Comparison and Application of Two Face Detection Algorithms

Xiaoxi Wei and Qian Yin

Abstract This chapter compares two face detection algorithms Viola-Jones and MTCNN that have certain similarities in the structure of the algorithm. They are both creative and significant objective detection algorithms though they are based on different frameworks. To have a more intuitive view of the merits and demerits of two face detectors, we test two detectors on a challenging dataset including 25 images divided into 5 categories and compare the results in terms of detecting accuracy, time consumed on the detection, and effects of nonface features on detection results. Besides, for the results of MTCNN, it is essential to observe the positions of the five feature points which are key indicators of accuracy. By comparing the results, it is clear that the detection rate of MTCNN is generally higher than that of Viola-Jones, but the detection speed of Viola-Jones is faster. Furthermore, Viola-Jones has a good overall performance in positive face detection and is better suited for fast detection in frontal view, whereas MTCNN is better suited for industrial-grade scenes requiring high accuracy.

Keywords Face detection · Viola-Jones · MTCNN · Comparison

1 Introduction

The purpose of this paper is to learn the merits and demerits, differences, and connections between face detection algorithms Viola-Jones [1] and MTCNN [2] by comparing them and testing them on a challenging dataset.

X. Wei (✉)
School of Mechanical, Electrical and Information Engineering, Shandong University, Shandong, China
e-mail: 201900800198@mail.sdu.edu.cn

Q. Yin
School of Biological and Agricultural Engineering, Jilin University, Changchun, China

C.-C. Wang, A. Nallanathan (eds.), *Proceedings of the 5th International Conference on Signal Processing and Information Communications*, Signals and Communication Technology, https://doi.org/10.1007/978-3-031-13181-3_5

Viola-Jones is a machine learning-based object detection algorithm. MTCNN is a deep-cascaded multitask framework that takes advantage of the inherent correlation between them to improve their performance.

Both Viola-Jones and MTCNN are cascade face detectors. From the paper [2], it can be known that the proposal of MTCNN is partly inspired by the Viola-Jones algorithm. The first algorithm trains multiple cascading classifiers to replace a single classifier, whereas the second algorithm employs a cascaded structure with three carefully designed networks.

The Viola-Jones framework is one of the most landmark achievements in the history of face detection, laying the foundation for the AdaBoost-based objective detection framework, while MTCNN is a representative of algorithms based on deep learning framework. So these two algorithms which have certain similarities in the structure but are based on different frameworks are chosen for comparison to delve into the differences in detecting results.

Although there are more than 10 years between the Viola-Jones detector and MTCNN, MTCNN takes both the advantages and shortcomings of Viola-Jones into account. MTCNN obviously improves the accuracy of face detection. The application characteristics of the two algorithms, as well as their advantages and disadvantages, can be generally revealed by testing the two algorithms on the challenging dataset.

With the outcomes it is discovered that the Viola-Jones algorithm is more for the frontal face. For the complete face, it has a high recognition rate. But for different poses or a variety of shields, it will be difficult. Although MTCNN can recognize faces in different poses and even occlusion well, it sometimes fails to recognize some obvious faces among multiple faces. To explore more, we select dozens of images on the Internet and divide them into five categories. Although the number is small, they all have their own characteristics and can bring challenges.

The main issues addressed in this report are as follows: (a) Find 25 more challenging images as a small dataset to test each of the two algorithms and compare the test results. (b) Pick 2–3 ROI regions in each image, test them, and compare the results. (c) Summarize the results of the two algorithms, their strengths and weaknesses, and the scenarios in which algorithms can be used.

2 Related Work

One of the contributions of the Viola-Jones algorithm is using integral images for fast computation of Haar-like features, which is motivated in part by the work of Papageorgiou et al. [3]. And the AdaBoost method proposed by Yoav Freund et al. [4] is used to build the classifier.

With the emergence of convolutional neural networks (CNNs), some people use them in a range of computer vision applications and make lots of impressive advances [5, 6]. Some CNN-based face detection algorithms were presented, inspired by the good performance of CNNs in computer vision tasks. Yang et al.

[7] build deep convolution neural networks for facial attribute detection to achieve high response in face areas, yielding candidate windows of faces. This technique, however, is very consuming in practice due to its sophisticated CNN structure. Li et al. [8] utilize cascaded CNNs for face detection. However, it has some limitations in terms of application, incurring additional computational costs and ignoring the intrinsic connection between facial landmark localization and bounding box regression. For the image pyramid constructed in the first step of MTCNN, the candidate window and its boundary box regression vector are obtained by a method similar to the Deep Dense Face Detector (DDFD) [9].

2.1 Viola-Jones

The overall flow of the algorithm can be summed up as follows:

- Application of Haar-like input features: thresholding the sum or difference of rectangular image regions.
- The new image representation technique speeds up the computation of values for a 45-degree rotation of the rectangular image region, and this image structure is employed in order to accelerate the computation of Haar-like input features.
- Using AdaBoost to create classifier nodes for the binary classification problem (high pass rate, low rejection rate) (face vs. nonface).
- Construct a filtered cascade of classifier nodes (in a filtered cascade, a node is a set of classifiers of the AdaBoost type). To put it in another way, the first set of classifiers is optimal and can pass through the region of images containing objects while allowing some images without objects to pass; the second set of classifiers is suboptimal and has a low rejection rate.

Viola-Jones achieves good performance while remaining real-time efficient. However, several studies [10–12] suggest that the performance of this detector may deteriorate significantly in practical applications with highly variable face morphology, even when using more advanced features and classifiers [2].

2.2 MTCNN

The mainframe of MTCNN is like a cascade with three neural network stages. These include proposal network (P-Net), refine network (R-Net), and output network (O-Net). The features of MTCNN come from each image that it is detecting. It can accurately identify face images with different poses and different occlusion.

MTCNN transforms the image to form an image pyramid and to detect faces of different sizes. The results are entered into P-Net for preliminary testing. The features of MTCNN are all learned from the image itself.

P-Net is a full convolution network. For the image pyramid constructed in the previous step, the candidate window and its boundary box regression vector are obtained by a method similar to Deep Dense Face Detector (DDFD). After the features of this network are input into three convolution layers, a preliminary face classifier is used to briefly determine whether this region is a face or not.

The second layer is called R-Net, and comparing with the first layer P-Net, a full connection layer is finally added. The candidate Windows screened by P-Net are input into this layer, and a large number of poor candidate Windows are filtered out. Finally, the results are further optimized by boundary box regression and NMS (non-maximum suppression) to output more reliable face results.

The last layer structure, O-Net, adds another convolution layer to R-Net. O-Net is a more complex convolutional network that has more input features and also has better performance, with the output of this layer as the final network model output. Similar to R-Net, both use a more complex structure for better optimization.

As shown above, both algorithms are very creative in their formulation, and their structures have similarities and differences. To better understand the strengths and weaknesses of these two algorithms, as well as the scenarios in which they are used, we select 25 more challenging images to test and compare the results of the two algorithms.

3 Experimental Setup

In this section, we write complete and working code based on the two algorithms and compile and test the code in a compiler to ensure that the code can run properly to obtain face detection results. We then select 25 images from the web to constitute a representative dataset for testing. To make the results more comprehensive and comparable, the images are divided into five categories with five images in each:

- Pose: This category contains images of people looking in various directions: side, top, and bottom. The objective is to evaluate the detector's accuracy in detecting faces at different viewing angles.
- Partial occlusion: In the presence of occlusion, the accuracy of face detection and face landmarks detection is reduced. There are five occlusion-affected images to investigate the extent to which occlusion affects the two face detectors.
- False pose: Images that resemble a human face can be detected as faces in some cases. Five images of monkeys, an avatar, a smiling face, and a dog are gathered to see if the two face detectors would recognize them as faces.
- Age: Given that some infants' facial features are not fully mature and the elderly are not as distinguishable as the young due to their age, five photos of infants, children, and elderly people are found to see if they could all be detected.
- Races: Images of people with various skin tones are selected to test the effect of skin tone on the two face detectors.

All of the images are input into the detector for testing, collecting the results, and comparing the results of the two different algorithms. To make the comparison more intuitive, the results are evaluated from three aspects: precision, recall, and FPS (frames per second). These metrics provide a more comprehensive assessment of the face detector's results. After obtaining these results, a more overall view of the performance of the two detectors can be obtained. As a result, it is more accessible to recognizing the strengths and weaknesses of each of the two algorithms, as well as the various scenarios in which they are applicable.

4 Experimental Results

The results are obtained by using two face detectors on each of the 25 images in this section. To visualize these metrics, we calculate the precision and recall for all images and test the time of detection for each image. The results are shown in Table 1 and Table 2.

4.1 The Results of Viola-Jones

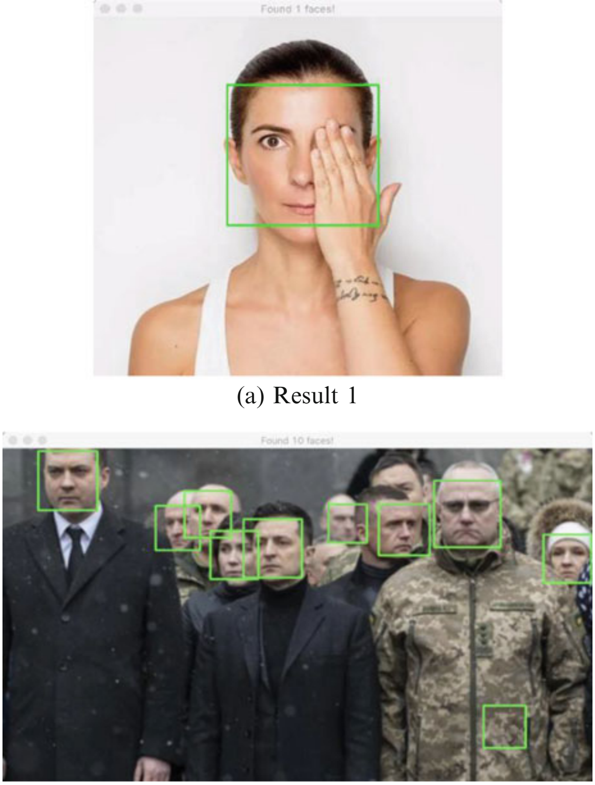
The results show that the Viola-Jones algorithm has a rather low side-face detection accuracy but a high front-face detection accuracy, while age and different races (skin color) have little effect on it, and the majority of images in these two categories are detected with high accuracy. As can be seen, the precision of "age"' reaches 100%. However, occlusion is a problem for the Viola-Jones detector, as evidenced by the fact that detection accuracy for images with occluded faces is as low as 38%, with

Table 1 The evaluation results of Viola-Jones

Category	Precision (%)	Recall (%)	FPS
Pose	15	10	23.85
Partial occlusion	38	36.36	21.71
False pose	40	40	20.11
Age	100	100	11.89
Different races	74.23	85.81	15.37
Average	53.45	54.43	18.58

Table 2 The evaluation results of MTCNN

Category	Precision	Recall	FPS
Pose	100	100	10.64
Partial occlusion	98.18	98.18	5.72
False pose	40	40	5.14
Age	100	100	1.85
Different races	79.55	91.91	1.89
Average	84.55	86.02	5.05



(a) Result 1

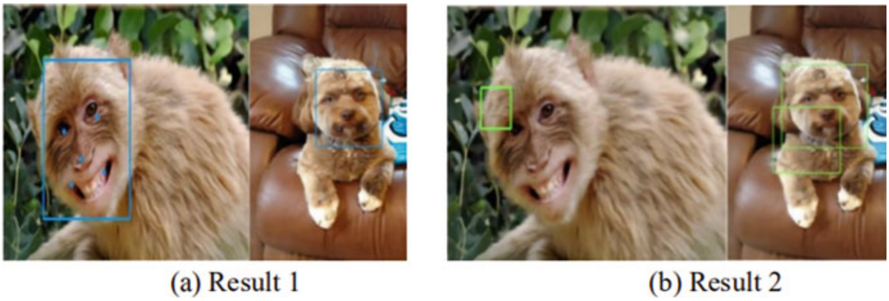
(b) Result 2

Fig. 1 The results of Viola-Jones. The image on the left is the only image that is detected in its entirety in “partial occlusion”; on the right, an area on the camouflage is misidentified as a human face. (**a**) Result 1. (**b**) Result 2

only one image in this category having an accurately detected face (as Fig. 1, Result 1). It is funny to see that the camouflage in one image is mistakenly detected as a face (Fig. 1, Result 2), so we select several more images of people wearing camouflage for detection. The detector is still disturbed by the camouflage, which proved to have a relatively large effect on the Viola-Jones detector.

There is also a type of image that resembles a human face. Images of monkeys and dogs in this category can be misidentified as human faces by Viola-Jones detector. We look for other pictures of dogs and monkeys for this purpose and discover that only pictures of animals with features that are very similar to human faces can be detected as human faces, other pictures cannot (Fig. 2, Result 2).

It is also possible to determine that the Viola-Jones is better at detecting frontal faces and faces with low occlusions by looking for commonalities in images with good detection results.



(a) Result 1

(b) Result 2

Fig. 2 The monkey and dog in MTCNN and Viola-Jones. The left two pictures are results of MTCNN and the right two are from Viola-Jones. It is interesting to find out that both algorithms have trouble dealing with monkeys and dogs. (**a**) Result 1. (**b**) Result 2

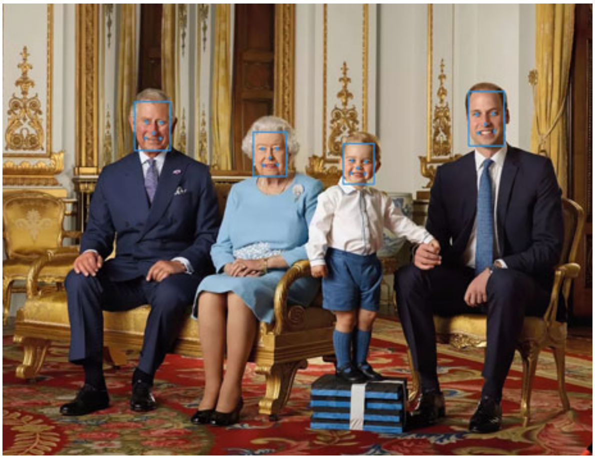


Fig. 3 Example of age. This picture is from the age images, it is clear that different ages do not make MTCNN confused. So does the partial occlusion

4.2 The Results of MTCNN

What is pleasantly surprising is that MTCNN recognizes faces in all poses, ages, and races with extremely high accuracy across all five image categories—just like Fig. 3. However, for some multi-face images, a face cannot be recognized but misjudged clothes. In addition, MTCNN is also low on false face recognition, recognizing monkey and dog faces as faces.

Specifically, for the identification of age categories, the accuracy reaches 100%. The same is true for poses (side face). But for different races, producing a face without recognition is only 79.55% accurate. For partial occlusion, the accuracy is only 98.18%. The results are shown in the Table 2. In the judgment of the "false pose," the correct recognition rate is as low as 40%. Especially for monkey and dog faces (Fig. 2, Result 1), the error rate is high. In subsequent tests with monkeys and dogs, the error rate is 66.7% for monkeys and 40% for dogs. We use different images of monkeys and dogs to further test algorithms. Along with the one depicted in

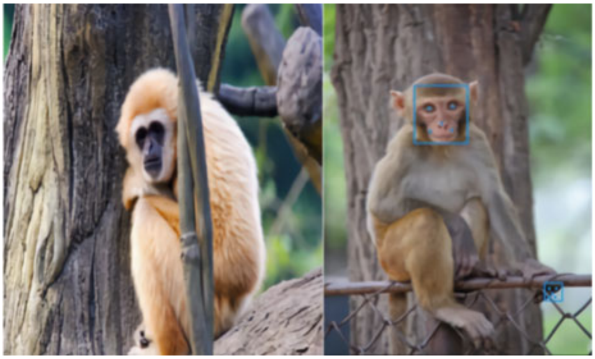


Fig. 4 Different faces with two results. The face of the monkey on the left is not detected as a face, whereas the face of the monkey on the second right is recognized as a face by the detector

Fig. 4, it is found that with a black face and a white circle around it, the left monkey is a true negative picture. However, the right picture is false negative. The reason of this may be the monkey face color. The left monkey is far away from a human being, but the right one is more similar to a human face.

When the data in the table is compared, it is clear that MTCNN has a higher precision and recall than Viola-Jones, implying that the first detector has a higher overall accuracy. However, when we extract 2–3 ROI regions from the images, such as undetected faces or regions misidentified as faces, all of the results are promising. Furthermore, the data in the table's last column shows that the average FPS (frames per second) of Viola-Jones is 18.58, while the MTCNN detects images at 5.05 frames per second. This suggests that in terms of detection speed, the Viola-Jones outperforms the MTCNN.

Based on the findings, a conclusion could be drawn. It appears that the Viola-Jones detector is significantly faster than the MTCNN detector. Both two detectors do poorly against animals such as monkeys and dogs (Fig. 2). And both of them are prone to misjudgments about certain parts of camouflage.

Analyzing the test set revealed that the algorithms have errors in animals due to hair occlusion. But based on subsequent tests and further analysis, the hair, while likely to have an impact, is not as big as expected. It is more likely that the algorithm does not look at the color of the site when detecting whether it is a face. If an animal's face is the same color, it may misjudge and recognize it as an adult face (Fig. 4). The same is true of camouflage clothing, which will be misjudged if there are patches of color that resemble a human face.

5 Conclusion

This paper compares the two object detectors of Viola-Jones and MTCNN. They are tested in terms of detection speed and accuracy on different types of images. According to the experimental results, although Viola-Jones is quick to test faces, only the frontal face has relatively reliable accuracy, and the other poses cannot be applied widely in practice. MTCNN goes a long way in this respect, as it can reliably test faces in different poses. But at the same time, MTCNN's program execution in this experiment depends on the image itself and how much time it takes.

For MTCNN, it is more used for testing faces in different poses and requires high accuracy, but for scenes that take less time to consider. Especially when images containing animals cannot be examined, errors can easily occur. For Viola-Jones, the application scope is narrower and can only be applied to the detection of front human face images, which takes less time and has low relative accuracy. But the same cannot be said for scenes that contain images of animals.

When comparing results, it is impossible to take into account all of the influences. What is certain is that MTCNN outperforms Viola-Jones in terms of detection accuracy and is less vulnerable to occlusion and other factors. However, Viola-Jones's overall detection rate is much faster than MTCNN's.

However, both algorithms fail to recognize the images that are very similar to human faces. They only detect whether it is a face or not based on a few features, but do not strictly distinguish between humans and other species or objects, which is most likely why, in some instances, animals or clothing are also detected as faces as well.

The algorithm could be improved in a variety of ways through the analysis. The underlying causes of why some animals (such as monkeys and dogs) are prone to facial misjudgment can be investigated. It is also potential to increase the precision of the results by expanding the training set and using a detector to separate clothes and other background from the face before performing face detection—in particular, from the aspect of facial color, differences between human faces and other objects, to better improve the performance of the detector.

References

1. P. Viola, M.J. Jones, Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
2. K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Proc. Lett. **23**(10), 1499–1503 (2016)
3. M.T. Pham, Y. Gao, V.D.D. Hoang, T.J. Cham, Fast polygonal integration and its application in extending haar-like features to improve object detection, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2010, June), p. 942949
4. Q. Zhu, M.C. Yeh, K.T. Cheng, S. Avidan, Fast human detection using a cascade of histograms of oriented gradients, in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, (2006, June), pp. 1491–1498
5. C.P. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, in *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, (1998, January), pp. 555–562
6. Y. Freund, R.E. Schapire, A decision- theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
7. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Proces. Syst. **25** (2012)
8. Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification. Adv. Neural Inf. Proces. Syst. **27** (2014)

9. B. Yang, J. Yan, Z. Lei, S.Z. Li, Aggregate channel features for multi-view face detection, in *IEEE International Joint Conference on Biometrics*, (2014), pp. 1–8
10. 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
11. 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
12. S.S. Farfade, M.J. Saberian, L.J. Li, Multi-view face detection using deep convolutional neural networks, in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, (2015), p. 643650