 R), \quad \forall\, p \in I_i, \tag{1}$$

where $N(p; P, \sigma^2 R)$ is the 2D Gaussian distribution centered at P with a small covariance. R is a matrix that decides the number of pixels in the Gaussian kernel. P_i is the set of dots. That says, if R is a 2×2 matrix, the density of the image is the sum of $D_i(p)$. And the object density at each pixel is estimated by the density function via:

$$D'_i(p; \alpha) = \alpha^{\mathrm{T}} x_p^i, \quad \forall\, p \in I_i, \tag{2}$$

where $x_p^i \in \mathbb{R}^K$ is the feature vector for pixel p of image I_i and α is the parameter vector, which can be regarded as the weighting vector for each codeword. The parameter is learned by minimizing the sum of the mismatches between the ground-truth and the estimated density functions over the training set:

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} \parallel \alpha \parallel^2 + \gamma_d \sum_{i=1}^{N} F(D_i(\cdot), D'_i(\cdot|\alpha)), \tag{3}$$

where γ_d is a hyperparameter controlling the regularization. The function $F(D_1, D_2)$ proposed in [25] measures the distance between the estimated density and its corresponding ground truth density, such that:

$$F(D_1, D_2) = \max_{B \in B} \left| \sum_{p \in B} D_1(p) - \sum_{p \in B} D_2(p) \right|, \tag{4}$$

$$m_i = \left\lfloor \sum_{p \in I_i} D'_i(p; \alpha) \right\rfloor \tag{5}$$

The solution turns to be a regression problem with α estimated by a quadratic programming solver [8] with a proper loss function. It means that the number m_i is a constraint for choosing the most likely m positive instances of i-th image.

4. Multiple-instance learning

For a standard MIL algorithm, the goal is to learn a function $f: R_k \to \{1, -1\}$ to predict the label. In contrast, in this paper our goal is to learn a score function $f: R_k \to (-\infty, +\infty)$, which measures how likely the instance can be a positive one (higher score means higher possibility). We adopt a density map based instance estimation to

achieve this goal, which is further integrated into an SVM based large margin classifier training, resulting in a so-called MIL-SVM formulation. Without loss of generality, we first give the formulation of the multi-instance SVM (mi-SVM) [26] as below:

$$mi-SVM \quad \min_{\{y_i\}} \min_{\beta,b,\xi} \frac{1}{2}\|\beta\|^2 + C\sum_i \xi_i \text{ s. t.} \quad \forall\, i\colon y_i(\langle \beta, x_i\rangle + b) \geq 1-\xi_i,$$

$$\xi_i \geq 0,\ y_i \in \{-1, 1\}. \tag{6}$$

Since there exist discrete instances, the goal of Eq. (6) is to find an MI-separating linear hyperplane, such that at least one positive instance from positive bags resides in the positive space, while all negative ones are in the opposite space. This formulation leads to a mixed integer programming problem, which is solved via a simple heuristic optimization. In particular, the score function is defined by the learned β and b as:

$$S(x_i) = \beta^T x_i + b \tag{7}$$

According to the score function $S(x)$, each instance has its own score. After listing them in descending order, the previous m instances are labeled as positive instances. We further show the detailed algorithm described above in Algorithm 1.

Algorithm 1. The proposed framework of multi-instance detector learning with density estimation.

Input:
n images' density map manually $D_1, D_2, \ldots, D_n$, the image I for predicting
1: extract SIFT feature r for each pixel of each image;
2: classify r into u classes using k-means algorithm;
3: quantize each pixel of the image I, calculate its density using Eq. (2)
4: calculate the sum of D_i as the number m of object in the image I;
5: initialize $y_i = L_j$, instance's label equals to the label which its bag in;
6: **repeat**
7: compute SVM solution β, b for dataset with the given label;
8: compute $f(i) = \langle \beta, x_i\rangle + b$ for x_i in positive bags(images);
9: set $y_i = sgn(f(i))$ for every $i \in I$, $Y_I = 1$;
10: **for** each positive bag B_I **do**
11: **if** $i \in I$ *and* $(1 + y_i)/2 == 0$ **then**
12: *compute* $i^* = \text{argmax}_{i\in I} f_i$
13: *set* $y_i^* = 1$
14: **end if**
15: **end for**
16: **until** The given labels have not changed
17: score each slide window of the image I using function $S(x_i)$ by w, b learn in step 7;
18: list the $S(x_i)$ in descending order;
Output:
the previous m slide window

We further present several groups of implementation details as below: the number of images we choose for training is 16. In our observation, it is not essentially true that more training images would result in better estimation of the density map. It is worth to note that, the objectives we aim to classify are vehicles, which cover dozens of pixels, and with regular shape. Therefore, it is reasonable to choose the Gaussian kernel in the above formulation, due to its simplicity and effectiveness. Besides, although HOG feature costs less time in extraction and description, SIFT feature generally performs better under the conditions of varying scales and rotations. In terms of the dictionary size, a bigger k means that it is more accurate when quantizing a pixel, which however costs more time in quantization.

As we can see, if there exist too many sliding windows in an image, the convergence speed will be slowed down. To solve this problem, we apply the superpixel based scheme proposed in [27], along with a practical scheme to extract road regions. In such a case, only candidate superpixels in the road regions are taken into consideration when detecting vehicles.

5. Quantitative evaluations

5.1. Datasets

We test our algorithm on a large-scale dataset collected and labeled (containing both region-based labels and instance-based labels about the existence of vehicles, the latter of which are only used for performance evaluation) from satellite images. To build this dataset, we collect 80 satellite images from Google Earth, in which each image is with 979×1348 resolution, covering the road maps in New York City. We ask a group of volunteers to manually label vehicle regions, which produces 1482 vehicle annotations in total. Both region and instance labels are given as vehicle rectangles, which are variant in visual appearances, imaging conditions, and camera viewing angles.

In our experiment, the instance size is 90×90 pixels, and hence the corresponding HOG feature size is 4356 ($11 \times 11 \times 36$). Direct usage of such high-dimensional features will lead to heavy computation burden for subsequent learning model. To address this problem, we perform Principle Component Analysis (PCA) algorithm to reduce the feature dimension, which significantly reduces the feature dimension into 300. To further deal with the challenges introduced by the scale and the orientation changes of vehicles in the HRRS images, we produce positive instance by using various angles, a variety of scales, and various types of vehicles. In such a way, we can partially deal with the problem of scale and orientation changes of vehicles.

5.2. Evaluation protocols

The accuracy of the proposed scheme is tested by using the *Precision–Recall curve* and the *mean classification error*. We compare the proposed scheme to a set of baselines and state-of-the-art approaches, including:

(1) *sMIL*(*convex*): The sparse MI learning (sMIL) algorithm uses both set and instance kernels to perform sparse coding based multiple instance learning [28].
(2) *stMIL*(*sound*): The sparse transductive MI learning scheme is another alternative scheme, which forces instances within bags to be outside the margin, and adds one constraint to make the problem nonconvex comparing to sMIL(convex).
(3) *sbMIL*(*sound*,*convex*): The Sparse balanced MI learning(sbMIL) scheme searches for a balancing parameter representing the fraction of positive examples in positive bags.
(4) *MI-SVM*(*sound*,*complete*): The MI-SVM scheme [10] approach effectively chooses a single witness or prime instance from both positive and negative bags in the dataset.
(5) *WSDDN*: It offers an end-to-end method for weakly supervised object detection using pre-trained CNNs, performing simultaneously region selection and classification [20].
(6) *mi-SVM*(*sound*,*complete under strong consistency*): The last scheme ensures that at least one instance label in each positive bag is positive.

For all the above approaches except WSDDN [20], HOG based descriptors [29,30] are adopted to extract features.

5.3. Parameter tuning

From the aerial images, we adopt a 1:5 leave-one-out (training vs.

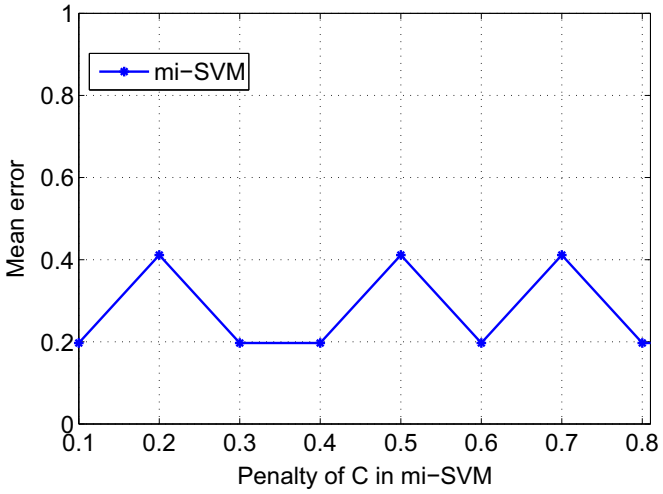

Fig. 2. Parameter tuning on the penalty C on mi-SVM.

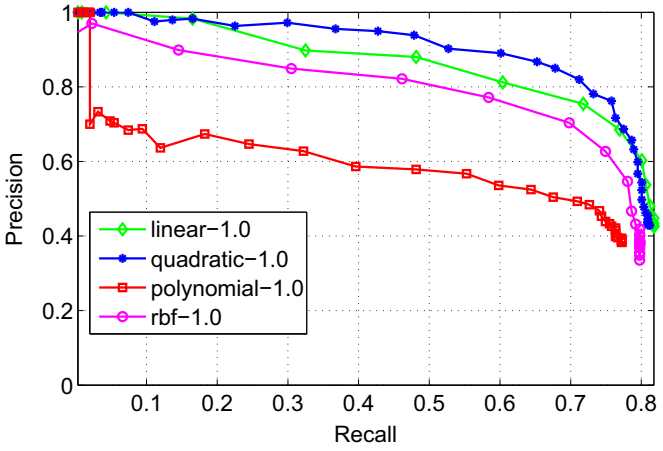

Fig. 3. The influence of a variety of kernel.

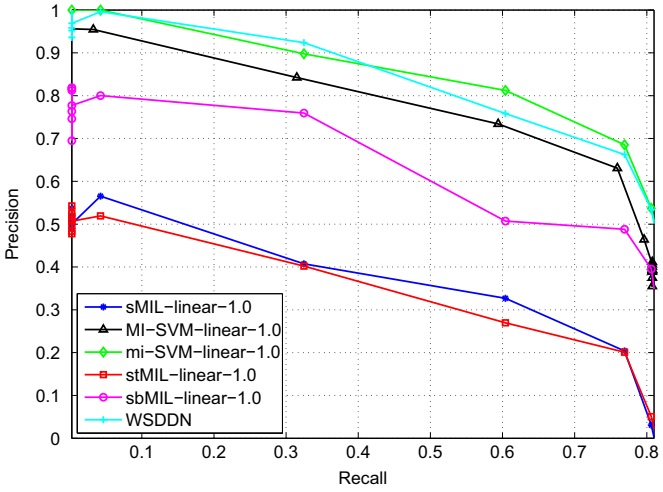

Fig. 4. Quantitative comparison to the state-of-the-art methods.

validation) setting in parameter tuning. We have identified and tuned the following parameters, which are shown to influent the overall performance of the proposed scheme:

(1) *Different kernels used in SVM*: We observe that the prediction precision is affected by using different kernels in SVM . As shown in Fig. 3, we tune to seek the best kernel by using the validation set

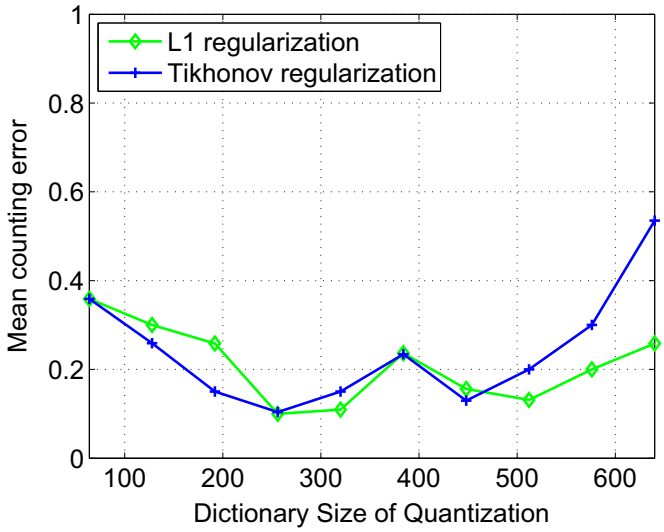

Fig. 5. The influence of size of dictionary.

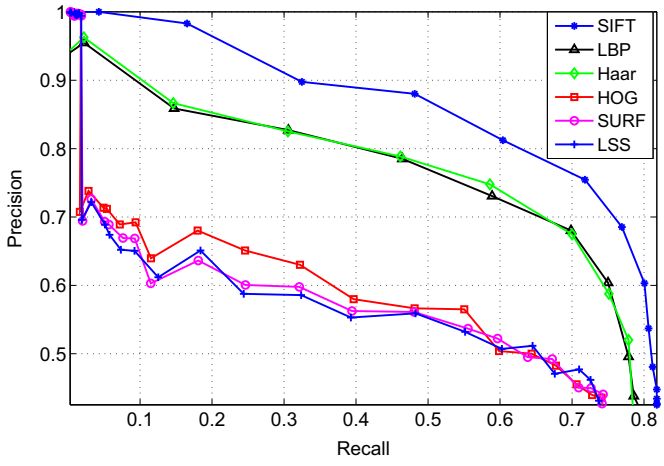

Fig. 6. Parameter tuning on different features.

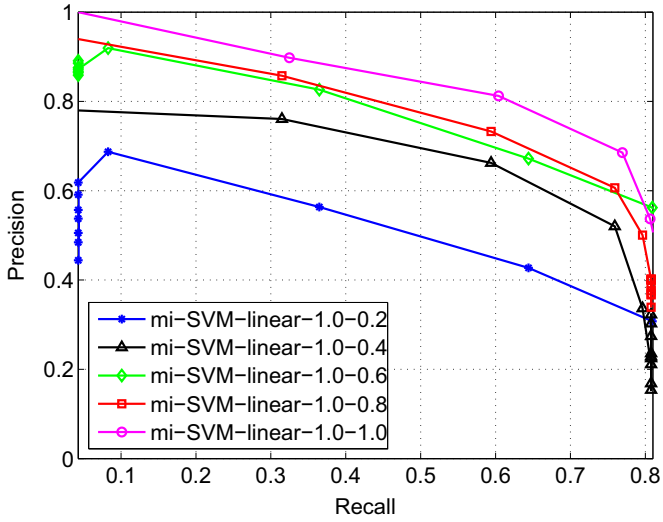

Fig. 7. Parameter tuning on different train sizes.

.

(2) *Penalty C in learning SVMs*: We identify that our scheme is robust (while with slight changes) to the setting of the penalty *C* in SVMs learning, as shown in Fig. 2.

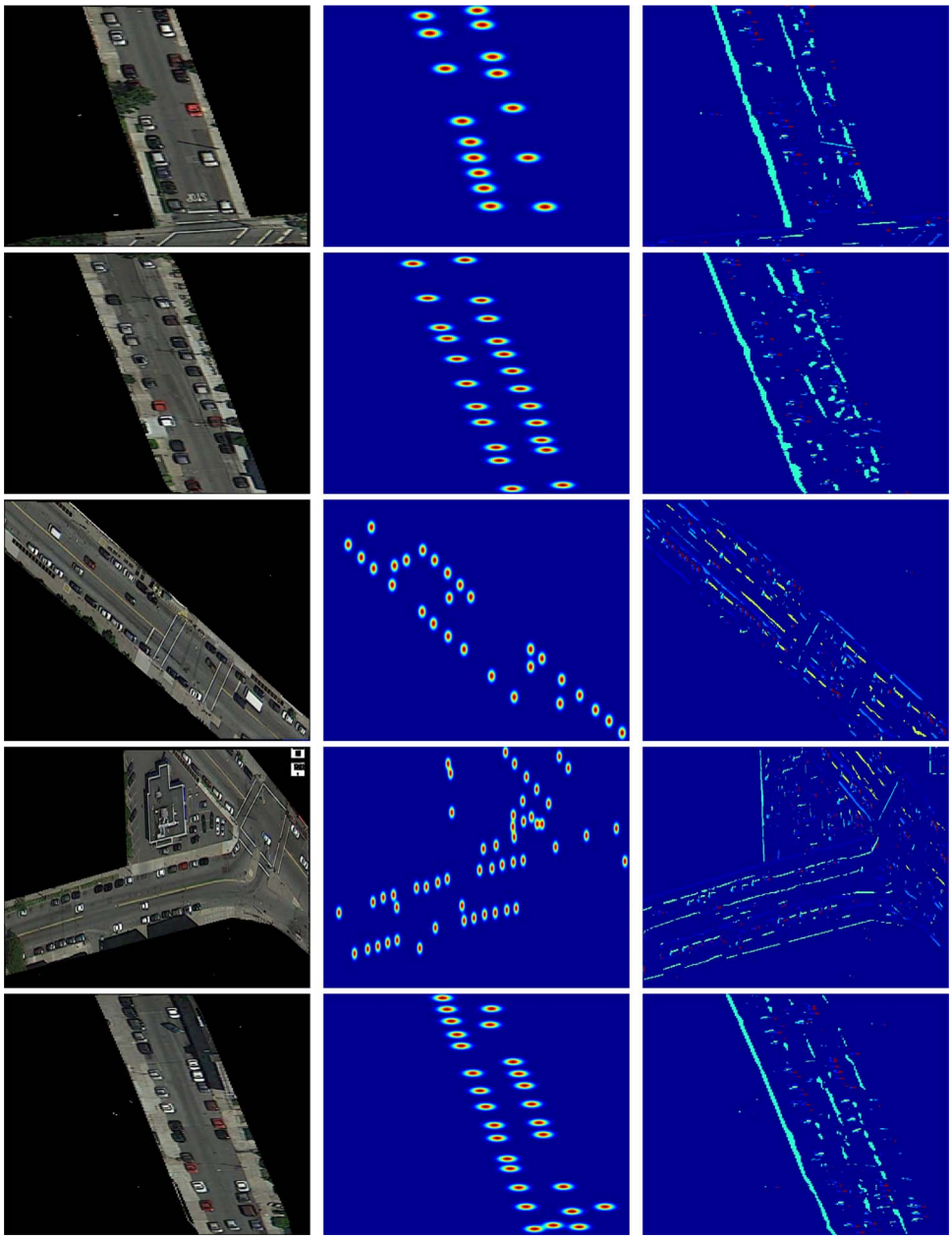

Fig. 8. The first column indicates the satellite image after filtering through the road, the second column indicates ground truth of density map, the third column indicates estimation of density map.

(3) *The size of dictionary*: We also tune the parameter of the dictionary size used in training, the results of which are shown in Fig. 5.
(4) *The size of training set*: We gradually reduce the training set using random selection, performance comparison of different scales are shown in Fig. 7.
(5) *Different features used in density estimation*: We have compared the efficiency of few different features like SIFT, LBP, HOG, Haar, SURF and LSS in Fig. 6, obviously SIFT feature suit the model most. And as Fig. 6 shows that extracting SIFT feature costs time less, so we choose SIFT feature as the meaning feature.
(6) *Comparison to other models*: Finally, we compare the proposed MIL-DD scheme to other alternative schemes as shown in Fig. 4. As can be observed from Fig. 4, the proposed mi-SVM method generally outperforms those competing methods. Although WSDDN achieves similar results as ours, it requires much more training time than our method.

5.4. Quantitative analysis

As shown in Fig. 4, the Precision–Recall curves have demonstrated that our approach achieves consistent and promising performance compared to the five baselines. There is a significant performance boost by replacing the stMIL-linear and sMIL-linear with the proposed method, both of which adopt the multi-instance classifiers. After a pretreatment like road extraction, WSDDN [20] almost has the similar result as ours. With the increase in the size of the picture, the training time of the depth neural network will increase dramatically. In addition, the proposed mi-SVM-linear further push the Precision–Recall curves to the state-of-the-art, as quantitative shown in the benchmark dataset.

5.5. Case study

We further conduct case study to investigate the quantitative

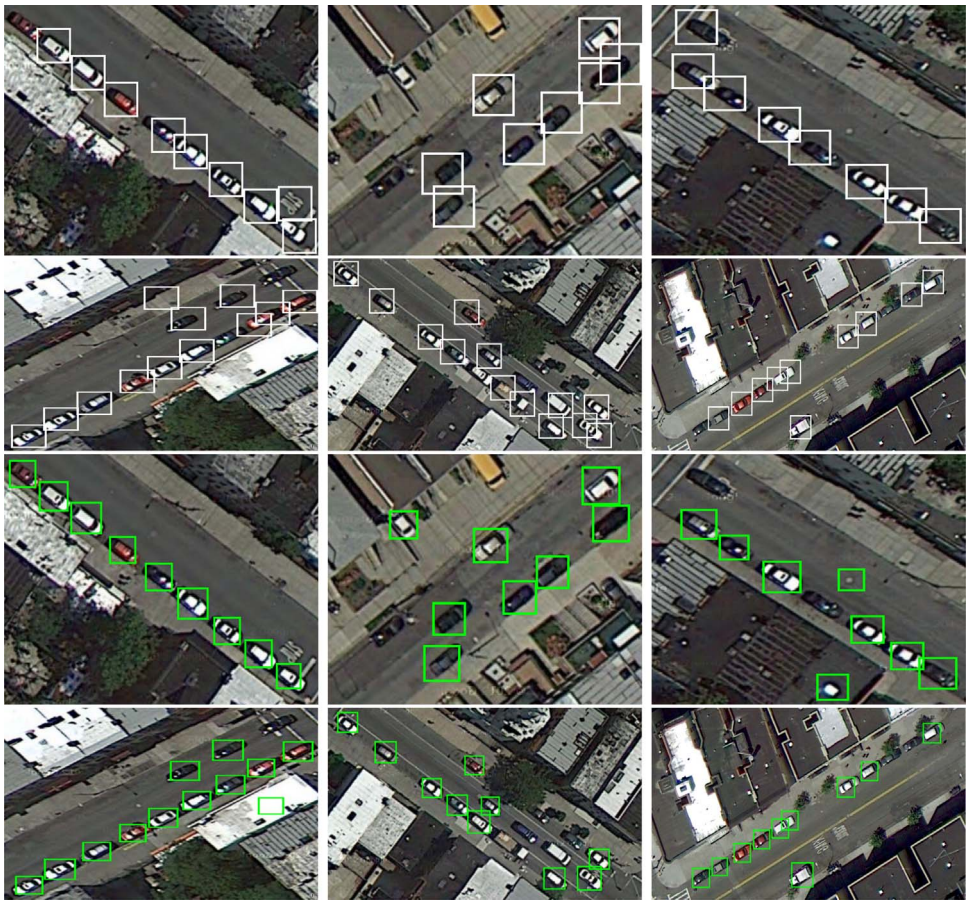

Fig. 9. Visualized vehicle detection results achieved by the proposed method and WSDDN. For each example, the first two rows show the predicted results of our proposed method, and the last two rows report the predicted results of WSDDN.

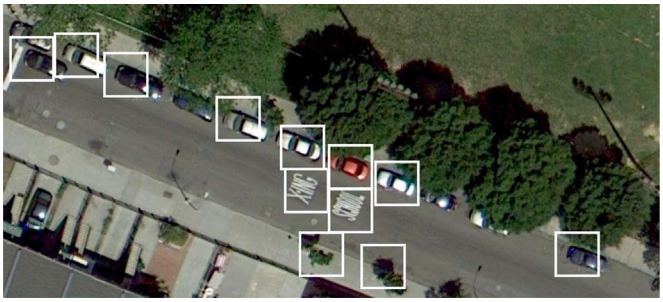

Fig. 10. Vehicle detection results of the proposed method.

performance reported above. We visualize the satellite image of road filter, the ground-truth of density map, as well as the estimation of the density map in Fig. 8. It is shown that vehicles in different directions can be correctly detected, the reason is that in the slide windows of training procedure, there exists vehicles in different directions.

We further show a list of visualized examples to include both correct and incorrect results. As shown in the lower right corner of Fig. 9, we locate 9 out of 11 vehicles in the street region. And there are two vehicles failing to be located. Meanwhile, comparing to the figure referred above, we further show a case of imperfect localization in Fig. 10. In such an extreme case, we only detect 8 vehicles out of 12, in which the false positive mainly happens in locations like tree and logo regions in the street. Despite such an extreme case, our method can get a high precision in most figures without too complicated background.

5.6. Memory and time cost

In the training stage, Litekmeans [31] is used to train the dictionary. This scheme computes inner maxima in (4), then constitutes a 2D maximum subarray, from which the box subarray of a given 2D array with the largest sum is identified. This problem has a number of efficient solutions, among which the most efficient ones [32] is an exhaustive search over one image dimension. This solution searches for the top and bottom dimensions of the optimal subarray, which is combined with the dynamic programming proposed in [33] to handle the 1D maximum subarray problem along the other dimension in the inner loop. This approach has the complexity of $O(|I|^{1.5})$, where $|I|$ is the number of pixels in the image grid.

In our experiments, the time bottleneck lied in the QP solver (3). We use $G = (V, E)$ to denote an undirected graph where V is the vertex set and E is the edge set. The Entropy Rate Superpixel Segmentation [27] algorithm is adopted as the QP solver, which has the complexity of $O(|V|\log(|V|))$. In addition, the work in [27] provides a speedup by a factor of 200–300 for image size 481×321 and on average requires 2.5 s. And the training stage of mi-SVM takes a lot of time about 3 h in a single PC with dual core 3.0 GHz CPU.

6. Conclusion

In this paper, we study the problem of vehicle detection in satellite imagery. Different from all existing works, we propose a novel scheme to label a sufficient amount of training instances (vehicle rectangles) in an extremely labor-light manner. To this end, only region-level annotations are required, which only requires the labels of whether this region contains vehicle(s) or not. To learn vehicle detector from

such region labels, we further design a novel weakly supervised, multi-instance learning algorithm to learn robust, instance-wise vehicle detectors from such "weak labels". Especially, a density estimator is adopted to estimate the instance density from the positive regions. Then, multi-instance SVM is trained to classify and locate vehicle instances from this density map. Extensive experiments are done on a large-scale satellite imagery collection with various imaging resolutions. We have demonstrated that the proposed scheme has achieved superior performance, by comparing to a set of state-of-the-art and alternative approaches.

Acknowledgments

This work is supported by the Special Fund for Earthquake Research in the Public Interest No. 201508025, the Nature Science Foundation of China (Nos. 61422210, 61373076, 61402388, and 61572410), the Open Projects Program of National Laboratory of Pattern Recognition, and the Xiamen Science and Technology Project (No. 3502Z20153003). The authors would also like to thank master students Feng Luo, Li Chen, and Yihan Sheng for experimental settings, data labeling, and quantitative evaluations.

References

[1] L. Cao, C. Wang, J. Li, Vehicle detection from highway satellite images via transfer learning, Inf. Sci. (2016). http://dx.doi.org/10.1016/j.ins.2016.01.004.
[2] H.-Y. Cheng, C.-C. Weng, Y.-Y. Chen, Vehicle detection in aerial surveillance using dynamic Bayesian networks, IEEE Trans. Image Process. 21 (4) (2012) 2152–2159.
[3] A.C. Holt, E.Y. Seto, T. Rivard, P. Gong, Object-based detection and classification of vehicles from high-resolution aerial photography, Photogramm. Eng. Remote Sens. 75 (7) (2009) 871–880.
[4] S. Hinz, Detection of vehicles and vehicle queues in high resolution aerial images, Photogramm. Fernerkund. Geoinformation (2004) 201–214.
[5] D. Lenhart, S. Hinz, J. Leitloff, U. Stilla, Automatic traffic monitoring based on aerial image sequences, Pattern Recognit. Image Anal. 18 (3) (2008) 400–405.
[6] K. Kozempel, R. Reulke, Fast vehicle detection and tracking in aerial image bursts, Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 38 (2009) 175–180.
[7] A. Kembhavi, D. Harwood, L.S. Davis, Vehicle detection using partial least squares, IEEE Trans. Pattern Anal. Mach. Intell. 33 (6) (2011) 1250–1265.
[8] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[9] V. Lempitsky, M. Verhoek, J.A. Noble, A. Blake, Random forest classification for automatic delineation of myocardium in real-time 3d echocardiography, in: Functional Imaging and Modeling of the Heart, 2009, pp. 447–456.
[10] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in Neural Information Processing Systems, 2002, pp. 561–568.
[11] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, H. Sun, Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8 (5) (2015).
[12] W. Yang, X. Yin, G.-S. Xia, Learning high-level features for satellite image classification with limited labeled samples, IEEE Trans. Geosci. Remote Sens. 53 (8) (2015) 4472–4482.
[13] T.G. Dietterich, R.H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artif. Intell. 89 (1) (1997) 31–71.
[14] Q. Zhang, S.A. Goldman, Em-dd: an improved multiple-instance learning technique, in: Advances in Neural Information Processing Systems, 2001, pp. 1073–1080
[15] Z.-H. Zhou, K. Jiang, M. Li, Multi-instance learning based web mining, Appl. Intell. 22 (2) (2005) 135–147.
[16] Y. Chen, J.Z. Wang, Image categorization by learning and reasoning with regions, J. Mach. Learn. Res. 5 (2004) 913–939.
[17] R.G. Cinbis, J. Verbeek, C. Schmid, Multi-fold mil training for weakly supervised object localization, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2409–2416.
[18] C.Wang, W. Ren, K. Huang, T. Tan, Weakly supervised object localization with latent category learning, in: Computer Vision–ECCV 2014, 2014, pp. 431–445.
[19] O. Russakovsky, J. Deng, H. Su, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[20] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, arXiv:1511.02853
[21] H. Bilen, M. Pedersoli, T. Tuytelaars, Weakly supervised object detection with convex clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1081–1089.
[22] X. Wang, B. Wang, X. Bai, W. Liu, Z. Tu, Max-margin multiple-instance dictionary learning, in: ICML (3), 2013, pp. 846–854.
[23] X. Wang, Z. Zhu, C. Yao, X. Bai, Relaxed multiple-instance svm with application to object discovery, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1224–1232.
[24] W. Shen, X. Bai, Z. Hu, Z. Zhang, Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images, Pattern Recognit. 52 (2016) 306–316.
[25] V. Lempitsky, A. Zisserman, Learning to count objects in images, in: Advances in Neural Information Processing Systems, 2010, pp. 1324–1332.
[26] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2002.
[27] M.-Y. Liu, O. Tuzel, S. Ramalingam, R. Chellappa, Entropy rate superpixel segmentation, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 2097–2104.
[28] R.C. Bunescu, R.J. Mooney, Multiple instance learning for sparse positive bags, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 105–112.
[29] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, 2005, pp. 886–893.
[30] M. Liu, S. Chen, D. Zhang, Attribute relation learning for zero-shot classification, Neurocomputing 139 (2014) 34–46.
[31] D. Cai, Litekmeans: the fastest Matlab implementation of kmeans, Software available at: ⟨http://www.zjucadcg.cn/dengcai/Data/Clustering.html⟩
[32] J. Bentley, Programming pearls: perspective on performance, Commun. ACM 27 (11) (1984) 1087–1092.
[33] J. Bentley, Programming pearls: algorithm design techniques, Commun. ACM 27 (9) (1984) 865–871.

Liujuan Cao is currently an assistant professor in the Department of Computer Science, Xiamen University.

Feng Luo is currently a master student in the Department of Computer Science, Xiamen University.

Li Chen is currently a master student in the Department of Computer Science, Xiamen University.

Yihan Sheng is currently a master student in the Department of Computer Science, Xiamen University.

Haibin Wang is currently a master student in the Department of Computer Science, Xiamen University.

Rongrong Ji is currently a full professor in the Department of Cognitive Science, Xiamen University.