ATLANTIS PRESS

Advances in Engineering Research, volume 149

2nd International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018)

Masked Face Detection Via a Novel Framework

Qiting Ye[a]

Beijing Key Laboratory of IOT information Security, Institute of Information Engineering, Chinese Academy of Sciences, China

School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China.

ayeqiting@iie.ac.cn

Keywords: masked face detection; convolutional Neural Network; deep learning

Abstract. Masked face detection has a large variety of application like community policing, criminals capture, etc. Meanwhile, it is a challenging problem for academia. Study on normal face detection lasts for decades and has reached a high level in recent years. Contrastively, masked face detection still requires a lot of further study. Compared to the normal face detection, masked face detection is harder to deal with due to the loss of key points and diversity of occlusion degree. Traditional detection algorithm and framework work ineffectively on the problem. This paper researches into the problem, gathers related images, builds up a dataset, and proposes a novel framework for masked face detection. The whole system contains four modules. Proposal module produces candidate boxes and extracts the feature of each region with the help of two pre-trained network. Then, classification module trains a four-layer full connection neural network. The network is aimed at predicting five parameters of each proposed region. Regression module is designed to gain a more accurate position of the proposal. Finally, cluster module combines the information of neighboring boxes to give the final detection result. This module is different from the NMS algorithm which is frequently used in many traditional detection frameworks. Compared to the NMS algorithm, the new module produces a more accurate position and increases the robustness of the whole framework. Experimental results on the dataset show that the proposed framework remarkably outperforms 6 state-of-the-arts by at least 16.8%.

1. Introduction

Face detection is always a hot topic and important problem in the field of artificial intelligence. In the past years, accuracy of face detection has been improved constantly with the help of deep learning [1]. However, in case of community policing, video surveillance, criminals capture, the faces tend to be sheltered partly by mask, scarf. Traditional face detection algorithm has a sharp decrease of accuracy when dealing with these masked faces.

There are several main questions in the masked face detection. Firstly, large dataset for masked face is lacked, which is unfavorable for the development of the research. Secondly, masked faces lack local facial feature, which is replaced by the variable noise. It results in failure in generalization for many face detection algorithms.

Toward the end, this paper firstly builds up a big dataset with 25876 training images and 4935 testing images. The masked faces in the dataset are various in skin color, posture, type of occlusion, degree of occlusion and sex. Meanwhile, this paper proposes a novel masked face detection framework. The system mainly contains four modules, which are related tightly. In proposal module, we employ two pre-trained networks to deal with the images. We gain a number of proposal regions using MTCNN [2]. And each region will be represented with a 4096-dimension vector using VGG-16 [3]. In Classification Module, we design a four-layer full connection network to finish the prediction task. In regression module, we fine-tune the position of the proposal region. We hope the IOU (Intersection-over-Union) between the proposal region and the perfect one can be as big as possible. In Cluster Module, we cluster the neighboring regions and analyze them. We discard the regions which have a high prediction value but does not overlap with any other one. Experiments shows this

module can improve the precision by 7% compared to the NMS (Non-Maximum Suppression) algorithm, which is generally used in other detection framework.

The main contributions of this paper are three folds:1) We present a dataset of masked faces that includes enough various images for research, 2) We propose a novel and complete detection framework for masked face detection, which outperforms 6 state-of-the-art face detectors, and 3) We conduct an analysis on the key challenges in masked face detection, which may be helpful for developing new face detectors in the future.

2. Related Work

Tracking back to the literature, previous works in face detection mainly rely on handcraft feature designs, such as Fisherface based on Linear Discriminant Analysis [4], Eigenfaces based on Principal Component Analysis [5], Harr-like features with cascade detector [6], and Gabor-like high dimensional features with Adaboost detector [7]. The traditional detection schemes tend to behave far from satisfactory in generalization especially when the environment changes radically. In recently years, the combination between convolutional network and GPU brings breakthrough in benchmark evaluations, such as LFW [8-10] and FDDB [11,12]. The success of deep learning works on the face detection brings inspiration for masked face detection.

Although research on face detection has been developing for decades, the framework and algorithm that are earmarked for masked face detection are rare. Lin et al. [13] modified LeNet [14] and designed a detection framework based on a 1000-training-sample dataset. The scenes in this dataset are single. MTCNN [2] is a face detection framework trained on the WiderFace [15]. Because some faces in this dataset has labels for occlusion, MTCNN [2] has a weak detection capability for masked faces.

In total, it still lacks a large dataset for masked face detection and a detection framework designed especially for masked face.

3. Dataset Build

In the preliminary investigation, we find that currently there is no large dataset for the research on masked face detection, For the convenience, we decide to construct a dataset. The images on masked face are gained mainly with two methods. 1) We download related images in batch using general search engine like Baidu, Bing and Sougou. Images of this part tend to be from news reports including many different scenes. 2) We get other images from social-network site like Weibo, Renren and Tieba. Images of this part tend to be selfies. The quality is high, and some images are dealt with photo filters. We delete the repeated images and extremely blurry images. Finally, we get 30811 images and each image contains one masked face at least.

Based on the images, we do the labeling work.

3.1Position Coordinates

We mark the position coordinates of masked face, eyes, occlusion and glassed. Some masked face in the image are hard to recognize even for human, and we give them a ‘invalid’ label. Invalid faces will not be counted when computing the average precision.

3.2 Information of Occlusion

We observe three areas that probably are covered: chin, mouth and nose. The degree of occlusion is decided by the amount of covered areas.

And the type of occlusion is also important for masked face detection. We define three types: simple occlusion, complex occlusion and physical occlusion. Simple occlusion is mainly in simple color and has few texture like white dust mask. Complex occlusion is mainly in multiple colors and has many texture like scarf and colorful mask. Physical occlusion means the face is occluded by hands, legs or other people.

3.3 Information of Face

We mark the skin color, the orientation of the face and sex which may be helpful for multi-task learning.

4. Detection Framework

Due to the multiple scenes, variable degrees of occlusion, various types of occlusion, detection task can be hard for tradition detection framework. We try to design a framework that can deal with the problem robustly. Our framework contains four modules and details are described as follows.

4.1 Proposal Module

Proposal module produces potential proposals from the images and extracts the feature. We take two pre-trained convolutional networks: MTCNN [2] and VGG-16 [3]. The P-net in MTCNN [2] can generate a large number of proposals. From the Tab. I, we can find the influence of different threshold. A high threshold results in a low recall while a low threshold results in many samples which can be a pressure for classification network. Because the later cluster module can handle the problem that how to choose the best one from many proposals, we choose a low threshold 0.4 to ensure the recall of the whole framework.

For the images in the dataset, after the process of P-net, we can get a set of proposals, which can be represented by a 3-unit vector P= (x, y, s). In this vector, x means horizontal ordinate, y means vertical ordinate, and s means scale of the proposal.

Table 1 Recall (%) on the training set

Orientation	Left	Left front	Front	Right front	Right
0.6	53.30	67.17	71.79	54.61	43.56
0.5	78.39	85.55	89.65	8.020	66.22
0.4	89.30	92.73	94.40	91.70	82.89

Then we compute the IOU of each proposal with its nearest true region Pt if it exists.

We divide all the proposals into two parts based on the IOU. If the IOU is greater than 0.5, the proposal is regarded as a positive sample. If not, it will be a negative sample.

We find the positive samples tend not to overlap with the true one completely. We hope the later modules can adjust the position to make a more precise proposal, so we compute the relative displacement horizontally and vertically and the variation of scale. For a proposal P= (x, y, s), and its corresponding region Pt= (xt, yt, st), we compute the variations. We define the label as a 5-dimention vector: L= (tag, IOU, (x - xt) / xt, (y - ytt) / yt, (s -st) / st), and tag here equals 1 if the proposal is positive or 0 if not. And For negative samples, the L will be set to (0, 0, 0, 0, 0) directly.

And for each proposal, we extract the feature by the help of VGG-16 [3] and get a 4096-dimention vector.

Combining the 4096-dimention feature vector F and 5-dimention label vector L, we can produce the final data prepared for the later modules.

4.2 Classification Module

In this module, we train a four-layer full connection network to finish the prediction task. The network accepts a input of 4096-dimention vector and gives a output of predicted vector Lp. We take L2 distance function as the loss function.

4.3 Regression Module

To make the proposal closer to the true region, we do the regression on each proposal on the base of predicted label.

After the classification module for testing set, we get predicted vector for each proposal. To address a more precise position, we calculate the new position combining the information of shifting.

We can find that the proposal has a higher IOU with the true region after regression. Fig. 1 shows two typical images after the aggression. Red boxes refer to the proposal that MTCNN [2] generates directly, blue boxes refer to the true region, and the yellow boxed refer to the proposal after regression.

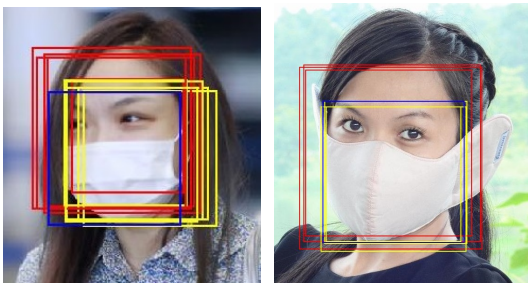

Fig. 1 Example of different face orientations in the dataset

Regression is very important for the recall of the whole framework. Since we use a lower threshold than that in evaluation, some proposals can be left out.

4.3 Cluster Module

In the general detection framework, NMS is employed to produce the final output. But it behaves not well in the masked face detection. Some local proposal has a higher predicted value, which can be a severe disruption. The former modules produce a number of proposal. The most proper proposal may not have the highest predicted value, and then NMS will choose a worse proposal.

We consider the problem deeply and find there are several proposals in the neighborhood of the true region and the outlier proposal tends to be discrete. We decide to cluster the proposals. Instead of the predicted value itself, the average predicted value of neighboring proposals is taken into consideration. And if the proposal has few neighbors, it will be discarded.

For each proposal, we observe its neighboring proposals which are predicted as positive region in the classification. We keep a vector Lp= (tag, predicted value, x, y, side) to represent some proposal. Here, tag indicates whether the proposal is valid and predicted value tells the confidence. They are both gained from classification module. We observe all the other proposals close to the proposal.

And we compute the average IOU for the current proposal with all the proposals in the neighboring set and get a new label. We add up the new label to the predicted vector. Based on the new vector, we choose the final proposal.

The cluster operation can lower the influence of outlier proposal and lessens the pressure of classification module. It is very common in the masked face detection that local face is mistaken as masked face. And the cluster module can handle it effectively. The module allows proposal module can choose a low threshold. The precision will descend without the cluster operation when there are too many proposals. Totally, the cluster module adds up to the robustness of the whole framework.

5. Experiment

In this section, we compare our framework and six famous detection frameworks including SURF [16], NPD [17], ZR [18], HH [19], HPM [20] and MTCNN [2]. From Tab. II, we can see that our approach significantly outperforms the other six models in masked face detection. While our AP reaches up to 76.8% over the testing set of the dataset, the second-best model, MT, only reaches an AP of 60.8%.

Table 2 Average precision (%) on the testing set

SURF	NPD	ZR	HH	HPM	MTCNN	OUR
16.1	19.6	41.6	50.9	60.0	60.8	76.8

6. Conclusion

In this paper, we build up a dataset for masked faces and propose a novel detection framework. The framework is based on convolutional neural network and contains four relative modules. The experiment on the dataset demonstrates the accuracy and robustness of the framework.

References

[1]. Y. LeCun, Y. Bengio and G. Hinton. Deep Learning. Nature volume 521, pages 436–444 (28 May 2015).

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

[3]. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[4]. P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[5]. M. Turk and A. Pentland, Eigenfaces for Recognition, Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.

[6]. P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, IEEE, Kauai, HI, USA, 2001, pp. I–511.

[7]. C. Liu, H. Wechsler, Gabor feature-based classification using the enhanced Fisher linear discriminant model for face recognition, IEEE Trans. Image Process. 11 (4) (2002) 467–476.

[8]. Y. Sun, X. Wang, X. Tang, Hybrid deep learning for face verification, in: Proceedings of the IEEE Conference on Computer Vision, IEEE, Portland, Oregon, USA, 2013, pp. 1489–1496.

[9]. V. Jain, E.G. Learned-Miller, Fddb: A Benchmark for Face Detection in Unconstrained Settings, UMass Amherst Technical Report, 2010.

[10]. 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.

[11]. Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[12]. M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of European Conference on Computer Vision, Springer, Zurich, Switzerland, 2014, pp. 818–833.

[13]. S. Lin, L. Cai, X. Lin and R. Li, Masked face detection via a modified LeNet, in: Neurocomputing, 218 (2016), pp. 197–202.

[14]. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[15]. S Yang, P. Luo, C. Chen and X. Tang. WIDER FACE: A Face Detection Benchmark. In CVPR, 2016.

[16]. J. Li and Y. Zhang. Learning SURF cascade for fast and accurate object detection. In CVPR, 2013.

[17]. S. Liao, A. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. IEEE TPAMI, pages 234–778, 2015.

[18]. X. Zhu and D. Ramanan. Face detection, pose estimation,and landmark localization in the wild. In CVPR, 2012.

[19]. M. Mathias, R. Benenson, M. Pedersoli, and L. V. Gool. Face detection without bells and whistles. In ECCV, 2014.

[20]. G. Ghiasi and C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. IEEE TPAMI, 20(1):23–38, 2015.