Mathematical Modelling of Multi-Task Cascaded Convolutional Networks: Face Detection Application

Dr S B Mohan, Associate Professor, ECE, S.A. Engineering College, drsbmohan@gmail.com

Abstract In a setting with little restrictions, it might be difficult to identify and align faces because to the wide variety of possible postures, illuminations, and occlusions. Recent research has shown that deep learning methods are capable of achieving outstanding results in both of these categories of problems. In this research, we provide a deep pipelined multi-task structure that, in order to improve the performance of the tasks, takes use of the intrinsic connection that exists between them. In specifically, our approach utilises a cascaded structure consisting of three stages of meticulously crafted deep convolutional networks. This strategy has the ability to enhance effectiveness automatically and does not need any human sample selection. While maintaining real-time performance, our approach achieves higher accuracy than the current state-of-the-art methods on the difficult FDDB as well as WIDER FACE benchmarks for face detection, as well as the AFLW standard for face alignment.

Keywords: identification of faces, mathematical programming, and convolutional neural networks are some of the terms used.

1. Introduction

In the most recent years, a number of (Duan,M., 2018), (Farfade, S.S., 2015), (Ranjan, R.,2017),(Yang, H.,2018) studies have been notably targeted toward this area of detection as well as assessment on human beings by employing face characteristics. This essentially refers to their utility in numerous applications, including the utilise of biometrics for network management, video monitoring systems, as well as other implementations related to safety. However, face detection must first be accomplished before face analysis operations can be carried out. Over the course of many decades, the problem of face recognition has been the subject of investigation in a number of research (Bell, S., 2016), (Farfade, S.S., 2015),(Li, H., 2015), (Qin, H., 2016), (Yang, S., 2015), (Yang, S., 2017). However, in spite of the important progress that has been made in the field, reliable facial recognition in such an unregulated atmosphere is still a comparatively unexplored area. This is because the aesthetic of a vast diversity of faces, which can be affected by obstruction, pose variations, low agreements, scale variants, luminance varieties, etc., is still a relatively obscure region.

Within the scope of this research study, we propose a novel approach to face recognition that makes use of (CNN) convolutional neural networks. In point of fact, the suggested technique proposes a novel CNN structure that enhances the design of the Accelerated R-CNN (Ren, S., 2015) by mixing both globally as well as local characteristics at several levels. The suggested procedure includes the following three stages: (1) feature collection by using a model that has been pre-trained, (2) the formation of a region of concern (ROI), as well as (3) the categorization of a (ROI) region of attraction as face or non-facial. The following is a condensed list of the primary contributions that this work has made: – When it came time to extract features, we made use of the ResNet50 model that had been pretrained [7]. The reality that it incorporates residual units that can extract both local and global data is one of its strongest selling points. During the process that included classifying regions of interest

(ROIs), we used a mixture of multi-scale featured maps to improve the feature arrays that were used to encode each ROI.

The remaining parts of the article are structured as described below. The second section is devoted to providing a concise summary of the works that are linked. The suggested approach is described in Section 3, which also identifies the method. The findings of the experiments are detailed in Section 4, which you may see here. The summary is presented in Section 5, which also includes some fresh ideas for consideration in future publications.

2 Related work

Face detection is a specific case of object detection. In literature, many methods have been proposed for face detection. These methods can be classified into hand craft feature-based

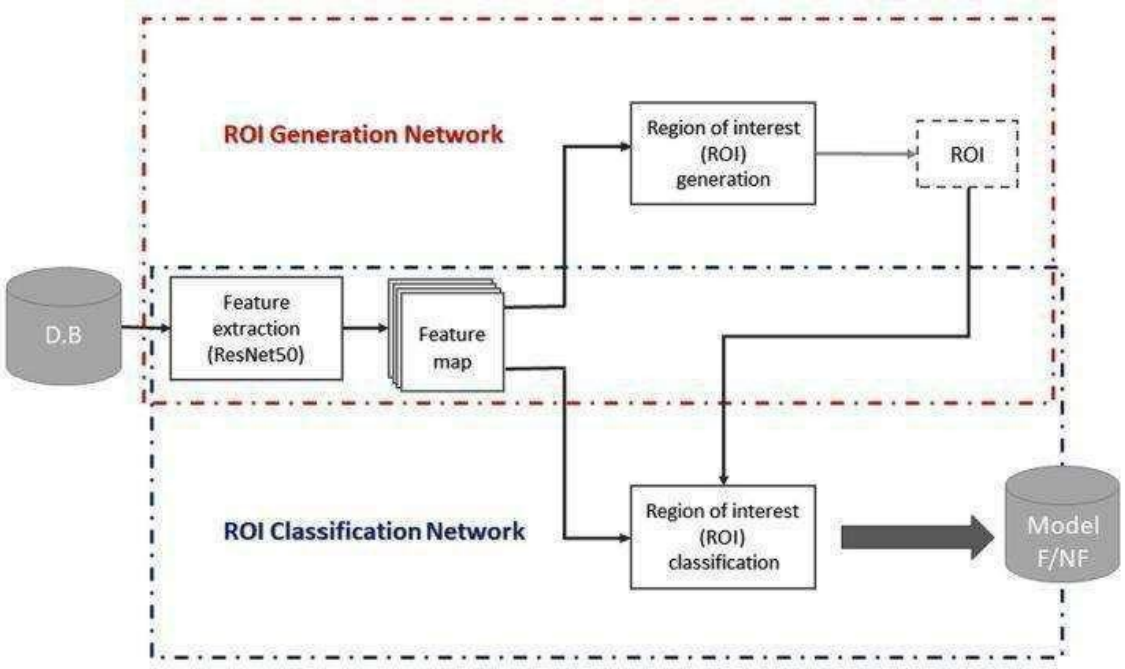

Fig. 1 The proposed method for face detection

Object recognition may be broken down into its component parts, one of which is face detection. In the available research, a great number of face detection strategies have been suggested. These techniques may be broken down into two categories: manually crafted feature-based techniques(Chen, D., 2014), (Ramanan, D., 2012), (Viola, P., 2004) as well as deep learning-based techniques that could automatically extract important characteristics. The results that can be acquired through the use of deep learning-based approaches perform noticeably better than those that can be produced through the use of hand-crafted feature-based methods (Zheng, Y., 2016). In point of fact, approaches that are based on machine learning are well-known for their ability to successfully capture complicated visual changes by using a significant quantity of training data. In furthermore, they do not need the selection of a particular feature extraction in any way. In order to solve the issue of multi-constraint face identification, many approaches based on deep learning have been developed. On the basis of the process of feature extraction, these approaches may be divided into two distinct

categories: scale-invariant characteristics techniques (Li, H., 2015),(Li, Y., 2016) as well as scale-variant characteristics techniques respectively.

Scale-invariant features methods

On the last layer of the CNN's convolution layer, the magnitude characteristics techniques gather as well as generate facial image. These operations take place inside the scale-invariant characteristics methods. (DDFD) Deep Dense Face Detector, that extends the well before AlexNet network(Krizhevsky, A., 2012) to address the challenge of face identification, was suggested by the authors in paper (Farfade, S.S., 2015) . The deep deformation face detector (DDFD) that was suggested does not need pose annotating or knowledge regarding facial landmarks, as well as it is capable of detecting faces in a broad variety of orientations utilizing a single system that is built on deeper segmentation neural systems. Regarding Yang (2015), we proposed something called the "Faceness-Net," in which the face is broken up into a number of different facial features, including the hairline, eyes, nasal, mouth, as well as beard. In point of fact, in order to build a partness mapping, five CNNs, every of which is pre-trained with the AlexNet framework , are training on each portion. The writers then integrated all of these maps into a single one in order to produce an extracted features that depicts the various parts of the facial. The Cascade-CNN were first presented by Li. (2015) using the similar framework. It is made up of six convolutional neural networks (CNNs), three of which are used for face/non-face categorization as well as three others for bounding box calibrating. Using this technique, target items are normalised into an uniform grid, and then multi-scale identification is performed using an imaging pyramids. Nevertheless, it was not enough to deal with variations in huge size and look.

A CNN as well as a 3D means face modeling are both included into a face identification approach that is described in (Li, Y., 2016) , which is referred by the term end-to-end multi-task exclusionary learner approach. This method detects faces. The pro - posed CNN is made up of two different parts: (1) the face suggestion component, which is used to produce facial expression bounding box suggestions by predicting facial key-points as well as the 3D improvement variables for every projected key-point; as well as (2) the face confirmation element, which is used for trimming as well as perfecting suggestions. Both of these parts are responsible for generating face lunging box suggestions.

Scale-variant features methods

The structure of scale variable characteristics technique, as opposed to learning sized businesses representational, mixes feature maps at various sizes as well as layers in order to extract and produce face regions.

A combined cascade network was suggested by Qin et al. (2016) for the purpose of learning scale-variant characteristics. In point of fact, examples from various scales are modelled independently by several networks, as well as the classification results are created by integrating the predictions of many networks together. The overall capacity of the network was enhanced by concurrently optimising cascaded levels, which resulted in increased performance when it came to shared convolution operation. Nevertheless, authors provided a face identification approach utilizing a single later part CNN model in (Bell, S., 2016). This method is comparable to the (SSD) Single Shot MultiBox Detection, which was published in (Liu, W., 2016). All of these techniques may be used using this technique: the multi-scale

previous layers combination, the multi-scale previous aspect ratios, and the confidence rectifying. The concatenated of multi-scale pattern extraction is an effective method for overcoming the challenges of spotting tiny faces. The computation time is cut down thanks to the multi-scale previous aspect ratios. In addition, the confidence correction works to enhance the detection performance. A deep network method referred to as "ScaleFace" was presented by Yang. (Yang, S., 2017), and it was designed specifically for the identification of faces over a broad range of sizes.

2. Proposed Methodology

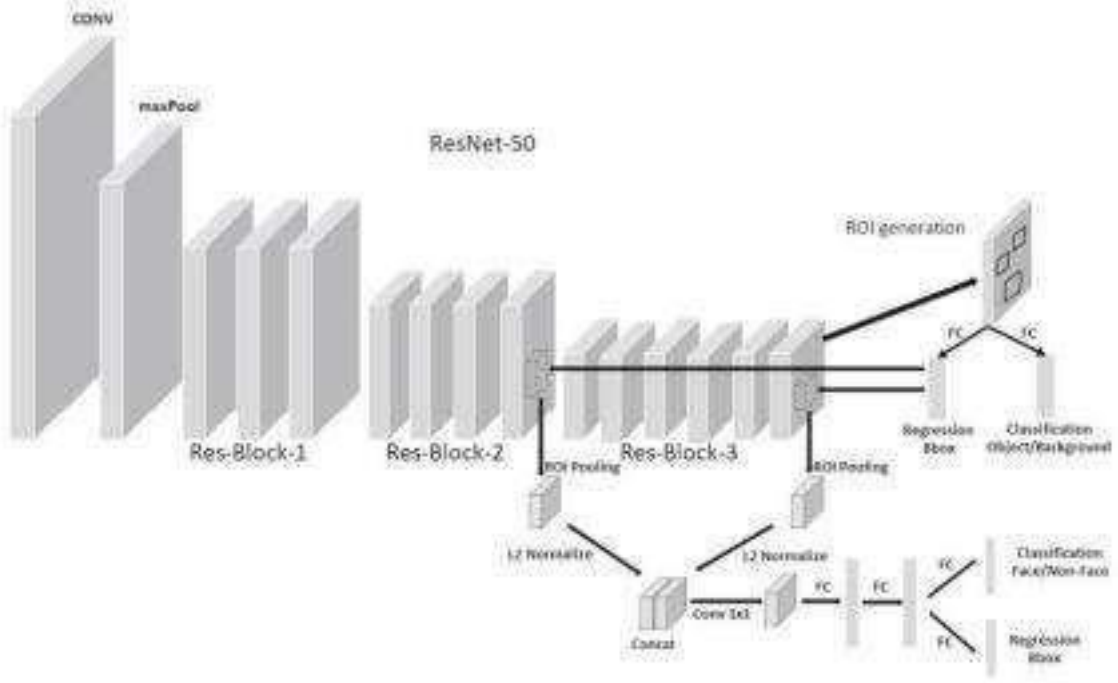

Fig. 2 Recommended Structure Chart

2.1 Feature extraction

In order to obtain face characteristics, deep learning is used since it enables the re-use of a CNN model that has already been trained as well as the utilisation of features that have been derived from a standard classification issue in order to customise them for our specific application. In contrast to other works that are dependent here on Faster R-CNN and use the well before VGG16 structure, we used the well before ResNet50 template as a fundamental network in our design. These other works are based upon that Faster R-CNN. One of the advantages of using this model is that it integrates activation functions that may extract both local as well as global data. In contrast, experimental investigations indicated that the well before ResNet50 model had a superior performance than that of other systems. The both ROI generation information as well as the ROI categorization system make use of the feature maps that are created by this approach as intimate association in their respective constructions.

2.2 Region of interest (ROI) generation

This stage receives as input a collection of characteristics from the ResNet50's most recent convolution layer and produces as output a set of regions of interest (ROIs). In order to

produce ROIs, we must first complete three subsidiary procedures, which are as follows: (1) the extraction of regions, (2) the categorization of regions into item as well as backgrounds, as well as (3) the filtering of ROIs. Exploitation of the Area In order to create region suggestions, the process of region extracting involves using a moving window to go through the most recent combination (conv) features maps that have been generated by the most recent shared convolutional layers. An approach that is based upon anchors was utilised by us so that we could have a reliable model for the scale fluctuation of ROI. As a result, we used k anchors of varying scales, every of which was positioned at a unique point inside the moving window on every one of the featured mappings. The anchors have been arranged such that each one is centred now at moving window. In our approach, we made use of the standard default structure of the classic Faster R-CNN anchoring. This arrangement is made up of nine anchors, each of which has 3 elements (128, 256, and 512) as well as 3 feature proportions (1:1, 2:1 and 1:2). Whenever a moving windows is used to traverse across the conv characteristic mappings, every place in the window yields nine anchoring. Those anchoring have been organised in the form of vectors, and they will serve as an intake for the process that involves classifying the area into item and backdrop.

Region classification into object/background

The process of classifying an area as either an item or a backdrop begins with the application of fully interconnected layers, followed by loss layers. In point of fact, the recently created anchoring are supplied into 2 fully twin FC layers, one of which is a categorization level (cls), while the other is a regressed level (reg).

2.3 Region of interest (ROI) classification

The next step is to categorise the ROIs as facing or non-face, depending on the situation. Because of this, we began by applying the ROI Pooling level, then moved on to the fully linked levels, as well as last we added the lost layer. In order to improve the ROI component vectors, we concatenated a number of different scales of featured mappings. The pooling layer is the parent layer from which the ROI pooled layers is derived; maximum pooling. Every layer receives as an entry the resulting feature mapping level of such pre-trained models as well as a listing of ROIs when the classic fast R-CNN algorithm is used. With the use of this layering, each ROI may be consolidated into a more manageable featured mapping of a predetermined size. This approach does not always optimum as well as may, at times, leave out certain significant characteristics since the featured mapping of the final conv level became less representational for tiny ROI patches . This is one of the reasons why this procedure occasionally leaves out some relevant characteristics. To get over this limitation, we recommend supplying the ROI Pooling utilizing convolutional layered characteristic mappings. These maps should make use of a variety of layers as well as scales to acquire worldwide as well as localized information.

2.4 Mathematical Modeling of CNN Architectures

There have been numerous CNNs developed for the purpose of facial recognition. Having said that, we have observed that its functionality may be constrained by the underlying facts: (1) Most filtering may not have a diverse set of weights, which may restrict their ability to create an exclusionary characterization. (2) When comparison to these similar multi-class objections recognition as well as categorization jobs, facial recognition is a

challenging binaries categorization problem. As a result, it may need fewer amounts of filtering but more discriminating of those filters than other categorization duties. In order to do this, we have decreased the amount of filtering as well as converted the 5x5 filtering into a 3x3 filtering. This allows us to decrease the amount of computation required while simultaneously increasing the depth, which results in improved achievement. When comparing to the earlier design described, these enhancements allow us to achieve higher effectiveness with a shorter amount of duration; the results are displayed in Table 1. We utilise the same information including both approaches so that we can do an objective evaluation. The structures of our CNNs may be shown in Figure 2.

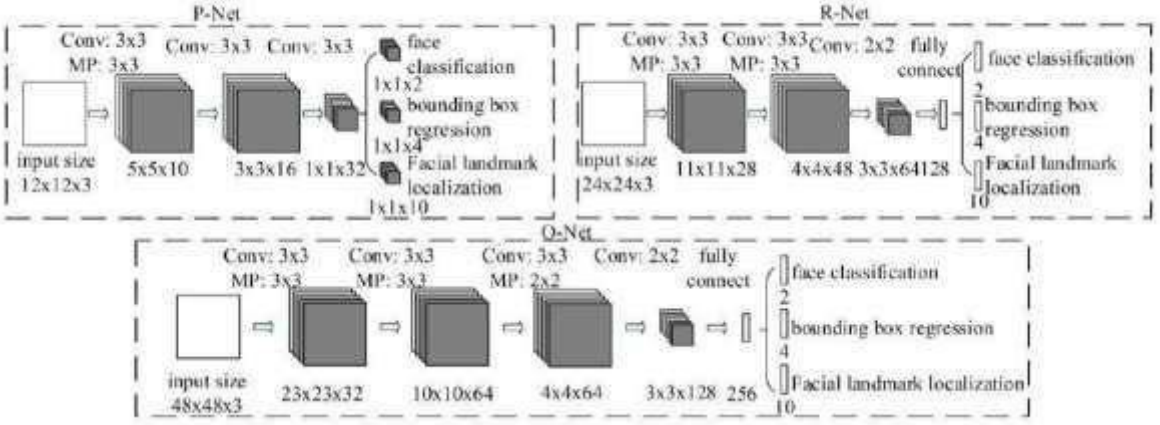

Fig. 2 P-Net, R-Net, as well as O-Net are the structures, as well as "MP" stands for "maximum pooled," while "Conv" refers to "complexity."

C. Training

In order to training their CNN detection, we use three different duties: face/non-face categorization, bounded boxes modelling, as well as facial marker localisation.

1) ***Face classification:*** The training goal is presented in the style of a challenge involving the categorization of two groups. We calculate the cross-entropy losses for every individual test as follows:

$$L_i^{det} = -(y_i^{det} \log(p_i) + (1 - y_i^{det})(1 - \log(p_i))) \quad \text{---------- (1)}$$

where is the likelihood that a sampling is a facial, as determined by the networking, and where does this likelihood come from? The title for the actual data is denoted by the expression.

2) Bounding box regression: We make a prediction about the offset that exists amongst each contender windows as well as the closest original data, which is defined as the left upper, height, as well as width of the surrounding boxed. The educational goal is presented in the manner of a regressed issue, as well as the following equation is used to calculate the Euclidean losses for every instance x i:

$$L_i^{box} = \left\| \hat{y}_i^{box} - y_i^{box} \right\|_2^2$$

---------- (2)

3) Facial landmark localization:In a manner similar to that of the boundary boxed regression job, the detecting of face landmarks is conceived of as a regressive challenge, as well as we want to minimise the Euclidean deficit:

$$L_i^{landmark} = \left\| \hat{y}_i^{landmark} - y_i^{landmark} \right\|_2^2$$

---------- (3)

wherein Y^{landmark} is just the ground truth location as well as the face landmark's position that was received from the networking. There seem to be five distinguishing features on the face, as well as they were the right eye, the left eye, the nostril, the border of the left lips, as well as the corners of the right lips.

4) Multi-source training:Facial expression, non-face, as well as partly alignment cheeks are all examples of the many kinds of training pictures used by CNNs throughout the processes of training. This is due to the fact that we implement a variety of tasks inside each CNN. In this particular instance, portions of the loss functions, namely equations (1) through (3), are not applied. As an illustration, in the case of the sample from the backgrounds area, we simply calculate, as well as the values for the remaining two loss are both set to 0. Immediately putting this into action using a sampling type indication is possible. The overarching goal of the training experience may thus be stated as:

$$\min \sum_{i=1}^{N} \sum_{j \in \{det, box, landmark\}} \alpha_j \beta_i^j L_i^j$$

---------- (4)

wherein the total amount of instances used for learning. indicates the significance of the assignment. They utilise in P-Net as well as R-Net, whereas O-Net is used for face landmarks localized that is more precise. represents an indication of the sampling kind. When training CNNs, it makes sense to use stochastic gradient descending given the circumstances presented here.

5) Online Hard sample mining:In the facial categorization problem, rather than following the conventional practise of doing hard sampling mining just after initial classifier has been learned, we execute difficult sample mined online so that we may be more flexible to the learning procedure.

In specifically, during each mini-batch, we sorted the loss calculated during the forward propagating stage from each of the instances as well as then choose the top seventy percent of those samples to use as tough instances. After then, during the period of reverse propagating, we will only calculate the gradient based on the difficult instances. This implies that we disregard the simple samples during learning since they contribute less to the overall improvement of the detection. Tests have shown that using this technique rather than manually selecting samples results in superior overall effectiveness.

4.1 The dataset and setup The suggested technique for face analysis is assessed using two different face samples (FDDBas well asPascal Faces). The FDDB facial database has 5171 face photos that have been evaluated. These face photos come from 2845 photographs that have a resolution display of 300 x 400 and were taken in the field. Pascal Face is some face datasets that sees a lot of application. There are a total of 1335 analyzed faces derived from 851 photos. Several well-known, time-tested algorithms, including Faceness,Structural Modelling, DDFD,google image Picasa, HeadHunter, as well as Face++, serve as benchmarks for evaluating our suggested solution. The ROC curve, also known as the Real Positive Rate-False Positive graph, as well as the PR curvatures, also known as the Precision-Recall curve, are both used as performance measures. On a computing device (running Ubuntu 14.04.01, Geforce GTX Titans X with 12G 4, as well as mxnet frameworks), the trials are carried out.

4.2 Experimental results

Figure 10 illustrates the curves of precision-recall for Pascal Faces datasets. It is clear that the suggested strategy is superior than both Faceness as well as DDFD in terms of roughly 2 percent points. Additionally, when tried to compare using the outcomes of facial recognition performed by several commercialized technologies (Picasa and Face++), it has a number of advantageous qualities. According to the reality that perhaps the non-face windows may be swiftly removed by the superficial network, the enhanced converters convolutional neural system has the potential to enhance the levels of accuracy of such detectors along with the efficiency with which it operates, as was noticed through their findings. The detecting outcomes of many approaches are shown in Figures 3 as well as 4, including Faceness, CCF, DDFD, DP2MFD, HeadHunter, Cascades CNN.

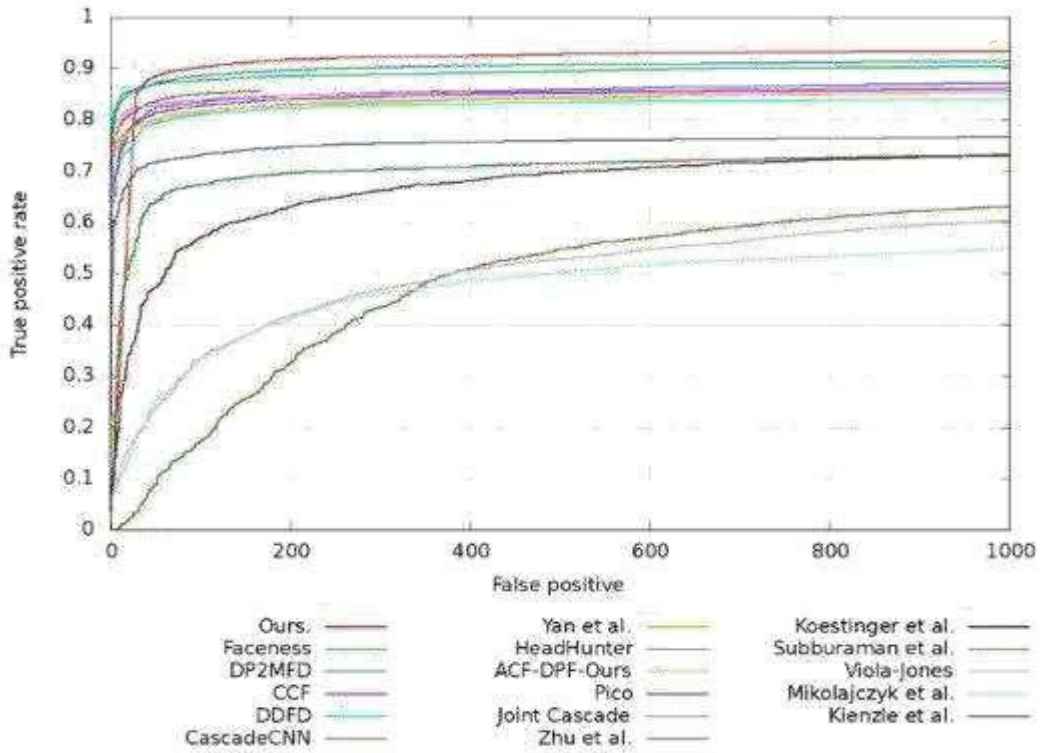

Figure 3 The disc ROC curves on FDDB face dataset

Table 1 presents a comparison of several CNN methodological approaches.

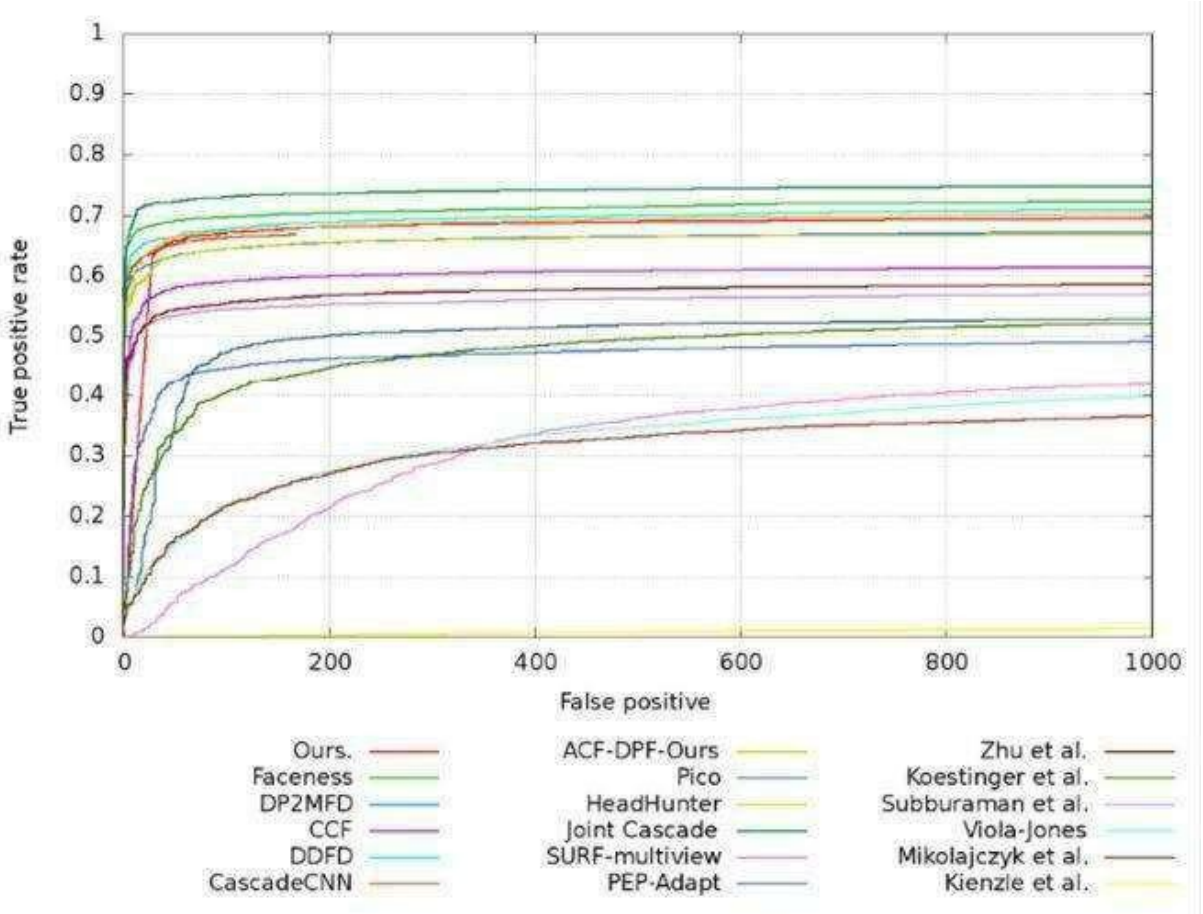

Figure 4 The contROC curves on FDDB face dataset

Table 1 : Comparison of CNN's

Group	CNN	300 Times Forward	Accuracy
Group1	12-Net	0.038s	94.4%
Group1	P-Net	0.031s	94.6%
Group2	24-Net	0.738s	95.1%
Group2	R-Net	0.458s	95.4%
Group3	48-Net	3.577s	93.2%
Group3	O-Net	1.347s	95.4%

Conclusion

The purpose of our article was to showcase a novel cascaded neural network convolutional that was developed towards face recognition. In the initial step of the process, we begin by providing a low-pixel option windows to a shallower convolutional neural system. This connectivity is able to create candidates window frames in a very short amount of time. Secondly, in order to optimize the eligible windows, the characteristics of 2 high pixels CNN models are mixed using a stacked threshold. Strong samples gathering as well as

combined practise both contribute to the overall improvement of the network. The PASCAL Faces as well as FDDB face datasets are used in the evaluation of our suggested technique. The findings of the experiments indicate that perhaps the strategy we have developed produces a higher level of effectiveness when comparing to alternative approaches.

References:

1. Bell, S., Lawrence Zitnick, C., Bala, K., Girshick, R.: Inside outside net: detecting objects in context with skip pooling andrecurrent neural networks. In: Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition, pp. 2874–2883(2016)

2. Berg, T.L., Berg, A.C., Edwards, J., Forsyth, D.A.: Who's in thepicture. In: Advances in Neural Information Processing Systems,pp. 137–144 (2005)

3. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade facedetection and alignment. In: European Conference on ComputerVision, pp. 109–122. Springer (2014)

4. Duan,M., Li, K., Yang, C., Li, K.: A hybrid deep learning cnn-elmfor age and gender classification. Neurocomputing 275, 448–461(2018)

5. Farfade, S.S., Saberian, M.J., Li, L.J.: Multi-view face detectionusing deep convolutional neural networks. In: Proceedings of the5th ACM on International Conference on Multimedia Retrieval,
pp. 643–650. ACM (2015)

6. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchiesfor accurate object detection and semantic segmentation.In: Proceedings of the IEEE Conference on Computer Vision andPattern Recognition, pp. 580–587 (2014)

7. He,K., Zhang, X.,Ren, S., Sun, J.:Deep residual learning for imagerecognition. In: Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, pp. 770 – 778 (2016)

8. Jain, V., Learned-Miller, E.: FDDB: a benchmark for face detectionin unconstrained settings.Technical ReportUM-CS-2010-009,University of Massachusetts, Amherst (2010)

9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classificationwith deep convolutional neural networks. In: Advances in NeuralInformation Processing Systems, pp. 1097–1105 (2012)

10. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutionalneural network cascade for face detection. In: Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition, pp. 5325–5334 (2015)

11. Li, Y., Sun, B., Wu, T., Wang, Y.: Face detection with end-to-endintegration of a convnet and a 3d model. In: European Conferenceon Computer Vision, pp. 420–436. Springer (2016)

12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y.,Berg, A.C.: SSD: Single shot multibox detector. In: European Conferenceon Computer Vision, pp. 21–37. Springer (2016)

13. Liu,W., Rabinovich, A., Berg, A.C.: Parsenet: looking wider to seebetter. arXiv preprint arXiv:1506.04579 (2015)

14. Qin, H., Yan, J., Li, X., Hu, X.: Joint training of cascaded CNN forface detection. In: Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, pp. 3456–3465 (2016)

15. Ramanan, D., Zhu, X.: Face detection, pose estimation, and landmarklocalization in the wild. In: Proceedings of the 2012 IEEEConference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. Citeseer (2012)

16. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: a deep multitasklearning framework for face detection, landmark localization,pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. 41(1), 121–135 (2017)

17. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards realtimeobject detection with region proposal networks. In: Advancesin Neural Information Processing Systems, pp. 91–99 (2015)

18. Sun, X., Wu, P., Hoi, S.C.: Face detection using deep learning:an improved faster RCNN approach. Neurocomputing 299, 42–50(2018)

19. Szegedy, C., Liu,W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D.,Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions.In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)

20. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput.Vis. 57(2), 137–154 (2004)

21. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by deexpressionresidue learning. In: Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition, pp. 2168–2177 , (2018)

22. Yang, S., Luo, P., Loy, C.C., Tang, X.: From facial parts responsesto face detection: a deep learning approach. In: Proceedings ofthe IEEE International Conference on Computer Vision, pp. 3676–3684 (2015)

23. Yang, S., Luo, P., Loy, C.C., Tang, X.:Wider face: a face detectionbenchmark. In: Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, pp. 5525–5533 (2016)

24. Yang, S., Xiong, Y., Loy, C.C., Tang, X.: Face detectionthrough scale-friendly deep convolutional networks. arXiv preprintarXiv:1706.02863 (2017)

25. Zheng, Y., Zhu, C., Luu, K., Bhagavatula, C., Le, T.H.N., Savvides,M.: Towards a deep learning framework for unconstrainedface detection. In: 2016 IEEE 8th International Conference on BiometricsTheory, Applications and Systems (BTAS), pp. 1–8. IEEE(2016)