BEYOND CONTEXT: EXPLORING SEMANTIC SIMILARITY FOR TINY FACE DETECTION

Yue Xi[⋆†] *Jiangbin Zheng*[⋆] *Xiangjian He*[†] *Wenjing Jia*[†] *Hanhui Li*[‡]

[⋆] School of Computer Science and Engineering, Northwestern Polytechnical University, P.R.China
[†]Global Big Data Technologies Centre (GBDTC), University of Technology Sydney, Australia
[‡]School of Data and Computer Science, Sun Yat-sen University, P.R.China

ABSTRACT

Tiny face detection aims to find faces with high degrees of variability in scale, resolution and occlusion in cluttered scenes. Due to the very little information available on tiny faces, it is not sufficient to detect them merely based on the information presented inside the tiny bounding boxes or their context. In this paper, we propose to exploit the semantic similarity among all predicted targets in each image to boost current face detectors. To this end, we present a novel framework to model semantic similarity as pairwise constraints within the metric learning scheme, and then refine our predictions with the semantic similarity by utilizing the graph cut techniques. Experiments conducted on three widely-used benchmark datasets have demonstrated the improvement over the-state-of-the-arts gained by applying this idea.

Index Terms— Tiny face detection, semantic information, metric learning, graph-cut

1. INTRODUCTION

Robust face detection is one of the ultimate components to support various facial related problems, such as face alignment [1][2], face recognition [3][4][5], face verification [6], and face tracking [7], etc. From the cornerstone by Viola-Jones [8] to the recent work by Hu et al. [9], the performance of face detection has been improved dramatically. The recent introduction of the WIDER face dataset [10], which contains a large number of small faces, exposes the performance gap between humans and the current face detection techniques due to a number of challenges in practice. Different from the classical face detection, tiny face detection mainly focuses on low-resolution, large scale variation and serious occlusion, as shown in Fig. 1. All of these challenges suggest the information on small objects is far too limited.

The existing methods for finding small objects in imageries can be grouped into three categories. The first group (e.g., [11]) aims to extract scale-invariant features using pre-trained deep networks. However, their performance drops dramatically as the target faces become too small. Another group tries to generate additional information inside the objects by interpolation. For example, the work in [9] demonstrated that interpolating the lowest layer of image pyramid was significantly beneficial for capturing small objects. The last group (e.g., [12]) seeks to incorporate information surrounding the objects (i.e., context) in order to improve the performance of tiny face detection. It is clear that computer vision needs additional contextual information to accurately classify small faces. Is there another way to improve the performance of small object detection?

Fig. 1. Tiny faces detected with our proposed approach (shown as yellow and green boxes) and the HR approach [9] (shown as green boxes).

Note that, the existing classification-based tiny face detectors simply apply a threshold on a classification score to determine whether the corresponding candidate is face or non-face, as shown in the first stage of Fig. 2. However, the optimal threshold is often difficult to obtain. In this paper, we propose a novel idea to exploit the semantic information (consisting of spatial locations, scales and textures) of a candidate's neighbors to classify a target to face or background. Specifically, based on such semantic information, we try to group all of the faces into one cluster, while backgrounds are kept far away from the cluster. For this purpose, we propose a Metric Learning and Graph-Cut (MLGC) framework, which carries out further classification on the candidates produced by other object detectors. Fig. 2 illustrates the framework of this idea.

Jiangbin Zheng and Xiangjian He are the co-corresponding authors for this paper.

We first obtain a high-recall classifier which aims to retrieve all of the targets in an image, but may unavoidably introduce lots of false positives. Our focus is to retrieve faces with low classification scores but remove these false positives. In order to do this, we design a metric 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 into faces or non-faces, 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 metric learning and graph-cut framework to exploit the semantic information between targeting objects' neighbors. Secondly, to depict local neighborhood relationships, we introduce a pairwise constraint into the tiny face detector to improve the detection accuracy. Thirdly, to realize such a pairwise constraint, we convert the problem of regression that estimates the similarity between different candidates into a classification problem that produces the scores of classification for each pair of candidates.

2. RELATED WORK

Face detection is a classic topic in computer vision. The pioneer work on the topic was published by Viola and Jones [8] who designed a cascade of weak classifiers using Haar features and AdaBoost for fast and robust face detection. Similar in spirit, numerous approaches have been developed to improve the performance with more sophisticated hand-crafted features [13] and more powerful classifiers [14]. However, these methods using non-robust hand-crafted features and optimized each components independently, and hence led to sub-optimal face detection results. Recently, face detectors based on *CNNs* [15][16][17] have greatly bridged the gap between human vision and artificial detectors.

Tiny face detection aims to detect a large number of small faces in crowded and cluttered scenes. It is totally different from detecting normal faces, because the cues for detecting a 3-pixel tall face are fundamentally different from those for detecting a 300-pixel tall face [9]. Bell [20] presented the Inside-Outside Net (ION) to model the context outside a region of interest and showed improvements on small object detection. Very recently, Hu and Ramanan [9] designed a foveal descriptor that captured both coarse context and high-resolution image features in order to effectively encode context information, which has achieved state-of-the-art performance on the WIDER FACE dataset. As we all know, it is not sufficient to detect small objects merely by extracting deep learning features from the texture inside an 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, ignoring the influence of various 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 alarms. To introduce local coherent relationships, we learn a metric to represent this coherence and use 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.

3. THE PROPOSED METHOD

Our goal is to integrate local coherent relationships into tiny face detection. In order to represent local coherent relationships, we define pairwise constraints, which are an equivalence constraint for pairs of data points belonging to same classes, and an inequivalence constraint for pairs of data points belonging to different classes.

As shown in Fig. 2, we present a metric learning and graph-cut (MLGC) approach for high-density tiny face detection. We first use a linear-SVM to estimate the similarity matrix among all candidates (Sect. 3.1) and then we construct a graph model and use the graph-cut algorithm to divide candidates into several groups (Sect. 3.2). Finally, we design a voting method to classify groups (Sect. 3.2).

3.1. Metric learning based on linear-SVM

Let $X = \{x_1, x_2, ..., x_N\}$ denote the set of N candidates (i.e., face or non-face bounding boxes). To introduce the pairwise constraint, we first build a similarity matrix $S = s(x_i, x_j), x_i, x_j \in X, i, j = 1, 2, ...N$, 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, we treat it as a classification problem and propose an unsupervised way to obtain the similarity score between two candidates. We use SVM to compute the similarity score between two candidates x_i and x_j based on multiple cues, i.e., the position, scale, classification score 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 [9]. 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 X_{Bottom} denotes the bottom 10% of the non-face patches in X. As shown in Fig. 2, in Stage 2 of our MLGC, we build a training set $\{(x_{11}^{'}, y_{11}^{'}), (x_{12}^{'}, y_{12}^{'}), ...(x_{nn}^{'}, y_{nn}^{'})\}$, $x_{ij}^{'} = \phi(x_i) - \phi(x_j)$, $y_{ij}^{'} = \{0, 1\}$. If $x_i, x_j \in X_{Top}, y_{ij}^{'} = 1$. If $x_i \in X_{Top}, x_j \in$

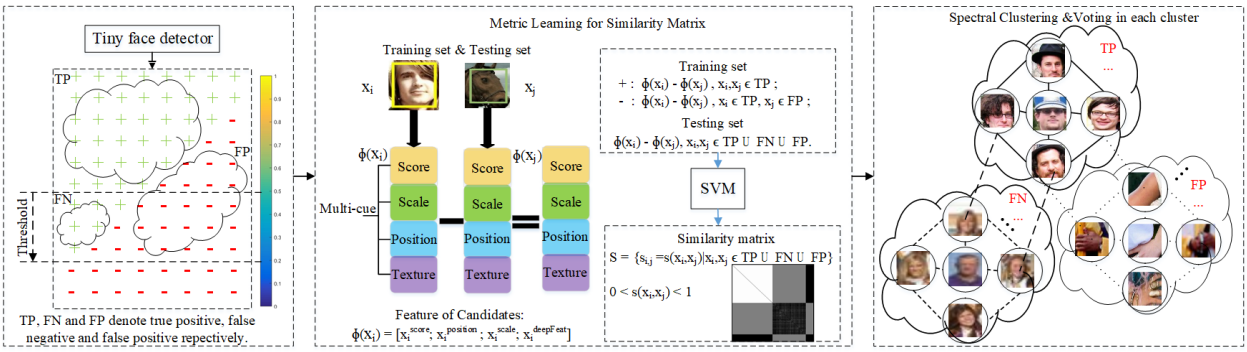


Fig. 2. The framework of our proposed MLGC for high-density tiny face detection.

$X_{Bottom}, y_{ij}^{'} = 0$. During the testing stage, we feed $x_{ij}^{'} = \phi(x_i) - \phi(x_j)$ 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 $X = \{x_1, x_2, ..., x_N\}$ and a similarity matrix S, our goal is to cluster X into different groups. Candidates are similar when they are in the same group, and are 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. Then, clustering X into groups can be reformulated with the graph model represented in Eq. 1. 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,

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

where $A_i \subset V, A_i \cap A_j = \emptyset$ and $A_1 \cup A_2 \cup ... \cup A_k = V$, $W(A_i, \bar{A}_i) = \sum_{m \in A_i, n \in \bar{A}_i} w_{mn}$, $w_{mn} = exp(-S_{mn}/2\delta^2)$ 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 $\{A_1, A_2, ..., A_k\}$ is reasonably large. So, we can transform Eq. 1 as follows:

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

According to [21],

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

where L is the Laplacian matrix, $H^T H = I$, and the indicator

$$H = \{h_1, h_2, ..., h_k\}$$

with

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

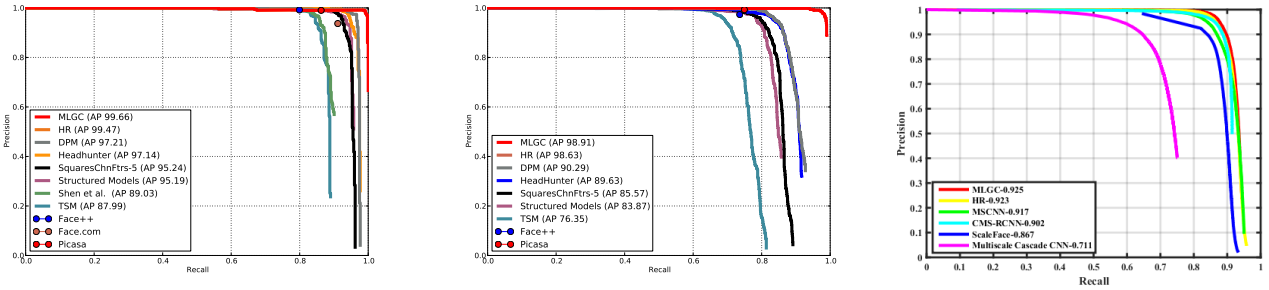
for $i = 1, 2, ..., N; j = 1, 2, ..., k$.

Eq. 3 is the standard form of a trace minimization problem. According to the Rayleigh-Ritz theorem [22], 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. So, we manage to cluster X into k groups $\{A_1, A_2, ..., A_k\}$. Finally, candidates in each group are classified to face or non-face class using voting.

4. EXPERIMENTS

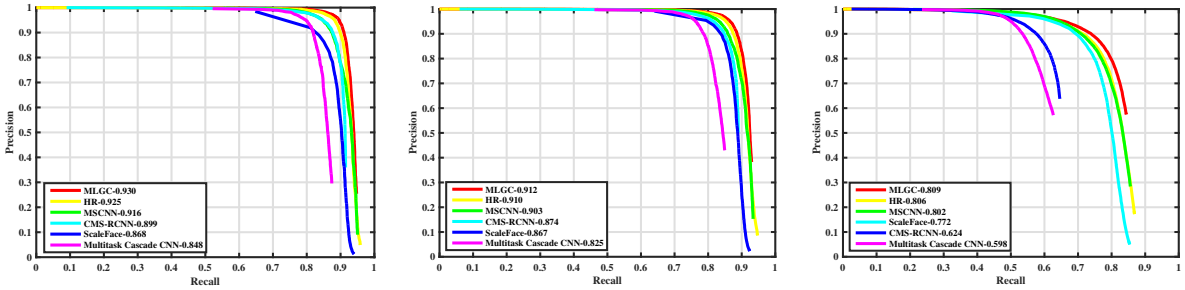
In this section, we first demonstrate the effectiveness of our proposesd semantic similarity metric and then evaluate the whole model on three widely-used face detection benchmarks, including WIDER FACE [10], Annotated Faces in the Wild (AFW) [23] and Pascal Faces [24].

To demonstrate the effectiveness of our proposed semantic metric (see Subsection 3.1) for similarity measurement, we create positive samples, i.e., the ground truth face regions, and negative samples which are patches randomly sampled from background, and evaluate the discriminative ability of the computed similarity matrix on the WIDER FACE validation set. The average precision in each image on the validation set is 79.58% in the testing set composed of both positive and negative samples, 72.25% in the set of positive samples only, and 86.75% in the set of negative samples only.



(a) On the AFW dataset (b) On the PASCAL FACE dataset (c) On the "easy" subset of WIDER FACE

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

4.1. The AFW and PASCAL FACE Dataset Results

The AFW dataset has 205 images containing in total 473 labelled faces. We evaluate our model against the HR [9], DPM [25], Headhunter, SquaresChnFtrs [26], Structured Models [24], Shen et al. [27], TSM [23] and commercial detectors (e.g., Face.com, Face++ and Picasa). As illustrated in Fig. 3(a), our MLGC 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 [25]. We also evaluate our model against the HR [9], DPM [25], Headhunter, SquaresChnFtrs [26], Structured Models [24], Shen et al. [27], TSM [23] and commercial detectors (e.g., Face++ and Picasa). As shown in Fig. 3(b), our MLGC outperforms all other detectors on PR curves.

4.2. Results Obtained on the WIDER FACE Dataset

The WIDER FACE Dataset is one of 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.

We compare our MLGC with the HR [9], MSCNN [28], ScaleFace [29], CMS-RCNN [18] and Multitask Cascade CNN [30]. The PR curves on the testing set is presented in Fig. 3(c), 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.

5. CONCLUSION

In this paper, aiming to improve the performance of tiny face detection, we have proposed a novel idea to exploit the semantic similarity between targeting objects' neighbors and created a pairwise constraint to depict such semantic similarity. Then, a framework which adopts the metric learning and graph-cut techniques has been formulated to boost the accuracy of existing tiny object classifiers. Experiments conducted on three widely-used benchmark datasets for face detection have demonstrated the improvement over the state-of-the-arts by applying this idea. For time efficiency, we take some time to improve tiny face detector's performance. The mechanism of our proposed framework is generic indicating that the framework has a great potential being applied on other small and generic object detectors.

6. REFERENCES

[1] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013.

[2] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016, pp. 146–155.

[3] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al., "Deep face recognition.," in *BMVC*, 2015.

[4] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[5] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *CVPR*, 2015.

[6] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *NIPS*, 2014.

[7] Minyoung Kim, Sanjiv Kumar, Vladimir Pavlovic, and Henry Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *CVPR*. IEEE, 2008, pp. 1–8.

[8] Paul Viola and Michael Jones, "Robust real-time face detection," *IJCV*, vol. 57, no. 2, pp. 137–154, 2004.

[9] Peiyun Hu and Deva Ramanan, "Finding tiny faces," in *CVPR*. IEEE, 2017, pp. 1522–1530.

[10] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang, "Wider face: A face detection benchmark," in *CVPR*, 2016, pp. 5525–5533.

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Cheng-Yang Szegedy, and Alexander C Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016.

[12] Quoc V. Le, Navdeep Jaitly, and Geoffrey E. Hinton, "A simple way to initialize recurrent networks of rectified linear units," *CoRR*, vol. abs/1504.00941, 2015.

[13] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li, "Aggregate channel features for multi-view face detection," in *IJCB*. IEEE, 2014.

[14] Minh Tri Pham and Tat Jen Chain, "Fast training and selection of haar features using statistics in boosting-based face detection," in *ICCV*, 2007, pp. 1–7.

[15] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua, "A convolutional neural network cascade for face detection," in *CVPR*, 2015.

[16] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li, "Convolutional channel features," in *CVPR*, 2015, pp. 82–90.

[17] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang, "From facial parts responses to face detection: A deep learning approach," in *ICCV*, 2015.

[18] Chenchen Zhu, Yutong Zheng, Khoa Luu, and Marios Savvides, "Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection," in *Deep Learning for Biometrics*, pp. 57–79. Springer, 2017.

[19] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015.

[20] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *CVPR*, 2016, pp. 2874–2883.

[21] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, 2007.

[22] Helmut Lutkepohl, "Handbook of matrices.," *Computational Statistics and Data Analysis*, 1997.

[23] Xiangxin Zhu and Deva Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*. IEEE, 2012, pp. 2879–2886.

[24] Junjie Yan, Xuzong Zhang, Zhen Lei, and Stan Z Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790–799, 2014.

[25] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[26] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool, "Face detection without bells and whistles," in *ECCV*. Springer, 2014, pp. 720–735.

[27] Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu, "Detecting and aligning faces by image retrieval," in *CVPR*. IEEE, 2013, pp. 3460–3467.

[28] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *ECCV*. Springer, 2016, pp. 354–370.

[29] Shuo Yang, Yuanjun Xiong, Chen Change Loy, and Xiaoou Tang, "Face detection through scale-friendly deep convolutional networks," *arXiv*, 2017.

[30] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, 2016.