Pattern Recognition Letters xxx (xxxx) xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Beyond context: Exploring semantic similarity for small object detection in crowded scenes☆

Yue Xi[a,b], Jiangbin Zheng[b,*], Xiangjian He[b], Wenjing Jia[b], Hanhui Li[b,c], Yefan Xie[a], Mingchen Feng[a], Xiuxiu Li[d]

[a] School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, PR China
[b] Global Big Data Technologies Centre (GBDTC), University of Technology Sydney, Sydney, Australia
[c] School of Computer Science and Information Security, Guilin University Of Electronic Technology, Guilin, PR China
[d] School of Computer Science, Xi'an University of Technology, Xi'an, PR China

ARTICLE INFO

Article history:
Available online xxx

MSC:
41A05
41A10
65D05
65D17

ABSTRACT

Small object detection in crowded scene aims to find those tiny targets with very limited resolution from crowded scenes. Due to very little information available on tiny objects, it is often not suitable to detect them merely based on the information presented inside their bounding boxes, resulting low accuracy. In this paper, we propose to exploit the semantic similarity among all predicted objects' candidates to boost the performance of detectors when handling tiny objects. For this purpose, we construct a pairwise constraint to depict such semantic similarity and propose a new framework based on Discriminative Learning and Graph-Cut techniques. Experiments conducted on three widely used benchmark datasets demonstrate the improvement over the state-of-the-art approaches gained by applying this idea.

1. Introduction

Small object detection in crowded scenes is very common in real world applications, such as face detection for large scale video surveillance, object detection in Remote Sensing Images (RSIs) for Earth Vision£ and small nodule detection and anatomy detection in medical image analysis [1–3]. In recent years, many efficient and accurate face detectors [4–12] have been proposed. These approaches have achieved impressive results on large and medium faces in most constrained scenes. However, their performance on handling tiny faces in crowded scenes is far from satisfactory, as shown in Fig. 1. On the other hand, object detection in RSIs [13–17] has remained as an unsolved problem, not only because of the scarcity of well annotated datasets of objects in RSIs, but also due to the huge variation in the scales of objects on the Earth's surface, as shown in Fig. 6. In this work, we focus on the practical task of Small object detection in Crowded Scenes (SCS), which is to estimate the bounding boxes of physical objects (such as a face, ship, airplane and large or small vehicle) from natural and aerial images with crowded scenes.

Recently, great progress on object detection has been achieved by convolutional neural network (CNN) based methods [18,19] that learn deep features from the regions of interest (RoIs) and perform classification based on the learned representation. However, they usually fail to detect small (*e.g.*, 10 × 10 pixel) and crowded (*e.g.*, 100 instances) objects. The main difficulty of SCS is that each small object contains no more than 100 pixels and lack sufficient details to be distinguished from similar background. For example, it is difficult to distinguish a face from hands, a ship from wave, etc. Another difficulty is that CNN-based detectors use convolutional layers with a stride of 8, 16 or 32 to sample objects. They are too coarse to extract useful features from tiny objects. What is worse, the pooling layers down-sample feature maps and tiny objects' features always disappear after several pooling layers [3]. Therefore, it is challenging for CNN-based methods to efficiently and accurately detect small and crowded objects.

The existing methods for SCS fall into three classes. The first class (e.g., [20]) attempts to use deep learning networks to extract scale-invariant features. However, their performance drops dramatically when targets become too small, because it is very difficult to learn sufficiently discriminative features from targets' poor-quality appearance at a small scale. Another class tries to generate additional information from inside the objects' bounding boxes. For example, the work in [12] demonstrated that interpolating the lowest layer of image pyramid was significantly beneficial for capturing small objects. However, increasing the scale of images often results

☆ **Conflict of Interest:** We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant inancial support for this work that could have influenced its outcome.

* Corresponding authors.
E-mail address: zhengjb@nwpu.edu.cn (J. Zheng).

https://doi.org/10.1016/j.patrec.2019.03.009
0167-8655/© 2019 Elsevier B.V. All rights reserved.

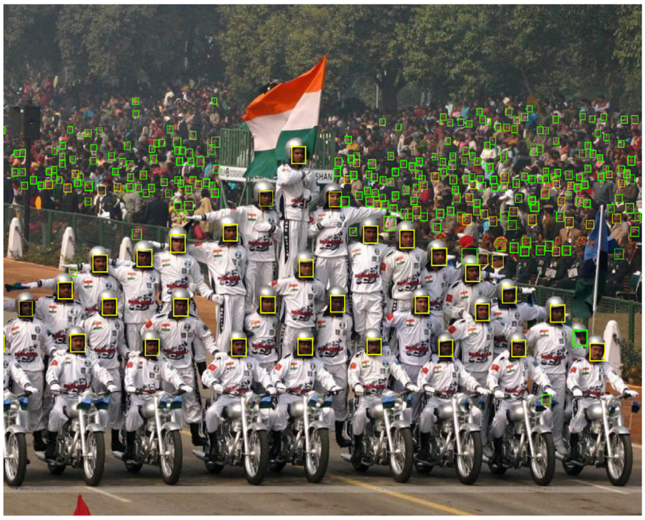


Fig. 1. Tiny faces detected with our proposed approach (shown as yellow and green boxes) and the HR approach [12] (shown as yellow boxes). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in heavy time consumption for both training and testing. The third class (e.g., [21]) seeks to incorporate the information surrounding the objects (i.e., context) to improve the performance of small and crowded object detection. Are there any other ways to improve the performance for SCS? It is commonly accepted that computer vision techniques need additional contextual information to accurately detect targets. However, how to effectively encode context is one of the practical challenges.

Note that existing classification-based detectors typically apply a simple threshold on its classification score to determine whether a candidate is a target or not, as shown in the first stage of Fig. 2. However, obtaining a threshold optimal to various scenarios is often difficult. In this paper, we propose a novel idea to exploit the semantic information (consisting of spatial location, scale and texture) of a candidate's neighbors to classify the candidate as object of interest or background. Specifically, based on such semantic information we try to group candidates containing the same class object into one cluster, while backgrounds are kept far away from the cluster. For this purpose, we propose a Discriminative Learning and Graph-Cut (DLGC) framework, which carries out a further classification on the candidates produced by other object detectors. Fig. 2 illustrates the framework of this idea.

We first obtain a high-recall classifier, which aims to retrieve all the targets in the image, but may unavoidably introduce lots of false positives. Our focus is to retrieve the targets with low classification scores while removing these false positives. In order to do this, we propose a discriminative learning method to learn a similarity matrix to evaluate the similarity of each pair of candidates. A graph model is built to represent the similarity matrix of these candidates. The graph-cut technique is utilized to divide the graph into several groups where candidates in the same group are similar and those in different groups are dissimilar to each other. Finally, the candidates in each group are classified as target or not, correspondingly, by voting.

The main contributions of this paper can be highlighted as follows. First, aiming to boost the detection performance, we propose a novel discriminative learning and graph-cut framework to exploit the semantic information between targeting objects' neighbors. Secondly, to depict a local neighborhood relationship, we introduce a pairwise constraint into tiny face detector to improve the detection accuracy. Thirdly, to realize such pairwise constraint, we convert the problem of regression that estimates the similarity between different candidates into a classification problem which produces scores of classification for each pair of candidates.

2. Related works

In this section we describe the existing works relevant to our approach discussed in this paper. They are organized and presented in three main aspects, i.e., small and crowded object detection, context in object detection, and discriminative learning for semantic similarity.

2.1. Small and crowded object detection

As a special case of small and crowded object detection, tiny face detection has attracted more and more interest. In face detection, since the pioneer work of [11] who designed a cascade of weak classifiers using Haar features and AdaBoost for fast and robust face detection, numerous approaches have been developed to improve the performance with more sophisticated hand-crafted features [22] and more powerful classifiers [23]. However, these methods just used non-robust hand-crafted features and optimized each component separately, and resulted in sub-optimal face detection frameworks. Recently, face detectors based on CNNs [24–26] have greatly bridged the gap between human vision and artificial detectors.

On the one hand, object detection in RSIs is a fundamental but challenging problem in Earth Vision. Various methods [13–17] have been proposed to address the task, including the classical methods such as [16,17]. Recently, some researchers have attempted to transfer deep-learning-based detection algorithms developed for natural scenes to RSIs [13–15]. G.-S. Xia et al. [13] introduced the dataset DOTA, which is the largest annotated object dataset with a variety of categories in RSIs. They also benchmarked deep-learning-based detection algorithms on DOTA as the baseline. K. Li et al. [14] proposed an end-to-end deep-learning-based detection framework. The framework combines the region proposal network and the contextual feature fusion network to handle the problem of object rotation variations and appearance ambiguity. Gong [15] designed a rotation-invariant Fisher discriminative CNN model to handle the problems of object rotation and between-class similarity. Different from their work, we focus on many small object instances crowded in the scene, such as in aerial images. SCS aims to detect a large number of small objects (*e.g.* small face, aircraft, ship, large or small vehicle) in crowded scenes. It is totally different from general object detection, because the cues for detecting a 3-pixel object are fundamentally different from those for detecting a 300-pixel object [12]. Li et al. [27] proposed a perceptual Generative Adversarial Networks to generate the super-resolved representation of a small object, which is similar to those of large objects in order to boost small object detection performance. Merely in the aspect of applications, the most related previous research to our work is [28], which presented a counting-detection framework with the combination of density map estimation and segmentation. While the images were crowded with small objects, the background is sample so that is effectively to estimate its density map. As we all know, it is not sufficient to detect small objects merely by extracting deep learning features from the texture inside the object region. One main drawback is that, these approaches have neglected local semantic information. We have observed that there exists local coherent relationships in terms of spatial location, scale, and texture in high-density tiny face detection, regardless of the viewpoints. For example, as shown in Fig. 1, face bounding boxes close to each other are similar in their scales and textures. Local semantic information helps tiny face detectors better eliminate false positive.

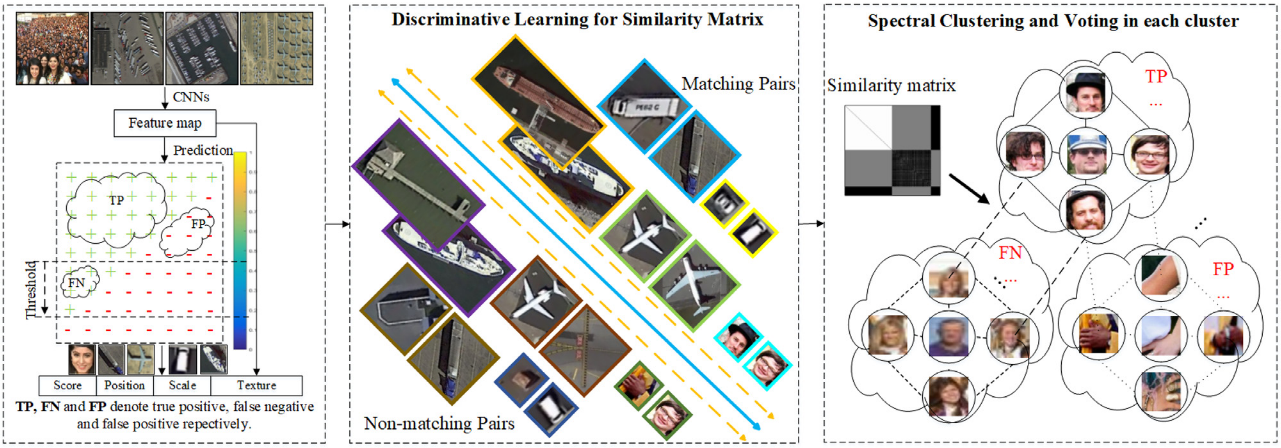

CNNs
Feature map
Prediction
TP
FP
FN
Threshold
Score
Position
Scale
Texture
TP, FN and FP denote true positive, false negative and false positive repectively.
Discriminative Learning for Similarity Matrix
Matching Pairs
Non-matching Pairs
Spectral Clustering and Voting in each cluster
Similarity matrix
TP
...
FN
...
FP
...

Fig. 2. The framework of our proposed DLGC. In the stage for discriminative learning, each candidate pair is shown as boxes with same colors each pair. Some pairs are true matches (top right), while other are false matches (bottom left). The green or yellow arrows pointing left and right means a max-margin separating hyperplane. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2. Context in object detection

Context is an effective cue for humans to detect small instance in crowded scene [12,29–32]. C. Zhu et al. [29] firstly introduced body context information to the Faster R-CNN object detector [33], which was the key to finding small instances. S. Bell et al. [30] presented the Inside-Outside Net (ION) to model context outside the region of interest and showed improvements on small object detection. Very recently, P. Hu et al. [12] designed a foveal descriptor that captured both coarse context and high-resolution image features in order to effectively encode context information. The descriptors has achieved state-of-the-art performance on the WIDER FACE dataset. C. Chen et al. [31] incorporated context information by concatenating deep learning features of objects and context region to boost small object detection performance. However, when small objects are crowded together, the context region may contain small objects. X. Tang et al. [32] designed a novel context anchor, named PyramidAnchor, to supervise a face detector to learn features from contextual parts around a face.

2.3. Discriminative learning for semantic similarity

Discriminative learning [34–38] for semantic similarity usually takes a pair of images as input and output a similarity score by calculating the distance between the images' features. In 1993, [34] first used discriminative learning to compute the similarity between a pair of images for signature verification. S. Cao et al. [35] proposed an SVM-based discriminative learning model to predict matching and non-matching image pairs for large-scale image matching. [38] used triplet samples for training the network which considered the images from the sample place and different places for place recognition. D. Yi et al. [36] split an image into three horizontal parts and trained three part-CNNs to extract features. The semantic similarity of two images was computed by the distance among those three features. Similarly, [37] combined CNN with some gate functions, which aimed to adaptively focus on the similar parts of input image pairs. Our problem is very different from theirs in the way that, the targets in our problem are very difficult to be split into parts, due to their very tiny size. To introduce local coherent relationships, our proposed method introduces a discriminative component to represent this coherence and uses the graph-cut algorithm to divide candidates into several groups, where candidates in the same group are similar, and dissimilar when they are in different groups.

Table 1
The Average Precision of Classification on the DOTA Validation set.

Methods	mAP	plane	small vehicle	large vehicle	ship
DLGC	64.37	92.58	47.49	66.51	50.88
Faster R-CNN	61.93	90.21	43.91	64.72	48.91

3. The proposed method

Our goal is to integrate the local coherent relationship into tiny face detection. To represent such a local coherent relationship, we define a pairwise constraint, which contains the equivalence constraint for pairs of data points belonging to the same classes, and the inequivalence constraint for pairs of data points belonging to different classes. To better estimate similarity of different candidate pairs, we use a max-margin separating hyperplane to learn a weighting of different features.

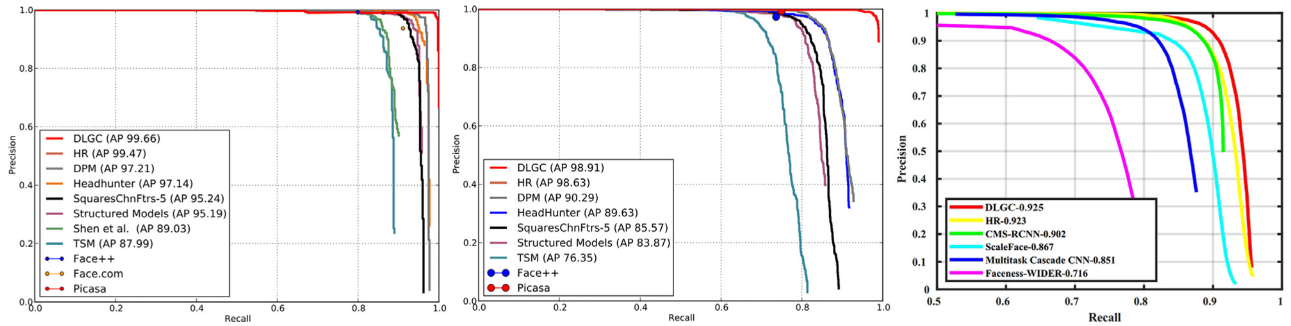
As shown in Fig. 2, we propose a DLGC framework for high-density tiny object detection. We first use a linear-SVM to estimate adjacency matrix among all candidates (Section 3.1) and then we construct a graph model and use the graph-cut algorithm to divide candidates into several groups (Section 3.2). Finally, we design a voting method to classify groups (Section 3.2).

3.1. Discriminative learning based on linear-SVM

Let $X = \{x_1, x_2, \ldots, x_N\}$ denote the set of the N candidates (i.e., face or non-face bounding boxes). To introduce the pairwise constraints, we first build a similarity matrix $S = s(x_i, x_j), x_i, x_j \in X, i, j = 1, 2, \ldots N$, where $s(x_i, x_j)$ represents the similarity between x_i and x_j. $s(x_i, x_j) = 1$ means that x_i has a strong resemblance of x_j, and $s(x_i, x_j) = 0$ means that x_i is completely different from x_j.

In order to obtain the similarity score between two candidates x_i and x_j, it is difficult to evaluate their similarity through regression, because it is difficult to label their ground truth similarity. Therefore we convert the problem into an unsupervised, binary-classification problem, namely whether they are similar or not. We use SVM to compute the similarity score between two candidates x_i and x_j based on multiple cues, i.e., the positions, scales, classification scores and deep features of the candidates, which are concatenated together into a feature vector $\phi(x_i)$. Note that classification scores and deep features of a candidate x_i are obtained from the tiny face detector [12]. During the training stage, we sort X by their scores in descending order. We suppose that X_{Top} denotes the top 10% of X which are face patches, while

Please cite this article as: Y. Xi, J. Zheng and X. He et al., Beyond context: Exploring semantic similarity for small object detection in crowded scenes, Pattern Recognition Letters, https://doi.org/10.1016/j.patrec.2019.03.009

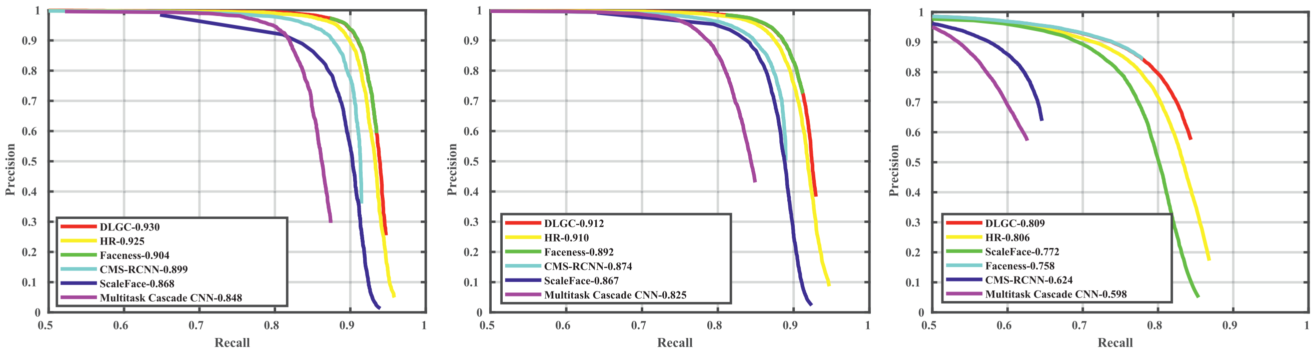


(a) PR curves on AFW dataset

(b) PR curves on PASCAL FACE dataset

(c) PR curves on “easy” subset of WIDER FACE dataset

Fig. 3. The precision-recall (PR) curves obtained using our proposed DLGC approach and the state of the arts.



(a) Val:Easy

(b) Val:Med

(c) Val:Hard

Fig. 4. The PR curves obtained on WIDER FACE validation set using our proposed DLGC approach and the state of the arts.

Table 2
The Average Precision of Classification on DOTA Testing set.

Methods	mAP	plane	small vehicle	large vehicle	ship
DLGC	61.73	83.59	57.89	54.31	51.14
Faster R-CNN	59.12	80.32	53.66	52.49	50.04
R-FCN	56.27	81.01	49.77	45.04	49.29
YOLO-V2	50.01	76.9	38.73	32.02	52.37
SSD	29.89	57.85	0.05	36.93	24.74

X_{Bottom} denotes the bottom 10% of X which are non-face patches. As shown in Fig. 2, in Stage 2 of our DLGC, we build a training set $\{(x'_{1,1}, y'_{1,1}), (x'_{1,2}, y'_{1,2}), \ldots (x'_{n,n}, y'_{n,n})\}$, $x'_{i,j} = \phi(x_i) - \phi(x_j)$, $y'_{i,j} = \{0, 1\}$. If $x_i, x_j \in X_{Top}, y'_{i,j} = 1$. If $x_i \in X_{Top}, x_j \in X_{Bottom}, y'_{i,j} = 0$. During the testing stage, we feed $x'_{i,j} = \phi(x_i) - \phi(x_j)$ to the SVM classifier, and then use the output score as the similarity score $s(x_i, x_j)$ between x_i and x_j. Thus, we build the similarity matrix S.

3.2. Graph-cut based on spectral clustering

Given a set of candidates $X = \{x_1, x_2, \ldots, x_N\}$ and a similarity matrix S, our goal is to cluster X into different groups where candidates in the same group are similar and dissimilar when they are in different groups. In this work, we adopt the graph-cut algorithm for this purpose. First, we build a graph model $G = (V, E)$ to represent X, where each vertex $v_i \in V$ represents a candidate x_i, and $e_{ij} \in E$ represents the similarity $s(x_i, x_j)$ between the corresponding candidates x_i and x_j. Now the problem can be reformulated with the graph model, i.e., we want to find a partition of the graph so that the weights of edges between different subgraphs are very low (indicating that points in different clusters are dissimilar from each other) and the weights of edges in the same group are very high (meaning that points within the same cluster are similar to each other). Formally,

$$cut(A_1, A_2, \ldots, A_k) = \frac{1}{2}\sum_{i=1}^{k} W(A_i, \bar{A}_i) \tag{1}$$

where $A_i \subset V, A_i \cap A_j = \emptyset$ and $A_1 \cup A_2 \cup \ldots \cup A_k = V$, $W(A_i, \bar{A}_i) = \sum_{m \in A_i, n \in \bar{A}_i} w_{mn}$, $w_{mn} = exp(-S_{mn}/2\delta^2)$ used to boost local neighborhood relationships.

However, the solution simply separates one individual vertex from the rest of the graph. To avoid unbalanced graph-cut situation that there is a large difference in sizes of subgraphs, we introduce the size of subgraph $|A|$ which is the number of vertexes in A to ensure the set of subgraph $\{A_1, A_2, \ldots, A_k\}$ is reasonably large. Therefore, we can transform Eq. (1) as follows:

$$cut(A_1, A_2, \ldots, A_k) = \frac{1}{2}\sum_{i=1}^{k} \frac{W(A_i, \bar{A}_i)}{|A_i|} \tag{2}$$

According to [39],

$$\arg\min cut(A_1, A_2, \ldots, A_k) = \arg\min_{H} Tr(H^T L H) \tag{3}$$

where L is the Laplacian matrix, $H^T H = I$, and the indicator $H = \{h_1, h_2, \ldots, h_k\}$ is

$$h_{i,j} = \begin{cases} \frac{1}{\sqrt{A_j}} & \text{if } v_i \in A_j \\ 0 & otherwise \end{cases} \tag{4}$$

where $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, k$. Eq. (3) is the standard form of a trace minimization problem. According to the Rayleigh–Ritz theorem [40], the solution is given by choosing the matrix U which contains the first k eigenvectors of L and then uses the k-means algorithm on U. Therefore, we manage to cluster X into k groups $\{A_1, A_2, \ldots, A_k\}$. Finally, the candidates in each group are classified as faces or non-faces using voting.

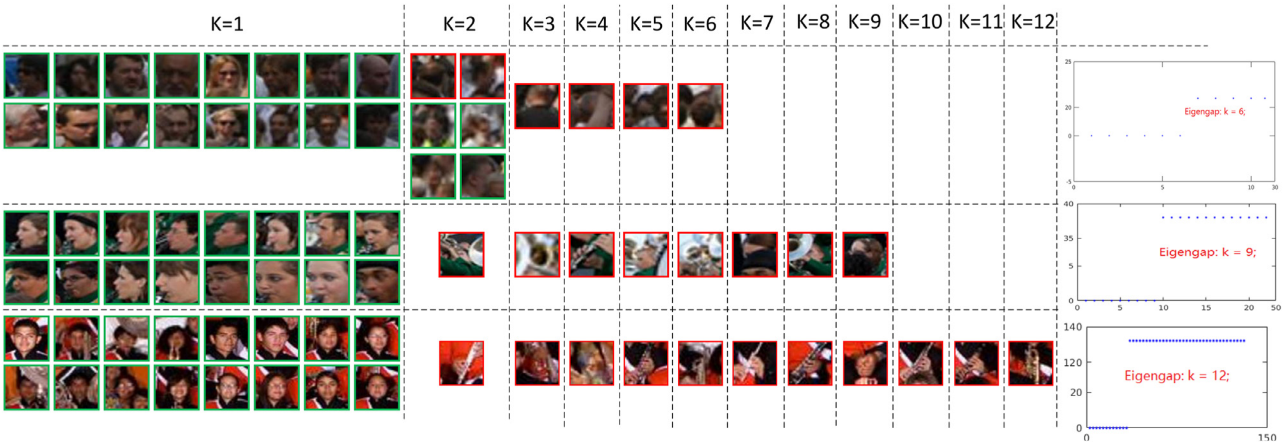
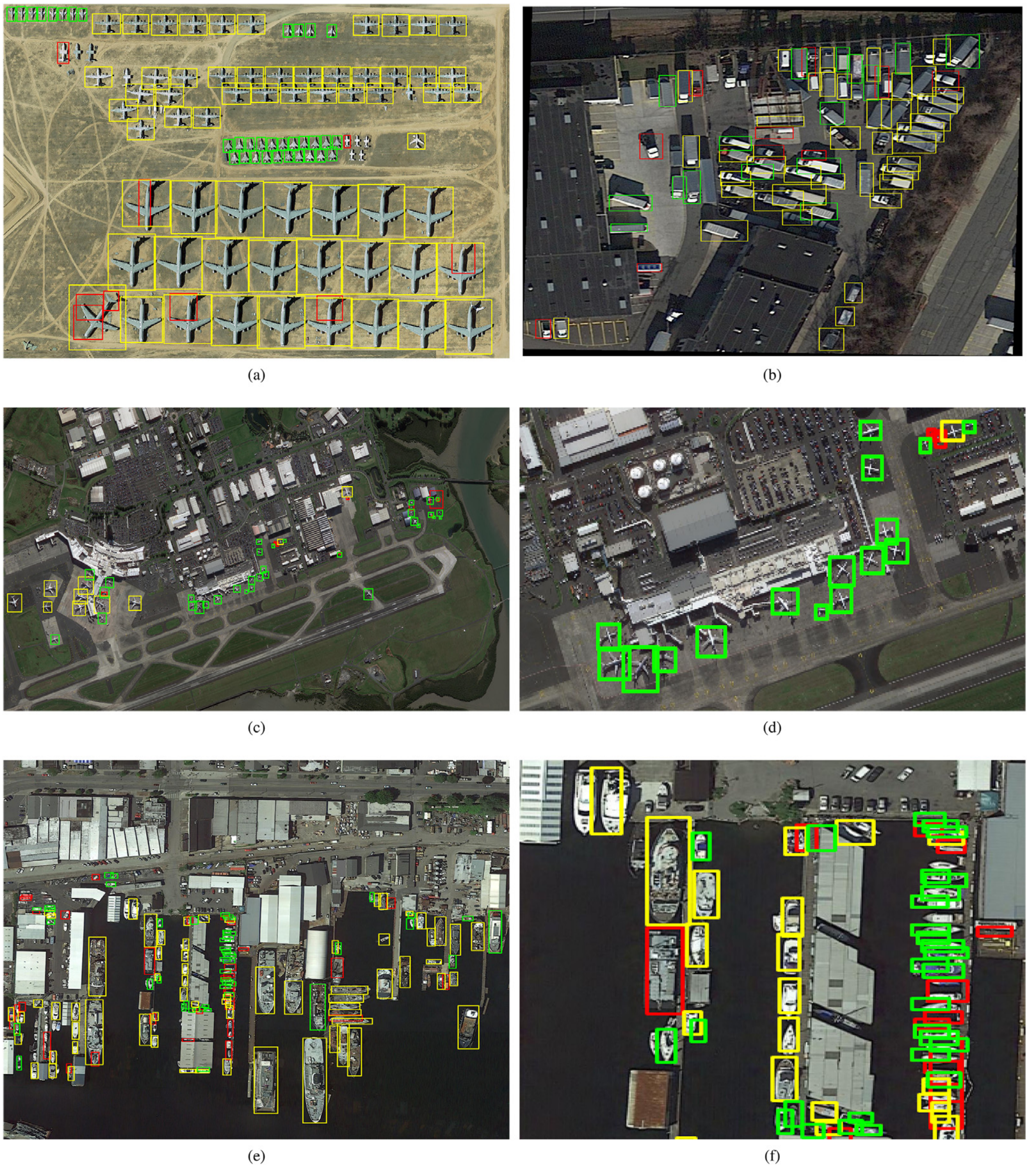


Fig. 5. Qualitative results of spectral clustering given similarity matrix between different candidates. Faces (shown as green rectangle) and background regions (shown as red rectangle) are clustered into different groups. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)


(a)

(b)

(c)

(d)

(e)

(f)

Fig. 6. Qualitative results of object detection with our method in RSIs (shown as yellow and green boxes) and Faster RCNN (shown as yellow boxes). Red boxes are introduced by reducing classification threshold and removed by our method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Experiments

In this section, we first introduce our experimental settings. We then present three groups of experiments to evaluate the performance of our method with comparisons to several baseline methods. We finally discuss the effectiveness of our proposed Discriminative Learning and Graph-Cut method.

4.1. Experimental settings

4.1.1. Datasets and evaluation metrics

To make a fair comparison with the state-of-the-art methods in object detection [12,20,29,41–51], we evaluate our method on three face detection benchmark datasets including WIDER FACE [52], Annotated Faces in the Wild (AFW) [48], Pascal Faces [46], and object detection in RSIs (DOTA) [13], which is the latest, largest and most difficult dataset for object detection in remote sensing imageries. In WIDER FACE, small faces of which the height ranges from 10 to 50 pixels account for about 50% of the data set. Similarly, in the DOTA dataset, small targets of which the height ranges from 10 to 50 pixels account for 57% of the data set. In terms of performance measurement, for the ease of comparison, in face detection we use the standard evaluation metrics precision-recall curves, whereas in object detection in RSIs, we use standard evaluation metrics Averages Precision (AP) and mean of AP (mAP) for evaluation, same as in the corresponding literature.

4.1.2. Implementation details

Our code is based on the publicly available finding tiny face framework [12] built on the Matconvnet [53] and Faster R-CNN framework [50,54] built on the Mxnet [55]. In order to extract texture features of candidates, we feed the result patches of the original object detection into original network and Faster R-CNN framework individually and use feature of res4b22x in resnet-101. The model is trained on a NVIDIA Quadro P2000, and Inter Core i7-7700HQ CPU @ 2.8 GHz.

4.2. Performance comparison

4.2.1. Results obtained on the WIDER FACE dataset

The WIDER FACE Dataset is the most challenging public face datasets due to the variety of face scales and occlusion. It contains 32,203 images split into training (40%), validation (10%) and testing (50%) set. The validation set and testing set are divided into “easy”, “medium”, and “hard” subsets according to the difficulties of the detection. Both images and annotations of the training and validation set can be available online[1], but annotations of its testing set are not released yet. So we should send our results to the database server for receiving precision-recall curves.

We compare our DLGC with the HR, Faceness, ScaleFace [42], CMS-RCNN [29] and Multitask Cascade CNN [43]. The PR curves on the testing set is presented in Fig. 3c, and our method outperforms HR by 0.2% in “easy” subset. The PR curves on the validation set is presented in Fig. 4 and our method outperforms the HR by 0.5%, 0.2%, 0.3%, in “easy”, “medium” and “hard” subsets respectively.

4.2.2. Results obtained on the AFW and PASCAL FACE dataset

The AFW dataset has 205 images containing in total 473 labelled faces. We evaluate our model against the HR, DPM [44], Headhunter, SquaresChnFtrs [45], Structured Models [46], Shen et al. [47], TSM [48] and commercial detectors (e.g., Face.com, Face++ and Picasa). As illustrated in Fig. 3a, our DLGC outperforms all other detectors on precision-recall (PR) curves.

The PASCAL FACE dataset contains 1,335 labeled faces in 851 images, which are collected from PASCAL person layout subset. Because this paper focuses on face detection, we ignore images without persons from the original dataset, similar like DPM [44]. We also evaluate our model against the HR, DPM [44], Headhunter, SquaresChnFtrs [45], Structured Models [46], Shen et al. [47], TSM [48] and commercial detectors (e.g., Face++ and Picasa). As shown in Fig. 3b, our DLGC outperforms all other detectors on PR curves.

4.2.3. Results obtained on the DOTA dataset

DOTA [13] contains 2806 aerial images, which are of the size in the range from about 800 × 800 to 10,000 × 10,000 pixels and contains 188,282 instances exhibiting a wide variety of scales, orientations, and shapes. Our experiments are conducted on four categories of objects in this dataset which contain mostly small, crowded objects, i.e., plane, ship, large and small vehicle.

We evaluate our model against the state-of-the-art object detection method Faster R-CNN [49], R-FCN[2] [50], YOLOv2[3] [51], SSD[4] [20]. The mAP and AP of each category on the validation set is presented in Table 1 and our method exceeds the base model by 2.75%. The mAP and AP of each category on the test set is presented in Table 2 and our method exceeds the base model by 2.61%. The improvement in AP of the plane and small vehicle category (more than 2%) is more than that of the ship and large vehicle category (less than 2%). This is because in crowded scenes, using the horizontal rectangle annotation, large vehicle and ships tend to introduce higher degree of overlapping compared to plane and small vehicles, as shown in Fig. 6.

Fig. 6 shows the effectiveness of our method. As it can be seen from the figure, it is difficult for the base model to detect small targets, but we have nearly recalled all the small planes (shown as blue boxes) and removed all false alarms (shown as red boxes).

4.3. Model analysis

We further analyze the effectiveness of the Discriminative Learning and Graph-cut of the proposed method and conducted more experimental analysis on The WIDER FACE Dataset.

4.3.1. The effectiveness of discriminative learning

We verify the effectiveness of estimating similarity of candidate pairs using our Discriminative Learning approach. Because the number of features 8202-D is large, we do not need to map data to a higher dimensional space. Therefore we utilize the LIBLINEAR library [56] with L2-regularized logistic regression and its parameter c 0.0625 to implement the Discriminative learning.The SVM for binary pairwise relationship classification is trained by our building dataset. To be specific, the training set consists of 10,500 pairs sampled from 50 images with crowded faces, 210 pairs generated from 15 faces and 6 background regions from those candidates by the famous Tiny face detector [12] per image. The test set similar to train set consists of 5,250 pairs. We use accuracy to evaluate the classification performance of trained SVMs. Its accuracy achieved 78.24% on the test set. Conclusion can be drawn that our discriminative learning method can effectively estimate the similarity between two candidates. The Average Precision achieved on WIDER FACE validation set with different confidence values for selecting candidate pairs is shown in Table 4. Consistent with our analysis, a confident value of 10% optimises the performance. So, we chose 10% in our experiments.

[1] http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/.

[2] https://github.com/daijifeng001/R-FCN.

[3] https://pjreddie.com/darknet/yolo/.

[4] https://github.com/weiliu89/caffe/tree/ssd.

Table 3
The accuracy of classification based on SVM.

Number of training samples	Accuracy on testing set
1,000	63.14%
2,000	69.70%
2,500	71.65%
5,000	75.66%
10,500	**78.24%**
20,000	78.60%

Table 4
The Average Precision of Face Detection on WIDER FACE validation set with difference confident values.

confident values	Easy	Med	Hard
5%	89.5%	88.2%	77.4%
10%	**93.0%**	**91.2%**	**80.9%**
15%	91.5%	90.5%	78.4%
20%	88.7%	88.4%	74.5%
25%	86.9%	86.9%	72.9%
30%	79.5%	81.7%	69.1%

Table 5
The accuracy of classification based on neural network.

Number of hidden layer	Accuracy on testing set
100	65.16%
500	66.41%
1,000	64.98%
2,000	67.34%

In addition, we also use neural network to estimate similarity between candidate pairs. The neural network is trained for 2000 epoches. The learning rate of network is set as 10^{-2}. The loss goal is set as 10^{-3}. The number of nodes in the first layer is 8,202-D. The number of nodes in hidden layer ranges from 100 to 2,000. The test results are as shown in Fig. 5.

Results: It is obvious that, compared with the NN, our Discriminative learning approach based on SVM has improved the accuracy by 10.90%.

4.3.2. The effectiveness of Graph-cut

In order to verify the effectiveness of spectral clustering, we evaluate the performance of face candidate clustering results with different number of clusters. We select about 50 images with crowded faces from the dataset. Table 3. To reduce the noise arose from similarity matrix, we regenerate a new one with ground truth and candidates as the input. Nearly all of faces are clustered into one group, as shown in Fig. 5. Note that intuitively all the background regions seem to have been clustered into one group, but essentially they are dissimilar with each other. So they will be clustered into several groups. We set the number of cluster as 4 and clustering results of candidates in one images is in a line. It shows that our spectral clustering can divide candidates into different groups, where candidates in the same group are similar and dissimilar when they are in different groups. All of the candidates in cluster 1 are faces. Note that there are several faces are grouped into groups which background regions lie in, because low resolution and background clutter lead that they are more similar to background regions.

The choice of the number of clusters is a general problem. We use the eigengap heuristic to choose the number k of clusters such that all eigenvalues $\lambda_1, \ldots, \lambda_k$ are very small, but λ_{k+1} is relatively large. According to perturbation theory, the eigenvalue 0 has multiplicity k, and there is a gap to the (k + 1)th eigenvalue $\lambda_{k+1} > 0$. For the cases in the first row of Fig. 5, the first six eigenvalues are approximately 0. Then, there is a gap between the 6th and 7th, which is $|\lambda_7 - \lambda_6|$ and is relatively large. According to the eigengap heuristic, this gap indicates that the data set contains six clusters. The same also applies to the results of the other two images (plotted in Fig. 5).

5. Conclusion

In this paper, aiming to improve the performance for SCS, we have proposed a novel idea to exploit the semantic similarity between targeting objects' neighbors and create a pairwise constraints to depict such semantic similarity. Then, a framework which adopts the discriminative learning and graph-cut techniques is formulated to boost the accuracy of existing object detectors. Experiments conducted on four widely-used benchmark datasets for target detection have demonstrated the improvement over the state-of-the-art by applying this idea. We have also verified the effectiveness of discriminative learning for estimating candidates' similarity and visualize graph-cut's result for clustering candidates. The mechanism of our proposed framework is generic indicating that it has a great potential being applied on other small and generic object detectors.

Acknowledgments

This work was supported by Rail Manufacturing CRC and Sydney Trains with University of Technology Sydney (Project ID PRO17-3968), HJSW and Research Development plan of Shaanxi Province (Project ID 2017ZDXM-GY-094, 2015KTZDGY04-01), National Natural Science Foundation of China (Project ID 61502382) and Chinese Scholarship Council.

References

[1] W. Zhu, C. Liu, W. Fan, X. Xie, Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification, in: Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on, IEEE, 2018, pp. 673–681.
[2] W. Zhu, Q. Lou, Y.S. Vang, X. Xie, Deep multi-instance networks with sparse label assignment for whole mammogram classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 603–611.
[3] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, X. Xie, Anatomynet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy, Med. Phys. (2018).
[4] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.
[5] X. Zhu, Z. Lei, X. Liu, H. Shi, S.Z. Li, Face alignment across large poses: A 3d solution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 146–155.
[6] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al., Deep face recognition., in: BMVC, vol. 1, 2015, p. 6.
[7] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[8] X. Zhu, Z. Lei, J. Yan, D. Yi, S.Z. Li, High-fidelity pose and expression normalization for face recognition in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
[9] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Advances in neural information processing systems, 2014, pp. 1988–1996.
[10] M. Kim, S. Kumar, V. Pavlovic, H. Rowley, Face tracking and recognition with visual constraints in real-world videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2008, pp. 1–8.
[11] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[12] P. Hu, D. Ramanan, Finding tiny faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1522–1530.
[13] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, Dota: A large-scale dataset for object detection in aerial images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[14] K. Li, G. Cheng, S. Bu, X. You, Rotation-insensitive and context-augmented object detection in remote sensing images, IEEE Trans. Geosci. Remote Sens. 56 (4) (2018) 2337–2348.
[15] G. Cheng, P. Zhou, J. Han, Rifd-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection, in: Proceedings of

Please cite this article as: Y. Xi, J. Zheng and X. He et al., Beyond context: Exploring semantic similarity for small object detection in crowded scenes, Pattern Recognition Letters, https://doi.org/10.1016/j.patrec.2019.03.009

the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 2884–2893.

[16] J. Han, D. Zhang, G. Cheng, L. Guo, J. Ren, Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning, IEEE Trans. Geosci. Remote Sens. 53 (6) (2015) 3325–3337.

[17] G. Cheng, J. Han, P. Zhou, L. Guo, Multi-class geospatial object detection and geographic image classification based on collection of part detectors, ISPRS J. Photogramm. Remote Sens. 98 (2014) 119–132.

[18] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[19] G. Gkioxari, R. Girshick, P. Dollár, K. He, Detecting and recognizing human-object interactions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018).

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.

[21] Q.V. Le, N. Jaitly, G.E. Hinton, A simple way to initialize recurrent networks of rectified linear units, arXiv:1504.00941 (2015).

[22] B. Yang, J. Yan, Z. Lei, S.Z. Li, Aggregate channel features for multi-view face detection, in: IEEE International Joint Conference on Biometrics,

[23] M.T. Pham, T.J. Chain, Fast training and selection of haar features using statistics in boosting-based face detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–7.

[24] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325–5334.

[25] B. Yang, J. Yan, Z. Lei, S.Z. Li, Convolutional channel features, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, 2015, pp. 82–90.

[26] S. Yang, P. Luo, C.-C. Loy, X. Tang, From facial parts responses to face detection: A deep learning approach, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3676–3684.

[27] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan, Perceptual generative adversarial networks for small object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1222–1230.

[28] Z. Ma, L. Yu, A.B. Chan, Small instance detection by integer programming on object density maps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3689–3697.

[29] C. Zhu, Y. Zheng, K. Luu, M. Savvides, Cms-rcnn: Contextual Multi-scale Region-based Cnn for Unconstrained Face Detection, in: Deep Learning for Biometrics, 2017, pp. 57–79.

[30] S. Bell, C. Lawrence Zitnick, K. Bala, R. Girshick, Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2874–2883.

[31] C. Chen, M.-Y. Liu, O. Tuzel, J. Xiao, R-cnn for small object detection, in: Asian Conference on Computer Vision, Springer, 2016, pp. 214–230.

[32] X. Tang, D.K. Du, Z. He, J. Liu, Pyramidbox: A context-assisted single shot face detector, arXiv:1803.07737 (2018).

[33] R. Girshick, Fast r-cnn, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[34] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a “siamese” time delay neural network, in: Advances in Neural Information Processing Systems, 1994, pp. 737–744.

[35] S. Cao, N. Snavely, Learning to match images in large-scale collections, in: European Conference on Computer Vision, 2012, pp. 259–270.

[36] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, 2014, pp. 34–39.

[37] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: European Conference on Computer Vision, Springer, 2016, pp. 791–808.

[38] M.A. Uy, G.H. Lee, Pointnetvlad: deep point cloud based retrieval for large-scale place recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018).

[39] U. Von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.

[40] H. Lutkepohl, Handbook of matrices., Comput. Stat. Data Anal. (1997).

[41] Z. Cai, Q. Fan, R.S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: European Conference on Computer Vision, 2016, pp. 354–370.

[42] S. Yang, Y. Xiong, C.C. Loy, X. Tang, Face detection through scale-friendly deep convolutional networks, arXiv:1706.02863 (2017).

[43] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process Lett. 23 (10) (2016) 1499–1503.

[44] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

[45] M. Mathias, R. Benenson, M. Pedersoli, L. Van Gool, Face detection without bells and whistles, in: European Conference on Computer Vision, Springer, 2014, pp. 720–735.

[46] J. Yan, X. Zhang, Z. Lei, S.Z. Li, Face detection by structural models, Image Vis. Comput. 32 (10) (2014) 790–799.

[47] X. Shen, Z. Lin, J. Brandt, Y. Wu, Detecting and aligning faces by image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3460–3467.

[48] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879–2886.

[49] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[50] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in neural information processing systems, 2016, pp. 379–387.

[51] J. Redmon, A. Farhadi, Yolo9000: Better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 6517–6525.

[52] S. Yang, P. Luo, C.-C. Loy, X. Tang, Wider face: A face detection benchmark, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5525–5533.

[53] A. Vedaldi, K. Lenc, Matconvnet: Convolutional neural networks for Matlab, in: Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 689–692.

[54] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 764–773.

[55] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, Z. Zhang, Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems, arXiv:1512.01274 (2015).

[56] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (Aug) (2008) 1871–1874.

Please cite this article as: Y. Xi, J. Zheng and X. He et al., Beyond context: Exploring semantic similarity for small object detection in crowded scenes, Pattern Recognition Letters, https://doi.org/10.1016/j.patrec.2019.03.009