Towards Domain Adaptive Vehicle Detection in Satellite Image by Supervised Super-Resolution Transfer

Liujuan Cao[†,‡], **Rongrong Ji**[†,‡], **Cheng Wang**[†,‡], **Jonathan Li**[†,‡]
[†] Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, 361005, China
[‡] School of Information Science and Engineering, Xiamen University, 361005, China
{caoliujuan,rrji,cwang,junli}@xmu.edu.cn

Abstract

Vehicle detection in satellite image has attracted extensive research attentions with various emerging applications. However, the detector performance has been significantly degenerated due to the low resolutions of satellite images, as well as the limited training data. In this paper, a robust domain-adaptive vehicle detection framework is proposed to bypass both problems. Our innovation is to transfer the detector learning to the high-resolution aerial image domain, where rich supervision exists and robust detectors can be trained. To this end, we first propose a super-resolution algorithm using coupled dictionary learning to "augment" the satellite image region being tested into the aerial domain. Notably, linear detection loss is embedded into the dictionary learning, which enforces the augmented region to be sensitive to the subsequent detector training. Second, to cope with the domain changes, we propose an instance-wised detection using Exemplar Support Vector Machines (E-SVMs), which well handles the intra-class and imaging variations like scales, rotations, and occlusions. With comprehensive experiments on large-scale satellite image collections, we demonstrate that the proposed framework can significantly boost the detection accuracy over several state-of-the-arts.

1 Introduction

Coming with the era of digital earth, nowadays have witnessed the explosive growth of satellite images, which opens a gate for various applications ranging from commercial to military. At the core of such applications lies the need to detect vehicles, which typically reside on the roadway from a high-angle shot. While extensive research has been done on vehicle detection in the aerial image domain (Hinz 2004; Holt et al. 2009; D.Lenhart et al. 2008; Kozempel and Reulke 2009; Kembhavi, Harwood, and Davis 2011), there is limited work on low-resolution satellite domain. However, with the evolution of imaging techniques, the recent satellite platforms have provided 0.5$\sim$1m resolutions by which vehicles are recognizable, for instance IKONOS (1m), QuickBird (0.61m), WordView (0.5m) etc. It is therefore emerging to detect vehicles in satellite images, which merit in high confidentiality comparing to using aerial images.

However, it is not an easy task at all. Key challenges mainly fall in two folds: First, comparing to aerial images, the spatial resolution of vehicles in satellite images are lower by an order of magnitude, *i.e.* typically $\leq$ 10 inch for satellite images. Therefor, it is almost infeasible to reuse features and detectors originally designed for the aerial image domain(Cheng, Weng, and Chen 2012a; Hinz and Baumgartner 2011; Choi and Yang 2009; Lin et al. 2008) in the satellite image domain. Second, as such data source is newly pervasive, there are few labels available, neither for moving vehicles (Cheng, Weng, and Chen 2012b) or still vehicles (Holt et al. 2009), in the satellite image domain to train vehicle detectors (Holt et al. 2009; Cheng, Weng, and Chen 2012b). While manual labeling might mitigate such a limitation, it is very labor-intensive due to the low resolution.

In this paper, we present a domain-adaptive vehicle detection framework for satellite image, which addresses both challenges by supervised, super-resolution transfer learning. Figure 1 illustrates the proposed framework. First, we adopt super-resolution to transfer the target region into high resolution to "augment" its visual appearance. This is achieved by a coupled dictionary learning, which jointly learns the feature encoding of both low- and high-resolution images to map image regions from satellite to aerial image domain. Notably, supervised labels are embedded into this transfer: We embed a linear detector loss to the objective function of dictionary learning, which enforces the augmented region to be more discriminative for the subsequent detector training.

Second, we detect vehicles from the augmented regions in the high-resolution aerial image domain. In this domain, there exist extensive vehicle detection algorithms, as well as rich labels. However, the detector robustness is still challenged by domain variations. We address this challenge by adopting Exemplar Support Machines (E-SVMs), which is a robust nearest neighbor classifier with instance-wised metric learning. As above, linear kernel is adopted to link E-SVMs loss into the previous coupled dictionary learning, forming a joint optimization between detection and super-resolution.

The proposed vehicle detection framework has been extensively evaluated in several large-scale satellite image data sets. We compare it to a set of state-of-the-art algorithms to demonstrate the significant performance gains.

The rest of this paper is organized as below: Section 2 reviews the related work. Section 3 presents the proposed supervised coupled dictionary learning algorithm. Section 4

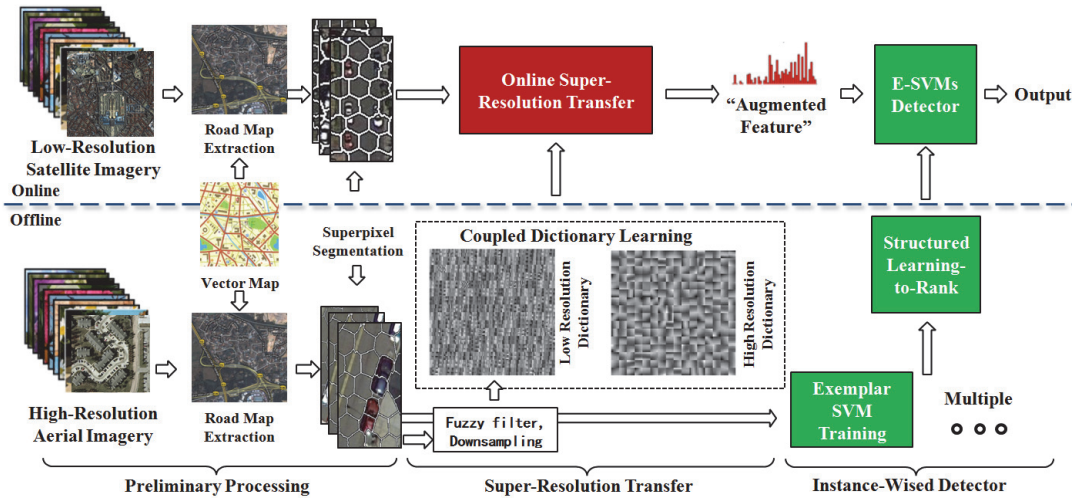

Figure 1: The proposed supervised super-resolution transfer for domain-adaptive vehicle detection in satellite images.

presents the E-SVMs based robust detection algorithm. Section 5 presents the quantitative evaluations in large-scale satellite image dataset. Finally, we conclude in Section 6 and discuss our future work.

2 Related Work

Vehicle Detection: There have been extensive works focused on vehicle detection in aerial image, which can be categorized into two folds by using, *i.e.*, (1) explicit models (D.Lenhart et al. 2008; Hinz 2004; Holt et al. 2009) that cluster pixels to regions then conduct template matching by using edge (Hinz 2004) or region (D.Lenhart et al. 2008; Holt et al. 2009) features, and (2) implicit models (Kozempel and Reulke 2009) that describe the intensity or texture surrounding the vehicle regions via contour (Kozempel and Reulke 2009) or HoG (Kembhavi, Harwood, and Davis 2011) features. Such features are bottom-up grouped to generate structured candidates for detection.

There is limited work on vehicle detection in satellite images. To the best of our knowledge, the work in (Eikvil, Aurdal, and Koren 2009) serves among the earliest ones for vehicle counting on QuickBird satellite images. Leitloff et al. (Leitloff, Hinz, and Stilla 2010) proposed a vehicle detection algorithm using sequential hypothesis in combination with Haar-like features. He et al. (He, Zhou, and Li 2011) proposed to extract roadways based on supervised classification with adaptive thresholds. Mantrawadi et al. (Mantrawadi, Nijim, and Young 2013) proposed to discover vehicle objects from satellite images by saliency based mining. Chen et al. (Chen et al. 2013) introduced parallel branches into the deep convolutional neural network to increase the detection speed and accuracy. However, all above works are designed for high-resolution images (Gerhardinger, Ehrlich, and Pesaresi 2005; Jin and Davids 2007; Sharma 2002; Zheng and Yu 2006), which is unsuitable for low-resolution images captured via the pervasive commercial satellites.

Super-Resolution: Over the past decade, learning based super-resolution (Freeman, Thouis, and Egon 2002) has been a research hot spot. It assumes the low-resolution images as transferred from high-resolution by losing high-frequency components. For instance, Baker and Kanade (Baker and Kanade 2002) proposed the face hallucination for face super-resolution. Yang et al. adopted non-negative matrix decomposition (Yang et al. 2008) and sparse coding (Yang et al. 2010) for face super-resolution. Kim et al. (Kim and Kwon 2010) presented a spare regression based algorithm that adopts ridge regression for super-resolution of a single-frame image.

Typically, coupled dictionaries are trained in above models to link the reconstruction of features or patches from both low- and high-resolutions (Yang et al. 2008; 2010). In this paper, we extend this setting into a "task-dependent" scenario, *i.e.*, to include the linear detection loss to enforce the reconstructed high-resolution patches to be more discriminative for the subsequent detection.

Transfer Learning: The proposed super-resolution transfer broadly relates to the transductive transfer learning (Pan and Yang 2010). One representative work in computer vision comes from (Kuettel, Guillaumin, and Ferrari 2012), which propagates segmentations over ImageNet dataset in a well-controlled manner. For another instance, Rohrbache et al. (Rohrbach, Ebert, and Schiele 2013) presented a novel transductive transfer for ImageNet classification with limited training examples. In the text domain, Zhao et al. (Zhao et al. 2013) proposed a crowdsourcing transfer learning scheme for tweets.

Instance-Wise Classification: Instance-wise classification has been recently popular due to its robustness. In principle, it numerates the potential variations from the training data to train individual classifiers, which enjoys high flexibility and generality. Here, linear classifiers are typically adopted to ensure online efficiency. Among various methods, the Exemplar SVMs (E-SVMs) proposed by Malisiewicz et al. (Malisiewicz, Gupta, and Efros 2011)

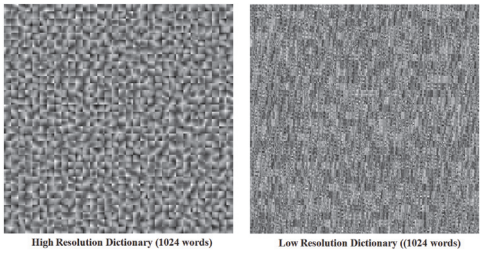

Figure 2: Visualized examples of coupled dictionaries.

serves as one of the most cutting-edge instance-wise classifiers, which combines the merits of both parametric and non-parametric (search-based) classifiers. In this approach, SVMs are trained for individual instances, which are aggregated as a robust nearest neighbor classifier with instance-wise learned metric for online classification. Since the discriminative classifier is able to detect the most unique features for each instance, E-SVMs has recently shown promising performance in object detection, cross-domain retrieval, and point cloud data parsing (Shrivastava et al. 2011; Wang, Ji, and Chang 2013).

3 Supervised Super-Resolution Transfer

In this section, we introduce the proposed supervised transfer learning. In preliminary, for the input low-resolution satellite image, we first extract roadways by aligning images with digital vector maps (detailed in Section 5). Then, superpixel based over-segmentation is done on roadways to extract potential vehicle regions, which are "augmented" to high-resolution and sent to E-SVMs for detection[1] (detailed in Section 4).

3.1 Supervised Coupled Dictionary Learning

To "augment" the superpixel extracted from the low-resolution satellite image, a supervised coupled dictionary learning is proposed. It extends the traditional dictionary learning to a task-dependent formulation, *i.e.*, the detection loss is embedded in the objective function of learning. In such a way, the reconstructed high-resolution regions is expected to preserve discriminative information for the subsequent domain-adaptive detection[2]. Our formulation is detailed below, which is consisted of two steps, *i.e.*, down sampling and coupled dictionary learning, the latter of which is further extended into a supervised learning setting:

Down Sampling: Let $X^h = \{x_1, x_2, ..., x_n\}$ be the set of n superpixels sampled from high-resolution aerial images. Note that if low-resolution images are available, this step can be skipped. Let $S = \{s_1, s_2, ..., s_n\}$ be the corresponding labels, where $s_i = 1$ denotes that the ith superpixel contains (or is part of) a vehicle, and -1 vice versa. Let $Y^l = \{y_1, y_2, ..., y_n\}$ be the corresponding low-resolution superpixels. Y^l is obtained by downsampling from X^h via,

$$Y^l = \Lambda \Theta X^h, \tag{1}$$

where Λ denotes the downsampling operation and Θ denotes the fuzzy filtering. By running Equation 1 over set X^h, we obtain the high-low resolution superpixel mappings $P = \{X^h, Y^l\}$. For each $\{{x_i}^h, {y_i}^l\}$ pair, vector ${x_i}^h$ refers to the unfolded pixel sequence of the high-resolution superpixel, and vector ${y_i}^l$ refers to that of the low-resolution superpixel.

Coupled Dictionary Learning: Coupled dictionary learning is to learn two dictionaries simultaneously to encode P. It typically adopts sparse coding to encode both high- and low-resolution data, as

$$\begin{aligned} V^h &= \arg\min_{\{V^h, U\}} \| F^h - UV^h \|_F^2 + \lambda \| U \|_1, \\ V^l &= \arg\min_{\{V^l, U\}} \| F^l - UV^l \|_F^2 + \lambda \| U \|_1, \end{aligned} \tag{2}$$

where F^h and F^l are features extracted from high- and low-resolution superpixels, V^h and V^l are the learned dictionary for high- and low-resolution superpixels. U denotes the shared coefficients between high- and low-resolutions, and λ denotes the tradeoff between regularization and reconstruction cost. Equation 2 links the learning of both high- and low-resolution dictionaries. Combined with sparse coding, the objective function can be formulated as:

$$\begin{aligned} \min_{\{V^h, V^l, U\}} \frac{1}{N} \| F^h - UV^h \|_2 + \frac{1}{M} \| F^l - UV^l \|_2 \\ + \lambda (\frac{1}{N} + \frac{1}{M}) \| U \|_1, \end{aligned} \tag{3}$$

where N and M are the dimensions of features extracted from high- and low-resolution superpixels respectively. To balance the scale differences between high- and low-resolution features[3] in Equation 3, the above formula is rewritten as:

$$\begin{aligned} &\min_{\{V^h, V^l, U\}} \| F_C - UV_C \|_2 + \lambda (\tfrac{1}{N} + \tfrac{1}{M}) \| U \|_1, \\ &s.t.\ F_C = \begin{bmatrix} \frac{1}{\sqrt{N}} F^h \\ \frac{1}{\sqrt{M}} F^l \end{bmatrix} \quad V_C = \begin{bmatrix} \frac{1}{\sqrt{N}} V^h \\ \frac{1}{\sqrt{M}} V^l. \end{bmatrix}. \end{aligned} \tag{4}$$

The objective function in Equation 4 is solved by Lagrange multiplier, which is transferred as a bi-convex optimization and solved by interactively learning dictionary U and coefficients (V^h, V^l).

Supervised Learning: We further extend the above formulation by incorporating a linear loss from the detector, which is trained in the high-resolution domain as detailed

[1] The roadway detection algorithm is orthogonal to our contribution. And the vehicles are not necessarily to be on roadways. A more comprehensive scheme is to adopt sliding windows to numerate all potential regions in the satellite image.

[2] "Domain-Adaptive" refers to needing no labels from the satellite image domain. Instead, all labels and detectors reside on the high-resolution aerial image domain. It can be also interpreted as a sort of transfer learning.

[3] The feature used here can be different, *e.g.* unfolded raw pixels, HoG descriptor, SIFT feature, or simply the color or textual statistical features.

in Section 4. Instead of using Equation 4 for unsupervised case, we rewrite Equation 4 as:

$$\begin{aligned} &\min_{\{V^h,V^l,U\}} \parallel F_C - UV_C \parallel_2 + \lambda(\tfrac{1}{N} + \tfrac{1}{M})\parallel U \parallel_1, \\ &\quad + \textstyle\sum_{i=1}^{n} s_i(W^T V_{s_i}^h U + \delta), \\ &s.t. \;\; F_C = \begin{bmatrix} \frac{1}{\sqrt{N}} X^h \\ \frac{1}{\sqrt{M}} Y^l \end{bmatrix} \quad V_C = \begin{bmatrix} \frac{1}{\sqrt{N}} V^h \\ \frac{1}{\sqrt{M}} V^l. \end{bmatrix}, \end{aligned} \tag{5}$$

where $\sum_{i=1}^{n} s_i(W^T V_{s_i}^h U + \delta)$ is the cost from n instance-wised linear classifiers. The learned dictionary U enforces the reconstructed high-resolution feature to be discriminative for the subsequent detection stage.

In online, given feature f^l extracted from a candidate superpixel, sparse coding is conducted to transfer f^l to $f^h = UV^h$, which is done by sharing reconstruction coefficients u between high- and low-resolutions:

$$\min \parallel U \parallel_0 \;\; s.t. \parallel V^l U - f^l \parallel_2 \ll \omega. \tag{6}$$

We further solve the above L_0 norm optimization to the L_1 norm, which revises the above formulation as:

$$\min \parallel U \parallel_1 \;\; s.t. \parallel V^l U - f^l \parallel_2 \ll \omega, \tag{7}$$

by which the augmented feature $\bar{f^h}$ in the high-resolution domain is obtained by $\bar{f^h} = V^h U$.

Subsequently, detectors are run on this augmented feature $\bar{f^h}$ to determine whether the candidate superpixel is vehicle. The detection scores from spatially nearby superpixels are aggregated with a non-maximal suppression. In the following section, we introduce the details of our instance-wised vehicle detector.

4 Exemplar-SVMs for Vehicle Detector

The design of vehicle detector in the high-resolution domain should pay special focus on the "cross-domain" variations. In other words, the detector should be robust enough against changes in spatial resolutions, visual appearance, imaging conditions, and camera viewing angles. We adopt an E-SVMs (Malisiewicz, Gupta, and Efros 2011) based, instance-wised detector scheme. In our observation, the vehicle appearances are highly changed due to the complicated intra-class and imaging variations. Therefore, it is a more practical solution to train instance-wised classifiers with instance-specific metrics to conduct a refined nearest neighbor classification (Malisiewicz, Gupta, and Efros 2011). The detailed formulations of our vehicle detector are given as below:

Training: Given m labeled high-resolution superpixels $\{S_i\}_{i=1}^m$, each of which is described by HoG feature and denoted as f_i. We train its Exemplar SVM with parameters (w_i, b_i) using a randomly sampled negative set N_i from superpixels *without* vehicles. We optimize the following convex objective function:

$$\begin{aligned} \Omega_E(w_i, b_i) = \parallel w \parallel_2 \; + \; C_1 h(w_i^T f_i + b_i) + \\ C_2 \textstyle\sum_{j \in N_i} h(-w_j^T - b_j), \end{aligned} \tag{8}$$

Algorithm 1: Domain-Adaptive Vehicle Detection in Satellite Images by Super Resolution Transfer

Offline:
Input: n superpixels: $X^h = \{x_1, x_2, ..., x_n\}$ with labels: $S = \{s_1, s_2, ..., s_n\}$
Downsampling: Do $Y^l = \Lambda\Theta X^h$ to generate low-res. superpixels: $Y^l = \{y_1, y_2, ..., y_n\}$
Super-Resolution Transfer:
Learn Coupled Dictionary by Equation 4
Do Pre-Detection by $\sum_{i=1}^{n} s_i(W^T F_{s_i}^h + \delta)$
Learn Supervised Dictionary by Equation 5
Instance-Wised Detector Training:
Train each *E-SVM* by Equation 8
Calibrate *E-SVMs* by Hinge Loss learned via structured learning-to-rank using Equation 10
Output: Low- and high-res. dic.s V^l, V^h, and the learnedE-SVMs $\{w_e, b_e\}_{e=1}^n$.
Online:
Given target low-res. superpixel y, do super-resolution to obtain high-res. coeff. u, send to E-SVMs detection in Equation 12.

where $h(x)$ is the Hinge loss, *i.e.*, *Max*(0, 1-*x*). To further guarantee the matching robustness, every superpixel S_i is flipped, translated and rotated to expand to more positive examples for training.

Calibration: After learning all of the E-SVMs with parameters $\{w_i, b_i\}_{i=1}^n$, we further calibrate their outputs. This is achieved by learning a sigmoid function (α_E,β_E) using the validation set, with the calibration test score is:

$$f(x|w_E, \alpha_E, \beta_E) = \tfrac{1}{1+e^{-\alpha_E({w_E}^T x - \beta_E)}}. \tag{9}$$

By thresholding the original SVM score -1 (negative border), the learned parameters of the Sigmoid function can be used to adjust parameters of each detector.

To calibrate, given a held-out validation set with ground truth labels $S_H = \{S_q\}_{q=1}^H$, we first apply the E-SVMs to get prediction scores $\{s_q\}_{q=1}^H$, and then collect SVMs with positive scores for re-ranking (only some of them are with the same label as S_q). For each E-SVM in S, we force superpixels with the same label of S_q to have larger score than others. This can be formulated as a structured learning-to-rank (Joachims 2002):

$$\begin{aligned} &\min \sum_i \frac{1}{2}||w||_2^2 + C \sum \xi_{i,j,k} \\ &s.t. \;\; \forall q_i, w^T \Phi(q_i, s_j) > w^T \Phi(q_i, s_k) + 1 - \xi_{i,j,k} \\ &\forall l(S_j) = l(S_{q_i}), \; l(S_k) \neq l(S_{q_i}), \xi_{i,j,k} \geq 0. \end{aligned} \tag{10}$$

Here $w = \left\{(w_i, b_i)\right\}_{i=1}^n$. $\Phi(q_i, s_j)$'s $(2j-1)$th and $2j$th dimensions are s_j and 1 respectively, which encodes weights and scores into a single vector. It can be learned by solving a cutting plane optimization (Tsochantaridis et al. 2005).

Hard Negative Mining: Different from regular E-SVMs that only finds nearly identical instances, we set a small C_2 in the training process to improve the generality, which allows similar but not exact examples to have positive scores.

Figure 3: Examples of vehicles extracted from high- (left) and low-resolution (right) imageries.

However, this may increase the number of false positives with different labels. To address this problem, given the decision boundary is only determined by the "hard" examples (the support vectors), we introduce a hard negative mining to constrain the decision boundary. We do the following:

1. Train E-SVMs, collect the prediction scores.
2. Add false positives into the negative examples and launch another round of E-SVM's training.

The above two steps are repeated until no new hard examples are found, or reaches a max. iteration.

Online Detection: To determine whether a target region S_q contains vehicle, we find the labeled superpixels with the k strongest responses from their E-SVMs as its k nearest neighbors in S, *i.e.*,

$$kNN(S_q) = \arg \max_{S_i \in S}^{k} F_i\big(\mathbf{w}_i^T x(S_q) + b_i\big), \qquad (11)$$

where F_i is the ranking function which is learned by using a linear function for each SVM to re-rank their output scores, as to minimizing the ranking error among different SVMs, *i.e.*,

$$F_i(x) = w_i^{(r)} \dot{x} + b_i^{(r)}. \qquad (12)$$

This linear mapping are learned in the previous calibration step and does not modify the relative order in each E-SVM, but re-scales and pushes the E-SVMs jointly to make their prediction scores comparable. We summarize the overall procedure of the proposed approach in Algorithm 1. It is worth to note that the proposed framework is general for other object detection tasks (in a setting of low-high resolution transfer) beyond detecting vehicles.

5 Experiments

5.1 Data Set and Ground Truth

We test our algorithm on both satellite and aerial image datasets. To build the satellite image 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. Correspondingly, we further collect 80 corresponding aerial images covering the same road map of New York City by zooming in the Google Earth into the finest resolution. We ask a group of volunteers to manually labeled vehicle regions with both low- and high-resolution images collected above, which produces 1,482 vehicle annotations in total. Figure 3 shows several groups of vehicle

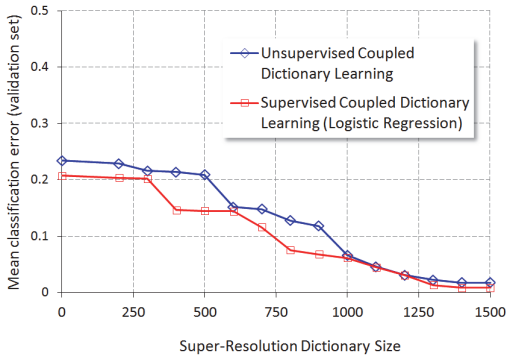

Figure 4: Parameter tuning on the dictionary size.

rectangles, which are variant in visual appearances, imaging conditions, and camera viewing angles.

5.2 Baselines and Evaluation Protocols

From the aerial images, we adopt a 1:5 leave-one- out (training vs. validation) setting for parameter tuning. Note that labels from the low-resolution satellite images are used for validation purpose only. The accuracy of the proposed scheme is tested by using precision-recall curves and mean classification error. We compare the proposed scheme to a set of baselines and state-of-the-art approaches, including: (1) k**NN-source**: It adopts kNN to search most similar superpixels and assigns labels by majority voting, which is operated on the source (low-resolution) domain. (2) **Linear SVM-source**: It adopts SVM with linear kernel for detection and operates on the source domain. (3) k**NN-target**: It adopts kNN to search most similar superpixels and assigns labels by majority voting, which is operated on the target (high-resolution) domain. (4) **Linear SVM-target**: It adopts SVM with linear kernel for detection and operates on the target domain. (5) **E-SVM-target**: It adopts Exemplar-SVMs for detection and operates on the source domain. (6) **E-SVM-target-calibrate**: It adopts Exemplar-SVMs for detection and operates on the source domain. The outputs of instance classifiers are further calibrated using Eq.(9). For all the above approaches, HoG based descriptors (Dalal and Triggs 2005) are adopted to extract features. It is quite clear that, even with the current "high-resolution" vehicle regions, the superpixels are still quite small, which by nature hesitates complex and premature detection models such as Deformable Part-based Model (Felzenszwalb, Girshick, and McAllester 2010) to be deployed.

5.3 Preliminary Settings

In preliminary, both satellite and aerial images are processed by applying a low-pass filter to remove noises. We then extract road maps from both satellite and aerial images. More specifically, a MapInfo/SHP format 2D-vector map and ArcGIS Engine is used to align vector maps and satellite (and aerial) images. Then, the road maps are extracted from images regions that coincide on roads of vector maps. Specially, active shape model (Cootes et al. 1995) is adopted

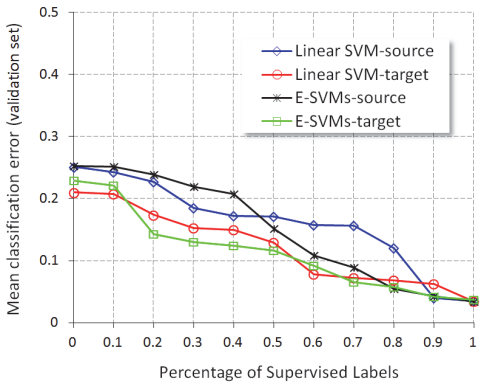

Figure 5: Parameter tuning on the per. of supervised labels.

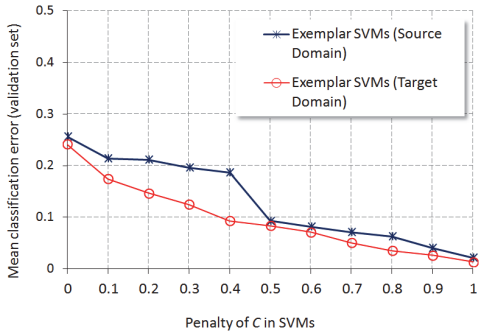

Figure 6: Parameter tuning on the penalty C on E-SVMs.

to extract precise boundaries. Subsequently, superpixel segmentation proposed in (Levinshtein et al. 2009) is adopted to segment the roads into superpixels, on which the vehicle detector is run. We tune a best segmentation scale to ensure the superpixel size is approximately the size of vehicles.

5.4 Parameter Tuning

We have identified and tuned the following parameters that affect the overall performance to tune the best performed detector: (1) *Dictionary size*: The proposed coupled dictionary learning is affected by the dictionary size. As shown in Figure 4, we tune to seek the best size by using the validation set. (2) *Percentage of supervised labels*: The percentage of labels used for supervised dictionary learning affects the final recognition accuracy, as shown in Figure 5. Note that the left end of x-axis 0 refers to the case of unsupervised learning, where significant accuracy drop can be observed. (3) *Penalty C in learning E-SVMs*: We identified that our scheme is robust (while with slight changes) to the setting of the penalty C in E-SVMs learning, as shown in Figure 6.

5.5 Quantitative Analysis

As shown in Figure 7, the Precision-Recall curves have demonstrated that our approach achieves consistent and promising performance comparing to the five baselines as introduced above. Especially, there is a significant performance boost by transferring from the source (low-resolution) domain to the target (high-resolution) domain.

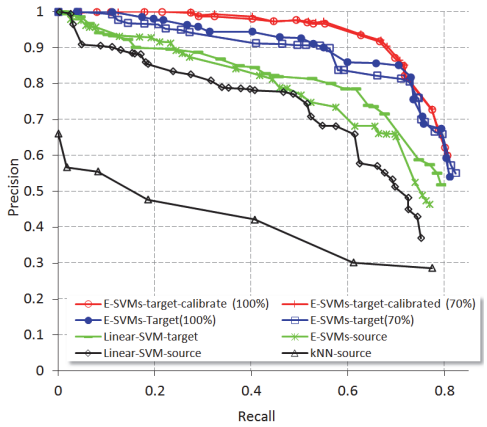

Figure 7: Quantitative Comparisons of PR curves between the proposed algorithm and alternatives.

In addition, another significant performance gain can be clearly observed by replacing kNN and Linear SVM based classifiers with the proposed E-SVMs, which shows the merits of using instance-based classifier for the problem of low-resolution, domain adaptive detection. The proposed learning-to-rank based calibration further push the Precision-Recall curves to the best one as evaluated in our entire experiment.

6 Conclusion

In this paper, we study the problem of vehicle detection in low-resolution satellite images. Two fundamental challenges exist, *i.e.*, limited training examples, as well as the difficulty to train robust detectors directly on the satellite domain. To this end, we contribute in the following aspects: First, we propose to transfer the detection problem into high-resolution aerial images, which is based on a supervised super-solution algorithm with coupled dictionary learning. Second, in the aerial image domain, a robust, instance-wised detection scheme using E-SVMs is proposed, which ensures the vehicle detector to cope with the variations caused by domain transfer. We have tested the proposed scheme extensively with comparisons to a set of existing and state-of-the-art approaches. With quantitative validations, significant performance gains have been reported. In our future work, we will pay attention to handling the multi-scale issue, *i.e.*, the detected windows can be further extended into a pyramid matching setting to ensure finding vehicles from satellite images with various spatial resolutions.

7 Acknowledgement

This work is supported by the Special Fund for Earthquake Research in the Public Interest No.201508025, the Nature Science Foundation of China (No. 61402388, No. 61422210 and No. 61373076), the Fundamental Research Funds for the Central Universities (No. 20720150080 and

No.2013121026), and the CCF-Tencent Open Research Fund.

References

Baker, S., and Kanade, T. 2002. Limits on super-resolution and how to break them. *IEEE Trans. on PAMI* 24(9):1167–1183.

Chen, X.; Xiang, S.; Liu, C.-L.; and Pan, C.-H. 2013. Vehicle detection in satellite images by parallel deep convolutional neural networks. In *IAPR Asian Conf. on Pattern Recognition*, 181–185.

Cheng, H.-Y.; Weng, C.-C.; and Chen, Y.-Y. 2012a. Vehicle detection in aerial surveillance using dynamic bayesian networks. *IEEE Trans. on Image Proc.* 21(4):2152–2159.

Cheng, H.-Y.; Weng, C.-C.; and Chen, Y.-Y. 2012b. Vehicle detection in aerial surveillance using dynamic bayesian networks. *IEEE Trans. on Image Proc.* 21(4):2152–2159.

Choi, J.-Y., and Yang, Y.-K. 2009. Vehicle detection from aerial image using local shape information. *Advances in Image Video Tech.* 227–236.

Cootes, T.; Taylor, C.; Cooper, D.; and Graham, J. 1995. Active shape models - their training and application. *CVPR* 38C59.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. *CVPR* 886–893.

D.Lenhart; Hinz, S.; Leitloff, J.; and Stilla, U. 2008. Automatic traffic monitoring based on aerial image sequences. *PRAI* 18(3):400–405.

Eikvil, L.; Aurdal, L.; and Koren, H. 2009. Classification-based vehicle detection in high resolution satellite images. *ISPRS J. of Photo. and Remote Sensing* 64:65–72.

Felzenszwalb, P.; Girshick, R.; and McAllester, D. 2010. Object detection with discriminatively trained part-based models. *IEEE Trans. on PAMI* 32(9):1627–1645.

Freeman, W.; Thouis, J.; and Egon, P. 2002. Example-based super-resolution. *IEEE Computer Graphics and App.* 22(2):56–65.

Gerhardinger, A.; Ehrlich, D.; and Pesaresi, M. 2005. Vehicles detection from very high resolution satellite imagery. *Int. Archives of Photo. and Remote Sensing* 36(3):W24.

He, X.; Zhou, L.; and Li, J. 2011. Extraction of traffic information in high resolution satellite images. *Urban Geo. Invest. Surveying* 3:49–51.

Hinz, S., and Baumgartner, A. 2011. Vehicle detection in aerial images using generic features, grouping, and context. *Pattern Recognition* 45–52.

Hinz, S. 2004. Detection of vehicles and vehicle queues in high resolution aerial images. *Photo.- Fernerkundung- Geo.* 3(4):201–213.

Holt, A.; Seto, E.; Rivard, T.; and Peng, G. 2009. Object-based detection and classification of vehicles from high-resolution aerial photography. *Photo. Eng. and Remote Sensing* 75(7):871–880.

Jin, X., and Davids, C. 2007. Vehicle detection from high-resolution satellite imagery using morphological shared-weight neural betworks. *IVC* 25:1422–1431.

Joachims, T. 2002. Optimizing search engines using click-through data. In *KDD*, 3133–142.

Kembhavi, A.; Harwood, D.; and Davis, L. 2011. Vehicle detection using partial least squares. *IEEE Trans. on PAMI* 33(6):1250–1265.

Kim, K., and Kwon, Y. 2010. Single-image super-resolution using sparse regression and natural image prior. *IEEE Trans. on PAMI* 32(6):1127–1233.

Kozempel, K., and Reulke, R. 2009. Fast vehicle detection and tracking in aerial image bursts. In *ISPRS City Models, Roads and Traffic*, 175–180.

Kuettel, D.; Guillaumin, M.; and Ferrari, V. 2012. Segmentation propagation in imagenet. In *ECCV*, 459–473.

Leitloff, J.; Hinz, S.; and Stilla, U. 2010. Vehicle detection in very high resolution satellite images of city areas. *IEEE Trans. on Geo. and Remote Sensing* 48:2795–2806.

Levinshtein, A.; Stere, A.; Kutulakos, K.; Fleet, D.; Dickinson, S.; and Siddiqi, K. 2009. Turbopixels: Fast superpixels using geometric flows. *IEEE Trans. on PAMI* 31(2):2290–2297.

Lin, R.; Cao, X.; Xu, Y.; Wu, C.; and Qiao, H. 2008. Airborne moving vehicle detection for video surveillance of urban traffic. *IET Trans. on Compute Vision* 2(1):1–12.

Malisiewicz, T.; Gupta, A.; and Efros, A. 2011. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 89–96.

Mantrawadi, N.; Nijim, M.; and Young, L. 2013. Object identification and classification in a high resolution satellite data using data mining techniques for knowledge extraction. In *IEEE Int. Systems Conf.*, 750–755.

Pan, S., and Yang, Q. 2010. A survey on transfer learning. *IEEE Trans. on Know.e and Data Eng.* 22(10):1345–1359.

Rohrbach, M.; Ebert, S.; and Schiele, B. 2013. Transfer learning in a transductive setting. In *NIPS*, 46–54.

Sharma, G. 2002. Vehicle detection and classification in 1m resolution imagery. In *Master Thesis, Ohio State U.*

Shrivastava, A.; Malisiewicz, T.; Gupta, A.; and Efros, A. 2011. Data-driven visual similarity for cross-domain image matching. In *SigGraph*.

Tsochantaridis, I.; Joachims, T.; Hofmann, T.; and Altun, Y. 2005. Large margin methods for structured and interdependent output variables. *Machine Learning Research* 1453–1484.

Wang, Y.; Ji, R.; and Chang, S.-F. 2013. Label propagation from imagenet to 3d point clouds. In *CVPR*, 3135–3142.

Yang, J.; Wright, J.; Huang, T.; and Ma, Y. 2008. Image super-resolution as sparse representation of raw image patches. In *CVPR*, 1–8.

Yang, J.; Wright, J.; Huang, T.; and Ma, Y. 2010. Image super-resolution via sparse representation. *IEEE Trans. on Image Proc.* 19(11):2861–2873.

Zhao, Z.; Yan, D.; Ng, W.; and Gao, S. 2013. A transfer learning based framework of crowd-selection on twitter. In *KDD*, 1514–1517.

Zheng, H., and Yu, Y. 2006. A morphological neural network approach for vehicle detection from high resolution satellite imagery. In *NIPS*.