Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220620

A Survey and Test of the Face Detection Model – Based on Yolov5

Neng Zhao[1], Zihan Qin[2]

[1]College of Automation, Nanjing University of Aeronautics and Astronautic, Nanjing, Jiangsu, China.
[2]Faculty of Science, University of Melbourne, Melbourne, Victoria 3010, Australia.

zn20010203@outlook.com

Abstract: Major surveys and papers illustrate how deep learning model performance in face detection and output many excellent tools and model to solve the problem. It boosts many people trying to use them on specific issues; they need to train the model and get the hyper-parameters. However, many people are not familiar with the structure of the model, and there are plenty of parameters that make experiments more complicated; the report tries to collect and conclude some methods of experiments. This report provides a survey of the recent research in deep-learning-based face detection and the corresponding form of design of the investigation. At first, this work will do a literature review about the deep learning model, separated into three kinds (CNN R-CNN SSD). They all symbolize their series and illustrate why the work recommends the Yolo series as a pipeline. Then, the work tries the most straightforward method, OVAT, which analyzes the results, then has a discussion, and tries to give an idea for future work.

Keywords: Bright field images; Face Detection; Yolo.

1. Introduction

Face recognition is a biometric recognition technology based on facial feature information. Compared with other biometric recognition technologies, face recognition has the advantage of being natural and undetected. The face recognition system mainly includes face reading, face locking, feature extraction, face recognition, and other four modules. The purpose of face recognition is to make the computer recognize people by facial features. Face recognition based on face features is mainly used for face detection and recognition in complex scenes. It is a research direction based on image understanding, pattern recognition and computer vision, statistics, and artificial intelligence. Since the 1970s, face recognition has become one of the research hotspots in the field of computer vision and biometric recognition. Face recognition has experienced three stages: early geometric feature algorithm, a combination of artificial feature and classifier, and deep learning. In the early days, recognition rates stayed around 40% and 50%, making it difficult to use commercially. Standard methods for face recognition include support vector product (SVM), principal component analysis (PCA), and neural networks. Deep learning has recently pushed face recognition to its peak, with Facebook's DeepFace and Google's FaceNet achieving good results.

The deep learning architecture includes deep neural networks, convolutional neural networks, deep

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2/20220620

reinforcement learning, etc. Its structure overcomes the classifier problem that linear classifiers cannot complete and has excellent classification performance, which can be applied to computer vision and natural language processing. In the computational vision, many problems exist, such as face detection, object detection, segmentation, etc. This report mainly focuses on face detection, promoted by the invention of some great deep learning models.

The idea of cascade classifiers inspires deep learning, and many early deep learning-based models are based on the Cascade-CNN architecture. But with the introduction of several novel architectures for general object detection, many deep learning-based models have recently moved to single-shot detection, R-CNN-based architectures, Feature Pyramid Network (FPN) models, etc.

2. Overview of different deep face detection models

2.1. CNN part

12-layer Cascade-CNN [1] is a CNN-cascade at multi-resolution, which detects and filters through CNN to localize and refine face candidates of smaller sizes. It contains a neat 12-layer structure, so the practical performance of the GPU is excellent. The downside is the GPU-dependent Titan Black and VGA resolution images: 640x480. Multi-task cascaded CNN [2] is created by joint face detection and alignment (coarse to fine pyramid face/background-position predicted with novel online mining strategy. It enables efficient CNN joint face detection and alignment unconstrained face detection solution scheme with higher accuracy while still being context-aware in real-time. Despite poor CPU performance, GPU relies on unoptimized pair problems. Inside cascaded CNN [3] is summed up in three points: 1. The face/non-face classifier Injects the same CNN. 2. Bidirectional contextual CNN that uses body part awareness to enhance detection. 3. Feed shallow layers with simple training data and vice versa. With context awareness, cascaded CNN's internally perform better than traditional architectural backbone models. Solve complex samples. Joint training of cascaded CNN [4] is better than greedy end-to-end optimization for cascade CNN. Use back propagation in training a CNN to train CNN cascade, with global optimization advantage. Cascaded convolutional networks [5] through multi-model feature extraction. Replace std conv with a set of separate combs, and residual structure can be inputted as Unfixed size. Therefore, it has Stronger generalizability and is less GPU reliant, while it cannot be real-time.

2.2. R-CNN part

Regions with CNN features based on the target detection algorithm [6] use a selective search method [7] to extract candidate boxes. RCNN has two apparent shortcomings in image processing. First, most of the extracted Region Proposal overlap each other, and the overlapping parts will repeatedly extract features, and CNN forward propagation must be performed separately. Therefore, after the CNN model is determined, it can only accept fixed-size input, which means the plot area is fixed. Fast-R-CNN improves on some of the flaws of RCNN. It draws on the idea of SPPNet and proposes two main methods based on SPP (Spatial Pyramid Pooling) detection [8]. Fast-ring has improved in two aspects. One is that the entire training process is not divided into stages and multi-steps, and the training efficiency is higher; the other is to change to ROI pooling, which improves loss stability. Faster-RCNN first proposed Region Proposal Network [9] and used the RPN network for candidate box extraction, which solved the problem search limited by Selective. Faster RCNN achieves high-precision detection performance through two-stage plus RPN. Compared with other one-stage detection networks, the two-stage network is more accurate and can solve multi-scale minor target problems. However, Faster-RCNN still has shortcomings that need to be improved. For example, the original RoI pooling of Faster RCNN will lead to a loss of accuracy, so the problem of changing this positioning pooling or feature scale output also needs to be improved.

2.3. Single shot detection models [10]

Single-stage detection is another popular and central direction of deep learning-based face detection.

DOI: 10.54254/2755-2721/2/20220620

Different from two-stage proposal-classification detectors, such as R-CNN models, SSD detect faces in a single-stage directly from the earlier convolutional layers in the classification

network. Table 1 shows the conclusion of the three parts of single-shot detection models according to their improvements, advantages, and disadvantages. In addition, Table 2 introduces the method used in the different techniques with their function.

Table 1. Summary of progress in single-shot detection models.

method	work they have done or improvement	advantage and disadvantage
SSH (2017)	An efficient convolutional detection module with different stride-> scale invariance Headless VGG-16->lighter and faster	D: compared to yolov3, less feature fusion layer-> detection on a tiny object is not as good as yolov3 Non-robust to more complex conditions, like occlusion, illumination, etc.
	Train\evaluate\test: WIDER Test: FDDB Evaluate: Pascal Faces	
S3FD (2017)	scale-equitable framework ->handle of different scale face Scale compensation anchor matching strategy->improve recall rate of small face	D: can be faster using a faster base network (replace VGG16 network)
	Evaluate: Annotated Faces in the Wild (AFW), PASCAL Face, Face Detection Dataset and Benchmark (FDDB), and WIDER FACE	
RefineFace (2020)	STR->regression STC, SML, FSM, RFE->classification	It can only run in real-time on GPU device
	experiments on the WIDER FACE, AFW, PASCAL face, FDDB and MAFA	

DOI: 10.54254/2755-2721/2/20220620

Table 2 Function in a different method

Method	Function\ special
FAN [11]	detect the occluded faces
DFS [12]	to handle hard faces
……	Small face problem
FaceBoxes [13]	a CPU real-time face detector based
Bai et al. [14] utilize GAN [15]	detect blurry small faces

Concluded different SSD face detection models have some common points

1. solve the small face problem
2. consider the detection speed and memory efficiency
3. all of them consider WIDER as the most challenging dataset
4. supervised learning means containing subjective

3. Methodology[16]

To discover optimal sets of experiments (parameters) that could maximize the ratio of model accuracy and thus minimize the total count of necessary investigations (in this case, train/validation of a model setting), this work uses statistical principles (p-value/confidence level) and mathematical models to explain the influence of parameter and parameter interaction on the output Y (in this case: loss or accuracy measures of the NN), and some mathematical modeling, 1. MLP 2. Various activation functions 3. Gradient calculation and chain rule 4. DOE-driven numerical simulation: (n-order optimization model).

As shown in Table 3, the used datasets are summarized in 4 parts: advantages, size, factors, and shortcomings.

DOI: 10.54254/2755-2721/2/20220620

Table 3 Introduction of different datasets

name	advantages	size	factors	shortcoming
WIDER	bounding box location accuracy [RefineFace] lots of faces in one image	393703 labeled faces	scale, pose occlusion, expression, makeup, and illumination.	Maybe only suit for those optimal in the small front...
UFDD	robustness	10897 face-annotations	rain, snow, haze, lens impediments, blur, illumination variations, and distractors.	small
Casual Conversations Dataset	Equal distribution	3011 subjects	Skin Tone, Gender, Age, Lightning	Human labeled(subjective)
VGGFace2 Dataset	Hefty (3.31 million)	9131 subjects	pose, age, illumination, ethnicity, and profession	Limitation: faces in this dataset detected by a specific model?

The experiment chooses the Casual Conversations Dataset as the database for our experiments because:

1. Each photo in this database has fewer objects, is easy to operate, and has a short experimental processing time.

2.The database considers skin color, light, gender, and age and is transparent. Providing your age and gender gives data a relatively unbiased view of age and gender.

The processing of the database mainly goes through several steps:

1.Write a program that randomly selects a certain number of photos of an object

2.Use labeling to manually calibrate 800 photos, customize my dataset (including 560 trains, 240 validate），and eight experiments need to be implied. Table 4 is shown the change in 8 experiments.

DOI: 10.54254/2755-2721/2/20220620

Table 4 Difference in 8 experiments

1	yolov5s	Img_size 128	mask_data.YAML –epoch 100 –batch-size 16
2	yolov5s	Img_size 256	
3	yolov5l	Img_size 128	
4	yolov5l	Img_size 256	
5	yolov5s	Img_size 128	face_data.yaml –epoch 100 –batch-size 16
6	yolov5s	Img_size 256	
7	yolov5l	Img_size 128	
8	yolov5l	Img_size 256	

4. Experimental results

The result conducted two broad categories of experiments, one of which was used for training, with images of faces with and without masks, and the Casual Conversations Dataset for testing.

Figure 1-6 shows the experiment results with label and precision.



Figure 1. face with label and precision

DOI: 10.54254/2755-2721/2/20220620

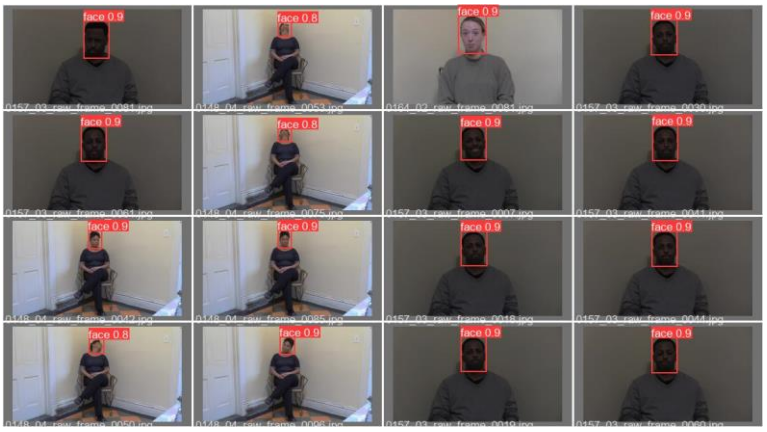


Figure 2. face with label and precision



Figure 3. face with label and precision



Figure 4. non-face with label and precision

DOI: 10.54254/2755-2721/2/20220620

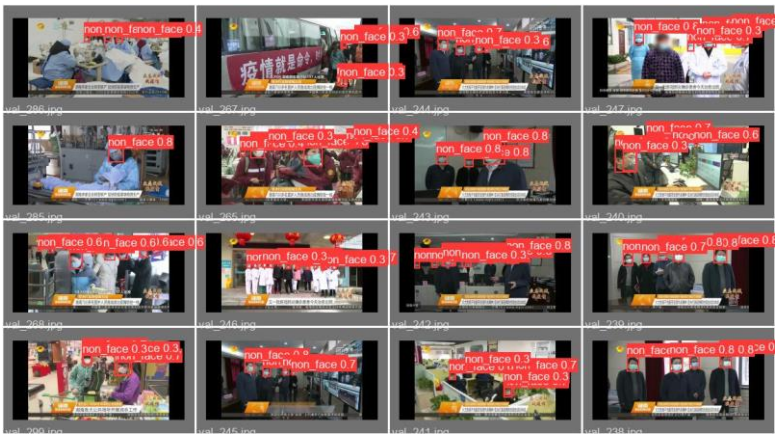


Figure 5. non-face with label and precision



Figure 6. non-face with label and precision

5. Discussion

Deep learning has experienced a long development and achieved great success in many fields. The basic models of deep understanding are roughly divided into three categories: multilayer perceptron models, deep neural network models, and recurrent neural network models.

Its representatives are DBN (Deep belief network), deep belief network, CNN (Convolution Neural Networks), convolutional neural networks, and RNN (Recurrent neural network), recurrent neural network. No matter what kind of network, they are often mixed into practical applications. For example, CNN and RNN are often connected to a fully connected layer before the upper layer output. It is difficult to say which category a network belongs to. However, there are still many areas for improvement in-depth, such as the limitations of deep learning, that is, it requires a lot of data, cannot handle hierarchical structures, and has limited capacity.

Deep learning models gradually become broader, deeper, and more prominent in continuous improvement. Still, the marginal utility of scale is reduced, and energy consumption and iteration efficiency have become more significant problems. And while the models are becoming more versatile, the algorithms are getting more and more normalized. In addition, interpretability, controllability, and predictability have not yet broken through, and understanding deep learning models is still fragile. This work also does a paper review, and the process of progress in machine learning is compared in 7 parts. Table 5 shows the result of the work.

DOI: 10.54254/2755-2721/2/20220620

Table 5 paper review of machine learning

Method	criteria	computation efficiency	small face	occlusion	illumination	CPU	GPU
Cascade-CNN Based Models	Li et al.12-layer Cascade-CNN	●				●	●
	Zhang et al. multi-task cascaded CNN	●				●	●
	Zhang et al. inside Cascade CNN	●	●				
	Qin et al. Joint training of cascaded CNN	●					
	Qi et al. cascaded convolutional networks	●					
R-CNN and Faster-RCNN Based Models	Jiang and Miller's application of a faster- RCNN model	●		●		●	●
	Wang et al. proposed" Face R-CNN.	●					
	Sun et al. proposed a face detection framework using an improved Faster-RCNN.	●	●	●	●		
	Zhu et al. proposed" CMS-RCNN.	●	●	●	●		●
	Wang et al. proposed a region-based, fully convolutional network for face detection.	●	●				
Single Shot Detection Models	SSH (2017)		●	●			●
	S^3FD (2017)	●	●				●
	RefineFace (2020)	●	●			●	●
	FAN			●			
	DFS	●					
	FaceBoxes	●	●				
	Bai et al. utilize GAN						
Yolo v-based Models	Yolo v1	●				●	●
	Yolo v2	●				●	●
	Yolo v3/4/5	●				●	●

From the perspective of network structure, YOLO's network structure is straightforward. For example, the neck part uses a simple SPP module consisting of 3 parallel branches. Each FPN structure of three different scales is fused for prediction.

From the perspective of training strategies, the work of the YOLO series uses a large amount of data enhancement, such as horizontal flipping, color space transformation, random cropping, multi-scale training, etc. These data enhancement methods can expand the training set and increase the

DOI: 10.54254/2755-2721/2/20220620

richness of data samples; in other words, data augmentation can make the model aware of more data. YOLO loads the pre-training weights and uses many data enhancement methods to train with significant epochs, with almost no parameter adjustment.

Therefore, the Yolo series dramatically improves the running speed. The Yolo series is the best choice when the experiment needs to process a large amount of data.

6. Conclusions

In this project, the work is to read the relevant literature on face detection methods, summarize the face detection methods based on different models, analyze their respective advantages and disadvantages according to the following indicators, and finally come to the choice of using Yolo. In addition, by understanding the experimental design, the experiment formulated the corresponding practical steps and methods, completed the recognition of faces and non-faces, faces and masks, and achieved a certain accuracy. Finally, the project discussed the future of face recognition technology. And direction analyzed the characteristics of the existing data sets and comprehensively sorted out the knowledge and practice of face recognition.

Acknowledgements

As all authors contributed almost equally to this work, author names are listed in noparticular order.

References

[1] H. Li, Z. Lin, X. Shen, J. Brandt, and 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.

[2] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[3] K. Zhang, Z. Zhang, H. Wang, Z. Li, Y. Qiao, and W. Liu, “Detecting faces using inside cascaded contextual cnn,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3171– 3179.

[4] H. Qin, J. Yan, X. Li, and X. Hu, “Joint training of cascaded cnn for face detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3456–3465.

[5] R. Qi, R.-S. Jia, Q.-C. Mao, H.-M. Sun, and L.-Q. Zuo, “Face detection method based on cascaded convolutional networks,” IEEE Access, vol. 7, pp. 110 740–110 748, 2019.

[6] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014

[7] Uijlings, Jasper RR, et al. "Selective search for object recognition." International journal of computer vision 104.2 (2013): 154-171.

[8] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[9] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems 28 (2015).

[10] Minaee, Shervin, et al. "Going deeper into face detection: A survey." arXiv preprint arXiv:2103.14983 (2021).

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolla´r, “Focal loss for dense object detection,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[12] L. Shi, X. Xu, and I. A. Kakadiaris, “Sanet: Smoothed attention network for single stage face detector,” in 2019 International Conference on Biometrics (ICB). IEEE, 2019, pp. 1–7.

[13] P. Hu and D. Ramanan, “Finding tiny faces,” in Proceedings of the IEEE conference on

DOI: 10.54254/2755-2721/2/20220620

computer vision and pattern recognition, 2017, pp. 951–959.

[14] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection," in Deep learning for biometrics. Springer, 2017, pp. 57–79.

[15] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li, "Detecting faces using region-based fully convolutional networks," arXiv preprint arXiv:1709.05256, 2017.

[16] Lasheras, F.S., Vilán, J.V., Nieto, P.G. and del Coz Díaz, J.J., 2010. The use of design of experiments to improve a neural network model in order to predict the thickness of the chromium layer in a hard chromium plating process. Mathematical and computer modelling, 52(7-8), pp.1169-1176.