ISSN: 0067-2904

A Review on Face Detection Based on Convolution Neural Network Techniques

Hayder Kadhim Dhahir*, Nassir Hussein Salman
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

Received: 27/4/2021 Accepted: 28/6/2021 Published: 30/4/2022

Abstract

Face detection is one of the important applications of biometric technology and image processing. Convolutional neural networks (CNN) have been successfully used with great results in the areas of image processing as well as pattern recognition. In the recent years, deep learning techniques specifically CNN techniques have achieved marvellous accuracy rates on face detection field. Therefore, this study provides a comprehensive analysis of face detection research and applications that use various CNN methods and algorithms. This paper presents ten of the most recent studies and illustrate the achieved performance of each method.

Keywords: Face detection, Convolutional neural networks, Face recognition, Deep learning.

مراجعة للكشف عن الوجه باستخدام تقنيات الشبكات العصبية التلافيفية

حيدر كاظم ظاهر*, ناصر حسين سلمان

قسم الحاسوب, كلية العلوم, جامعة بغداد, بغداد, العراق

الخلاصة

يعد اكتشاف الوجه أحد التطبيقات المهمة لتقنية القياسات الحيوية ومعالجة الصور الرقمية. تم استخدام الشبكات العصبية التلافيفية (CNN) بنجاح مع نتائج رائعة في مجالات معالجة الصور الرقمية بالإضافة إلى التعرف على الأنماط. في السنوات الأخيرة ، حققت تقنيات التعلم العميق على وجه التحديد تقنيات الشبكة العصبية التلافيفية معدلات دقة رائعة في مجال اكتشاف الوجه. لذلك ، تقدم هذه الدراسة تحليلاً شاملاً لأبحاث وتطبيقات اكتشاف الوجه التي تستخدم طرق وخوارزميات الشبكات العصبية التلافيفية المختلفة. سنلقي نظرة على عشر من أحدث الدراسات المتنوعة ونوضح الأداء المحقق لكل طريقة.

1. Introduction

Face detection is a well-known topic in computer vision because it is a necessary step in a variety of applications, including face recognition [1], facial performance capture [2], eye-tracking [3], facial expression analysis [4], facial expression transformation [5] and so on. There are unlimited interesting interdisciplinary applications [6–11] in the field of animation that are not limited to the conventional fields. However, as can be seen in Figure 1, several factors such as occlusion, illumination, and the variety of faces raise significant difficulties in face detection [12].

*Email: hayderkadhim.90@gmail.com

The Convolutional Neural Network (CNN) achieved incredible results. It is already one of the most well-known neural networks in the field of deep learning, attracting both business and academic interest. Using Convolutional Neural Networks, image processing has enabled people to achieve things that were previously thought to be impossible.

Face recognition has gotten a lot of popularity and is one of the most promising image processing technologies. Human faces are visual stimulants that are meaningful, multidimensional, and complex. It's difficult to build a computational system for face recognition [2]. Face detection is a critical component of a face recognition systems since it helps to concentrate computing energy on the portion of an image that contains a face. It is essential to perform face detection to extract appropriate data for face and conduct analysis of facial expression before using a facial recognition technology.

Due to the difficulty of the face detection technique, many applications focused on the detection of human face have been created recently. For example, applications for cell phones, security systems, intelligent robots, digital monitoring, PC cameras, laptop, and digital cameras. These applications perform a huge part in our lives. Nonetheless, the applications' algorithms are complex, making it difficult to satisfy real-time frame-rate specifications. Many approaches for enhancing the efficiency of face detection have been suggested over the last decade.

This review was written to concentrate on the research that has been done on face detection using a convolutional neural network as it will be illustrated in this paper. Some challenges in face detection are shown in Figure 1.

Figure 1- Some challenges in face detection [12].

2. Literature Review

In the recent years, sizable research have been conducted for the purpose of face detection. In this section, we will focus on recent research on Convolutional Neural Networks for the mentioned purpose above.

2.1 *Hierarchical Convolutional Neural Network (HCNN)*

HCCN is a two-layer detector, the first layer is a spatial pyramid pooling (SPP) based that removes the fixed size bonds of the network and decrease the computational complexity. The classifier is named FaceHunter [13]. The second layer is CNN refine structure, which consist of two levels, deep CNN and the FaceHunter classifier as shown in Figure 2.

After the image run through the FaceHunter classifier, initial detection results of the face are produced. The FaceHunter will directly output the positive faces, while the non-faces

(Negative outputs) will not get deleted in the first layer, they will be refined by the CNN refine structure.

Some of the results with semi high score are regressed by the deep CNN then run through the FaceHunter classifier of the second layer to output the positive results and delete the negatives[13].

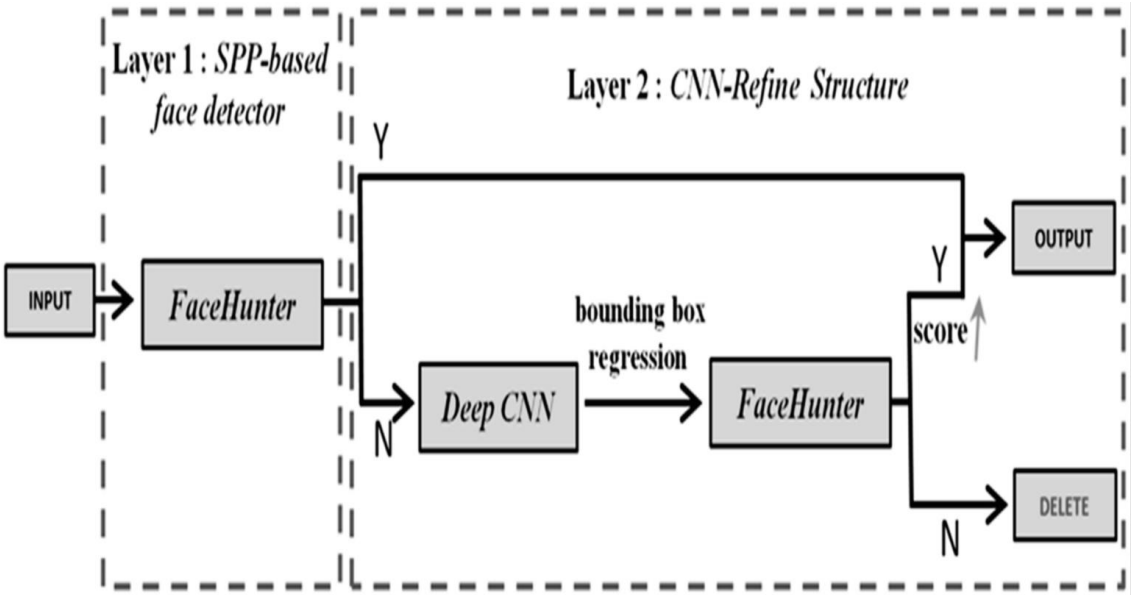


Figure 2-HCNN architecture [13].

2.2 *Cascaded Convolutional Neural Network (CCNN) based on Separable Residual Module (SRM)*

This CCNN framework uses a cascaded architecture consisting of three stages of deep CNN to enhance detection accuracy [14]. This network can predict faces in a coarse-to-fine manner, which uses a low-resolution picture and increase the resolution gradually into a finer picture. This method uses a combination of residual structure and separable convolution instead of the standard convolution. The SRM is a new module used to reserve the benefits of the residual structure performance and lower the computation complexity.

SRM adds a shortcut connection directly to the separable convolution module. The channel of input is decreased to ¼ of the first input using a 1*1 convolution and make a feature extraction using a convolution of 3*3 channel by channel. The 1*1 convolution connects and combines the features and restore the channels to their original number as shown in Figure 3.
The rectified linear activation function (ReLU) is a linear function that outputs a zero if the input is negative and will give a direct output if the input is positive.[14]

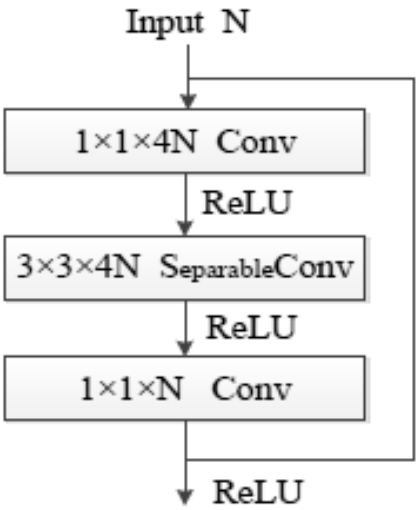


Figure 3- Separable Residual Module (SRM) [14].

This network of CNN is based on the theory of regression, it replaces Multitask convolutional neural network (MTCNN)'s standard convolution with SRM and removes the facial landmarks task. So only bounding box regression and classification tasks are performed during the detection stage. The detection precision is enhanced using the CCNN and the technique of the channel expansion in the SRM. The separable convolution depth-wise type is used to lower the amount of the computation to preserve a fast speed of detection.

2.3 *An improved faster Region Convolutional Neural Network (Faster RCNN)*

RCNN [15], Fast RCNN [16] and Faster RCNN [17] were proposed by the same author for object detection. An improved Faster RCNN was based on these three mentioned studies but with some improvements including the number of different strategies, such as hard negative mining, feature concatenation, model pre-training, multi-scale training and calibration of key parameters [18].

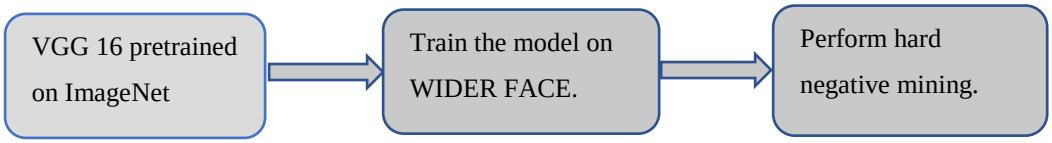
This method uses the same framework of the Fast RCNN, which include two parts, the first one is Region Proposal Network (RPN) to produce a list of some region proposals that mostly contain objects or can be named as Regions of Interest (RoI). the second part is Fast RCNN, which is used for classifying the outcome regions of the first part into a specific object and then the boundaries of the regions get refined.

The common parameters are shared between the two parts in the convolution layers, and they are used for the purpose of feature extraction.

In the beginning the CNN of Faster RCNN is trained using a dataset of WIDER FACE [19] as the first step as shown in Figure 4. The pre trained model would be tested using the same dataset to produce hard negatives. In the training procedure second step, the hard negatives are run through the network. Less false positives are produced as a result of using the hard negatives for training.

The FDDB dataset is used to fine-tune this model. In the final process of the fine-tuning, the multi-scale training is applied. To further enhance the performance, a feature concatenation technique is used. For the entire training processes, the same training technique of the Faster RCNN is used for its strong performance and simplicity.

An optional final step, the bounding boxes that are detected as a final result are converted from rectangular to ellipses because the human face regions are more elliptical.



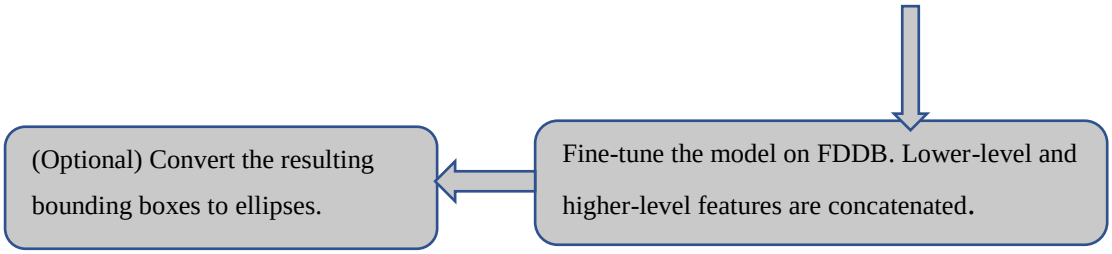


Figure 4- The proposed training procedure [18].

2.4 *Cascade Framework with Head-Shoulder Information*

This framework is a multi-resolution that utilizes parallel CNN cascades for the purpose of detecting faces in a large scene such as crowded areas, security cameras, stadiums, and stations as it's hard to detect faces in these scenes [20].

It benefits from the information of face together with the information of the head-with shoulder to solve the large scene problem. In this framework the main purpose is head with shoulder information incorporation into the cascade. The main face detector framework comprises of two alongside CNN cascades; The first is "small size cascade", which detects the faces with lower than 20*20 pixels scale. It is preferable to use a network with a small-scale for the purpose of speed as the sliding windows are large in number for the detection of small faces. The second is "big size cascade", which detects faces with higher than 20* 20 pixels scale. It takes more time for sliding windows to be processed in this cascade as the large size of the network and more accurate [20].

Figure 5- detecting different scales of faces by large and small windows [20].

2.5 *An improved Joint cascaded CNN model for face detection*

This joint cascaded CNN is proposed to address the drawbacks of the convolutional techniques [21]. The aim is to merge calibration and training to learn these tasks in the same

framework cascade to lower the candidate's number at a later stage [22]. It also enhances the accuracy by applying many algorithm simulations and analyzing the C-CNN structures.

The cascade CNN is constructed in the same way as Li et al's, [23], with three detection stages in cascade and a calibration stage in each of the three detection stages. Each calibration stage's output is used to modify the location of the window of detection for input to the next stage. The main concept is to improve accuracy by examining current C-CNN structures and implementing different improvements such as data augmentation, model and parameter optimization, and drop-out adjustment.

The input to a neural network is transformed into hidden layers series. Every hidden layer is made up of a group of neurons, each of which is entirely connected to the neurons in the earlier layer. The output layer is the final fully connected layer in a classification system, and it represents the class scores [24].

The cascaded CNN works at different resolutions, rejecting background regions in the low-resolution stage and calculating the low number of difficult candidates in the high-resolution stage [24] [25]. The calibration stage of CNN is implemented at later stages after any of the stages of detection in this structure to improve localization efficacy and reduce the candidate's number [25].

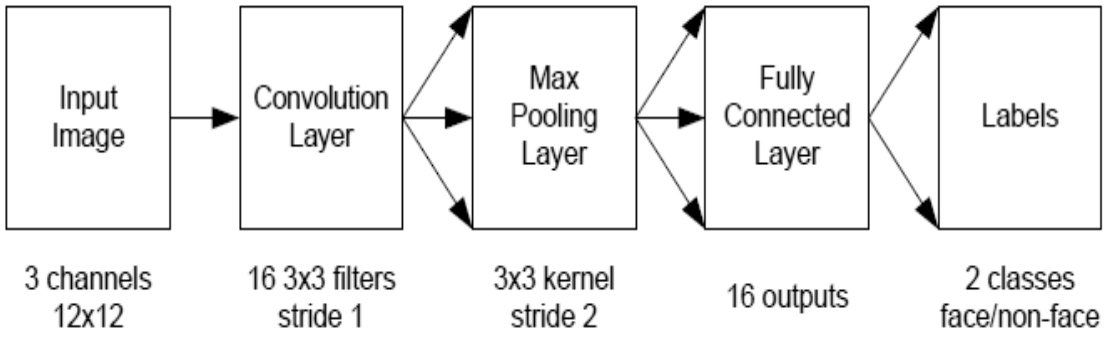


Figure 6-12_net C-CNN structure [22].

2.6 *Two-stage Cascaded Convolutional neural network*

There are two major stages in this cascaded convolutional neural network approach. A low-pixel candidate window is being used as a direct input in the first step so that the superficial convolutional neural network will obtain the candidate window rapidly. The window of the previous step is resized and then used as an input to the related network layer in the second step.

For the training period, hard samples are conducted by joint online training, and the dataset is tested using the soft non-maximum suppression mechanism. Using three input images with varying resolutions (12×12, 24×24, and 48×48) to define this cascade CNN for the purpose of face detection. To create an image pyramid, the input image is resized to various scales. At first, a (Fully convolutional proposal network, FCPN) named a micro network is used to remove a lot of non-face windows.

The remainder of the candidate window is then sent as an input to the second stage, which is (Multi Scale Network, MSN). MSN-24 denotes a branch with a 24 x 24 size input, and MSN-48 denotes a branch with an input size of 48 x 48. MSN-24's fifth layer convolutional characteristics, such as (probability distribution information), are later fused with MSN-48's. To accomplish the two tasks, bounding box regression and face classification, hard sample

mining and joint training for multiple cascade stages would be carried out. The framework is shown in Figure 7.[26]

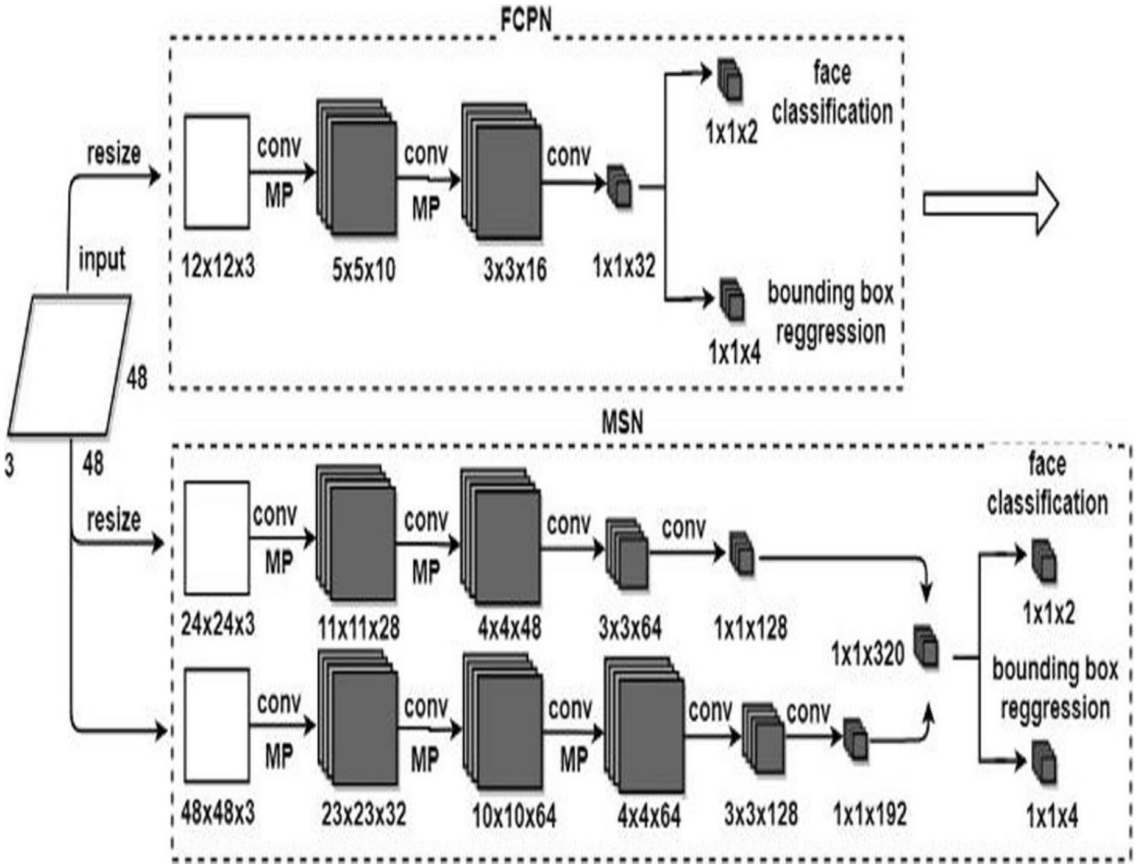


Figure 7-The framework of this two-stage cascaded method [26].

2.7 *Pre-Identification and Cascaded Detection (PMCD)*

This pre-identification mechanism is designed to detect small faces, which are extremely difficult to detect. It can significantly decrease background and other non-essential data. The cascade detector uses two phases of deep CNNs to detect small faces efficiently, i.e., the region of interest (RoI) represents the face-area candidates that are pre-identified based on the pre-identification mechanism and a real-time pedestrian detector. The list of RoI candidates is fed into the next sub-network rather than the entire picture.

The second sub-network is constructed as a shallow network to take advantage of the above mechanism and maintain high precision and ultimate time performance. This technique is constructed from two aspects, the first part is pre-identification detection, and the second part is face detection.

In the first part, a deep CNN is used to identify the pedestrian's bounding box, followed by a self-adaptive mechanism to pick RoI candidates. Next an image pyramid is built to adapt various sizes of faces as a face detection network input.

During the face detection section, the image pyramid will be forwarded to the multitask neural network for face detection. Because the preceding steps drastically minimize background interference, this detector only needs a shallow structure to work effectively in real time. This is a cascaded framework that allows the two sub-networks to be trained independently.

2.8 *A fast face detection method using CNN based on discriminative complete features (DCF)*

This is a fast approach for face detection that uses the sliding window technique on feature maps directly prior to the fully-connected layer, where a particular CNN is elaborately constructed as in Figure 8 [28].

The trained CNN model should be able to map the entire input picture to a feature space where the obtained features on all sliding window can be linearly separated. After that, by conducting classification system upon the feature space, the complexity of computation of this method for face detection system is greatly reduced.

The outcome of the layer prior to the classification layer is selected as the new features because the classification layer of the CNN architecture is normally a linear classifier. Until the classification layer, all the layers behave as a nonlinear mapping mechanism. The sparse discriminative features and the nonlinear mapping function for face detection are the two main components of this method.

Raw data is projected to the Euclidean space using the nonlinear mapping function, the sparseness of features aids linear separation. The max-pooling and convolution layers serve as a nonlinear mapping mechanism, while ReLU constraining the outcome features to be sparse. Local contrast normalization (LCN) layers deliver competitive and compact features at the same time. Local competition across neighboring features that are in a feature map is applied in the LCN layers, which are motivated by computational neuroscience frameworks [29]. The LCN layers have been empirically shown to minimize error rates and speed up supervised learning significantly [28]. Figure 8 shows the framework of this method.

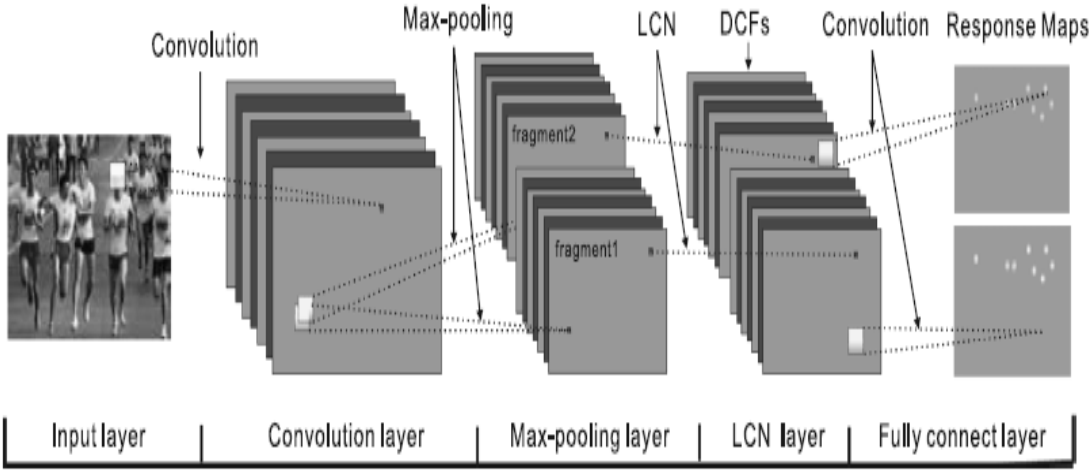


Figure 8-An overview of the framework of DFC based method [28].

2.9 *A lightweight face detector by integrating the convolutional neural network with the image pyramid*

To achieve quick and precise identification, this method uses a single-stage detection model with an incredibly lightweight CNN. This system, in particular, has a structure that integrates the CNN and the image pyramid such that the measured features can be fully used. Because of weight sharing, the size of the network will remain small.

It analyzes the detection performance of anchors with different scales and save the most powerful anchors in our model. Furthermore, every ground truth face is given a 0-or-1 weight through training to prevent complicated training samples that the simple networks cannot learn. This model is an end-to-end system that uses an image as input and outputs the estimated face box shape and coordinates.

The image pyramid is used, and the entire image pyramid is used as the network's input. Since the network's overall stride is 8, the size of the input image must be able to be divided by eight. For ease of interpretation, a 640x480 image resolution is used as an example.

Since the image pyramid's scaling factor is set to 0.5, the image size of level 2 is 320 x 240 pixels, the size of image level 3 is 160 x120 pixels, and so forth. For level 1 and level 2, the outputs of the network are 80x60 and 40x30, respectively.

The network's outcome is each boundaries adjustment and anchor's confident scores, which can be quickly transformed to estimated face boxes.

The highly overlapped estimated boxes are eliminated using non-maximum suppression (NMS) [30].

2.10 *Proposal pyramid networks for fast face detection using deep convolutional neural network*

Deep convolutional neural networks (DCNN)-based face detectors have significantly enhanced detection accuracy in the wild. The speed of detection remains the most important bottleneck, making it impossible to install these face detectors on resource-constrained platforms. To solve this problem, the Proposed Pyramid Network (PPN) for rapidly producing face candidates, which decreases the complexity of computation in face detectors that uses cascaded DCNN, is used. [31].

PPN is a lightweight multi-branch fully convolutional network that uses one image as input and produces face proposals of different sizes in terms of separate branches at the same time. It avoids the typical image pyramid model in this way, resulting in a very high processing speed. Based on PPN, a three-stage face detector using a cascaded DCNN is implemented to test the efficacy of this method.

The Proposal Pyramid Network (PPN) is the first step, which generates face proposals of multi-scale. Both the second stage, RNet-24, and the third stage, RNet-48, are networks with dual-task that optimize PPN proposals and estimate offsets of respective bounding boxes. Three levels for this detector are used to balance accuracy and speed.

More stages result in higher accuracy, although fewer stages result in faster detection. Most cascaded network-based face detection methods have three stages [32,33,34]. The aim of network architecture is to maximize speed while retaining accuracy.

PPN is a fully CN of 11 branches in the first stage, all branches are built to be lightweight. Each of the PPN's last ten branches, in fact, has one PReLU and only two convolutional layers. Nonetheless, the general performance is assured. Since higher-level branches of more complex architecture have a better ability to classify.

To speed up the reasoning speed of RNet-24 and RNet-48 as far as possible by improve convolutional layer's strides and abandon using the pooling layers (Average or Max pooling) to understand sampling operations driven by Springenberg et al. [35].

Im2col [36] is used for simplifying and converting complex convolution to matrix-matrix multiplication. Nonetheless, for the same number of parameters, a convolutional layer takes longer time compared to a fully connected layer [31].

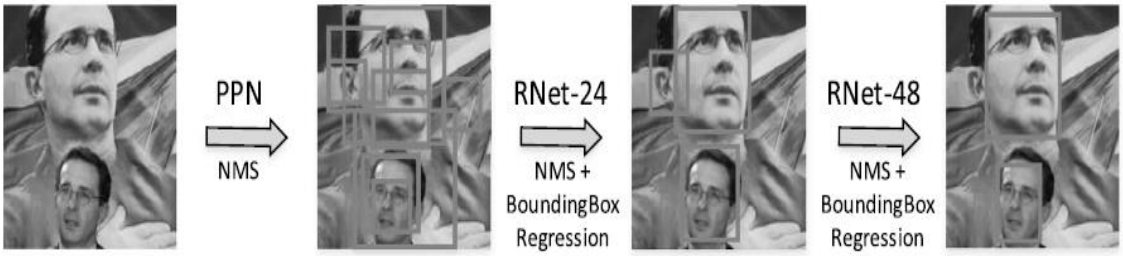


Figure 9-The overall pipeline of this PPN based method [31].

3. Comparisons between results of different convolutional neural network techniques for face detection based on accuracy and speed.

We have discussed ten of the most recent convolutional neural network techniques for face detection and their framework, and to show the performance of each technique, we will compare their accuracy results that have been achieved on different datasets as shown in Table 1.

Table 1- comparison of all mentioned methods based of the achieved accuracy and speed.

N	Method name	Dataset	Accuracy	Speed / fps
1	Hierarchical Convolutional Neural Network (HCNN) [13].	AFW [37]	94.1%	4
		FDDB [38]	76.5%	
		Pascal faces [39]	80.69%	
2	Cascaded Convolutional Neural Network (CCNN) based Separable Residual Module (SRM) [14].	WIDER FACE [19]	70.11%	28
		FDDB	78.2%	
3	An improved faster Region Convolutional neural network (Faster RCNN) [18].	AFW	85.1	9
		FDDB	84.2%	
4	Cascade Framework with Head-Shoulder Information. [20].	Shanghaitech private dataset	90.2%	Data unavailable
5	An improved Joint cascaded CNN model for face detection [22].	AFW	89%	Data unavailable
6	Two-stage Cascaded Convolutional neural network [26].	Pascal	93.92%	Data unavailable
		FDDB	91%	
7	Pre-Identification and Cascaded Detection (PMCD) [27].	Private date	84.8%	34.6
		Caltech Pedestrian dataset 18	70.12%	
8	A fast face detection method using CNN based on discriminative complete features (DCF) [28].	AFW	91.6%	Data unavailable
9	A lightweight face detector by integrating the convolutional neural network with the image pyramid [30].	WIDER FACE	84.1%	50
10	Proposal pyramid networks for fast face detection using deep convolutional neural network [31].	FDDB	92.7%	60
		Pascal	93.24%	
		WIDER FACE	78.8%	

4. Results discussion

We saw in the previous table ten of the most recent convolutional neural network techniques for face detection that achieved various accuracy rates, which are very high compared to the older techniques.

Hierarchical Convolutional Neural Network technique has achieved 94.1%, which is the highest accuracy rate of all the mentioned techniques, and it was performed on AFW dataset. While the accuracy rate dropped to 80.69% on pascal faces dataset and to 76.5% on FDDB dataset, the structure of HCNN wasn't complex as it uses only one layer of detection and another layer for refinement.

CCNN based Separable Residual Module results were the lowest of the mentioned techniques, though it used three layers of detection, which increased the complexity.

Faster RCNN has achieved 84%-85% accuracy rates but it is one of the fastest face detectors as it uses Region Proposal Network (RPN) to produce a list of some region proposals that mostly contain objects in the first part of the structure.

Cascade Framework with Head-Shoulder Information has achieved 90.2% on Shanghaitech private dataset. Since it is used to detect faces in large scene, it's hard to compare it with the other techniques but it's suitable for this kind of scenes.

An improved Joint cascaded CNN model for face detection achieved 89% on AFW data set, which is lower than HCNN in 5%. Also, it uses three detection stages and three calibration stages which make it more complex than HCNN.

Two-stage Cascaded Convolutional neural network achieved 91% on FDDB and 93.92% on pascal faces datasets, which are higher than the results of HCNN on those datasets, though the frameworks consist of two detection layers making it more complex and slower than HCNN but with better and more consistent results.

Pre-Identification and Cascaded Detection has achieved 84.8% on a private date and 70.12% on Caltech Pedestrian dataset 18 dataset for the purpose of detecting small faces in images which is much harder than usual faces.

A fast face detection method using CNN based on discriminative complete features has achieved 91.6% on AFW and it is a fast approach as it uses the sliding window technique on feature maps directly prior to the fully-connected layer, where a particular CNN is elaborately constructed. Lower results on AFW than HCNN but a faster technique.

A lightweight face detector by integrating the convolutional neural network with the image pyramid has achieved 84.1% on WIDER FACE dataset, which is the highest on this dataset, it is a very fast approach and uses only one detection stage and image pyramid.

Proposal pyramid networks for fast face detection using deep convolutional neural network achieved 92.7% on FDDB and 93.24% on pascal, which are very high on those datasets, this method's result outperforms HCNN's results on those datasets and better results on Pascal dataset than the two-stage Cascaded Convolutional neural network while only 0.68% less than it on FDDB dataset. This technique is faster than HCNN and the two-stage CNN as it uses the image pyramid approach.

5. Conclusion

Face detection has gotten a lot of attention from scientists and researchers in pattern recognition, biometrics, and computer vision in recent years.

In this paper we provided a review of face detection using various convolutional neural network techniques.

Each technique achieved varied results on different datasets due to several factors such as occlusion, illumination, and the variety of faces on the different images.

Some techniques are used to detect specific type of faces in images. For example, Cascade Framework with Head-Shoulder Information is used to detect faces in large scene while Pre-Identification and Cascaded Detection is used to detect small faces in images.

We also concluded that more complexity and more detection layers does not always lead to a better accuracy. On the other hand the more complexity means a slower approach.

Proposal pyramid networks for fast face detection using deep convolutional neural network achieved a very high accuracy and it is fast approach, which make it better than other techniques in terms of accuracy and speed as the results were very reliable on different datasets, while the two-stage Cascaded Convolutional neural network technique also achieved a high result, but it is more complex and slower. In contrast Hierarchical Convolutional Neural Network only got good result in one dataset and the accuracy dropped on other datasets so the Proposal pyramid networks for fast face detection using deep convolutional neural network is the better option for the mentioned reasons.

Even though the accuracy ratings of different face detection models differed, they have high level of reliability and can accurately detect faces, making them worthy of further study and use in practice. We believe that there is still more that can be done to improve performance.

6. References

[1] L. M. Dang, S. I. Hassan, S. Im, J. Lee, S. Lee, and H. Moon, "Deep learning-based computer-generated face identification using convolutional neural network," *MDPI*, 13-Dec-2018.

[2] L. M. U. of Houston, L. Ma, U. of Houston, U. of H. V. Profile, Z. D. U. of Houston, Z. Deng, U. of Washington, É. de T. Supérieure, U. Technologies, and O. M. V. A. Metrics, "Real-time hierarchical facial performance capture: Proceedings of the ACM SIGGRAPH symposium on interactive 3D graphics and Games," *ACM Conferences*, 01-May-2019.

[3] S.-J. Kang, "Multi-user identification-based eye-tracking algorithm using position estimation," *MDPI*, 27-Dec-2016.

[4] W. Bai, C. Quan, and Z. Luo, "Uncertainty flow facilitates zero-shot multi-label learning in affective facial analysis," *MDPI*, 19-Feb-2018.

[5] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/off," *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '09*, 2009.

[6] T. Weise, S. Bouaziz, H. Li and M. Pauly, "Realtime performance-based facial animation," *ACM transactions on graphics (TOG)*, 25-Jul -2011.

[7] H. Li, T. Weise and M. Pauly, "Example-based facial rigging," *ACM transactions on graphics (tog)*, 26-Jul-2010.

[8] L. Wei and Z. Deng, "A practical model for live speech-driven lip-sync," *IEEE computer graphics and applications*, 9-Sep-2014.

[9] H. Li, J. Yu, Y. Ye and C. Bregler, "Realtime facial animation with on-the-fly correctives," *ACM Trans. Graph*, 1-Jul-2013.

[10] C. Ouzounis, A. Kilias and C. Mousas, "Kernel projection of latent structures regression for facial animation retargeting," *arXiv preprint arXiv*, 30-Jul-2017.

[11] L. Ma and Z. Deng, "Real-Time Facial Expression Transformation for Monocular RGB Video," *InComputer Graphics Forum* Feb-2019.

[12] A. Kumar, A. Kaur and M. Kumar, "Face detection techniques: a review," *Artificial Intelligence Review*, Aug-2019.

[13] D. Wang, J. Yang, J. Deng and Q. Liu, "Hierarchical convolutional neural network for face detection," *In International Conference on Image and Graphics* Aug-2015.

[14] R. Qi, RS. Jia, QC. Mao, HM. Sun and LQ. Zuo, "Face detection method based on cascaded convolutional networks," *IEEE Access*, 12-Aug-2019.

[15] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *InProceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[16] R. Girshick, "Fast r-cnn," *InProceedings of the IEEE international conference on computer vision*, 2015.

[17] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, 6-Jun-2016.

[18] X. Sun, P. Wu and SC. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, 19-Jul-2018.

[19] S. Yang, P. Luo, CC. Loy and X. Tang, "Wider face: A face detection benchmark," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[20] C. Peng, W. Bu, J. Xiao, KC. Wong and Yang M, "An improved neural network cascade for face detection in large scene surveillance," *Applied Sciences*, 8-Nov-2018.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein and AC. Berg, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, Dec-2015.

[22] U. Joseline and HY. Lee, "Improved Joint Cascaded CNN for Face Detection," *International Journal of Engineering Research and Technology*, 2018.

[23] H. Li, Z. Lin, X. Shen, J. Brandt and G. Hua, "A convolutional neural network cascade for face detection," *InProceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[24] H. Qin, J. Yan, X. Li and X. Hu, "Joint training of cascaded CNN for face detection," *InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[25] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, L. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," *InProceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

[26] W. Yang, L. Zhou, T. Li and H. Wang, "A face detection method based on cascade convolutional neural network," *Multimedia Tools and Applications*, Sep-2019.

[27] Z. Yang, J. Li, W. Min and Q. Wang, "Real-time pre-identification and cascaded detection for tiny faces," *Applied Sciences*, 9-Jan-2019.

[28] G. Guo, H. Wang, Y. Yan, J. Zheng and B. Li, "A fast face detection method via convolutional neural network," *Neurocomputing*, 28-Jun-2020.

[29] N. Pinto, DD. Cox and JJ. DiCarlo, "Why is real-world visual object recognition hard?," *PLoS Comput Biol*, 25-Jan-2008.

[30] J. Luo, J. Liu, J. Lin and Z. Wang, "A lightweight face detector by integrating the convolutional neural network with the image pyramid," *Pattern Recognition Letters*, 1-May-2020.

[31] D. Zeng, H. Liu, F. Zhao, S. Ge, W. Shen and Z. Zhang, "Proposal pyramid networks for fast face detection," *Information Sciences*, 1-Aug-2019.

[32] I. Kalinovskii and V. Spitsyn, "Compact convolutional neural network cascade for face detection," *arXiv preprint arXiv*, 6-Aug-2015.

[33] H. Li, Z. Lin, X. Shen, J. Brandt and G. Hua, "A convolutional neural network cascade for face detection," *InProceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[34] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, 26-Aug-2016.

[35] JT. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, "Striving for simplicity: The all-convolutional net," *arXiv preprint arXiv*, 21-Dec-2014.

[36] K. Chellapilla, S. Puri and P. Simard, "High performance convolutional neural networks for document processing," *InTenth International Workshop on Frontiers in Handwriting Recognition*, 23-Oct-2006.

[37] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," *In2012 IEEE conference on computer vision and pattern recognition*, 16-Jun-2012.

[38] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," *UMass Amherst technical report*, Jan.2010.

[39] J. Yan, X. Zhang, Z. Lei and SZ. Li, "Face detection by structural models. Image and Vision Computing," 1-Oct-2014.