Facial Detection and Recognition Among Heterogenous Multi Object Frames

Alka Londhe, Kuldeep Mehta, Ashitosh Bhinge, Aditya Deshmukh
Department of Computer Engineering,
Pcooe, Pune, India.

***Abstract*— An individual's face plays a really important role in social interaction, conveying a person's identity and has a high degree of variability in its appearance. To solve the problem of this variance. Currently discovering different ways to find or identify facial features have been a really active research field. Different applications are being developed using different face detection techniques such as Convolution Neural Networks (CNN), Region Based Convolution Neural Networks (RCNN), You Only Look Once (YOLO), Single Shot Detection (SDD) can be used for applications such as live face filters, face recognition-based screen unlocking systems, etc. These applications also use pattern recognition techniques such as Linear Discriminant Analysis (LDA), Principal component Analysis (PCA), Histogram Oriented Graph (HOG) can be combined with .Both face detection and face recognition techniques can be combined to efficiently detect faces and give really accurate and faster results which can be used in security purposes, interactive gaming, human computer interaction.**

Keywords—Faces, CNN, YOLO, RCNN, Hog, Patterns, Secutrity, Filters, Detection, Recognition.

I. INTRODUCTION

A face detection includes classifying image into two classes: one with faces (targets), and other containing the background (clutter) which needs to be removed. Commonalities exist between faces, they vary differently in terms of age, skin color and facial expression, this becomes difficult due to this commonality. The further problem is complicated by differing lighting conditions, image qualities and geometries, partial occlusion and disguise is also a possibility. A face detector should be able to detect the presence of any face under different set of lighting conditions in any background condition. The face detection analysis can be broken into two tasks. The first task is a classification task that takes some random regions of image as input and outputs a binary value of yes or no, indicating if there are any faces present in the image. The other task is the face localization task which is to take an image as input and output the location of any face or faces within that image as some bounding box/boxes with (x, y, width, height).

Smart robots can be built by automatic facial expression application. These bots can be used in various applications like interactive games and service center. There are six universal expressions according to Ekman they are fear, disgust surprise, anger, sadness and happiness. Face variances can be observed to recognize these expressions. For example, we can say a person is happy which can be identified as a gesture of smile by tightened eyelids and raised lips corners. A person's internal states, social communication and intentions are indicated by change in facial expressions. Many applications in many areas like human emotions analysis, natural Human computer interaction, image retrieval and talking bots have a large effect on them by automatic facial expression detection.

Face Recognition with Histogram of Oriented Gradients using CNN detection has been an impacting issue in the technological community as human beings fined facial expressions one of the most natural and powerful means to express their intentions and emotions. Last stage of the system is facial expression detection. There are basically three steps in training procedure in expression recognition systems named as feature learning, classifier construction and feature selection. Feature learning stage is first, feature selection is second and the last one is classifier construction. Only learned facial expressions variations among all features are extracted after feature learning stage. Facial expression is then represented by the best features which are chosen by feature selection. Not only maximizing inter class variation but they also should try to minimize the intra class variations of expressions not only maximizing inters class variation but they also should minimize the intra class variations of expressions. Because same expressions of different individuals in image are far from each other in pixel's space so minimizing the intra class variation of expressions is a problem.

Techniques that can be used for facial detection are YOLO, SDD, RCNN, Faster RCNN.

II. RELATED WORK

Several approaches have for the face detection and recognition. One of the earlier approaches for addressing the face detection task was proposed in [10] and later extended in [11] and many more. A cascade-based method was recently proposed in [12] for detecting faces on a given image. This method improved detection performance by takes in face alignment step in a cascade structure.

CNN works well for many different vision related tasks by deeply learning the features of data. For instance, Alex et al. [13] and He et al. [14] use CNN to classify the ImageNet dataset [8] and achieve the accuracy of 83% and 94.3%, respectively. Farabet et al. [15] show the impressive performance of CNN on scene labeling. R-CNN [17] improves mean average precision (map) by more than 30% relative to the previous best result on the VOC 2012 dataset [16], and it achieves a map of 53.3%. Although CNN has shown good performance for object detection, its computational efficiency is not high. OverFeat [18] obtains the multi-scale dense features by using the sliding window strategy. Although since the detection task of OverFeat requires running the classifier and regressor networks across all the possible locations and scales this task in OverFeat consumes a ton of time. To improve on this efficiency, R-CNN first generates many class-independent proposal windows(regions), and then extracts features on the regions with the trained CNN on a multi-scale images. Afterwards, a score is assigned to each proposal window by applying a linear SVM (Support Vector Machines) to the features. R-CNN takes about 15 seconds to run a detector on a 500 × 375 image. Reading and capturing of the images and videos is done using OpenCV and web service for image registration using images or live camera has also been proposed in the system. This method currently achieves the best performance on images and live camera yet.

Table1: Comparison between face detection techniques

METHOD NAMES	Frames per second	Speed second/image
R-CNN	.05fps	20s/image
Fast R-CNN	.5fps	2s/image
YOLO	45fp	22ms/image

III. TECHNIQUES

Yolo (you only look once) is used to design an object detection algorithm for a real time video and also in real time it is used to detect what objects are where. It is one of the best and new technique for multiple object detection. Yolo takes a completely different approach to detect an image, it looks at the image just once but in a clever way. Yolo divides the image into a grid of 13 by 13 cells. Suppose an image is displayed then usually we have to detect all the classes that are in the image. Each of this cells i.e. of the 13 by 13 cells is responsible for predicting five bounding boxes. Yolo also outputs a confidence score that tells us that how certainly that predicted bounding box actually encloses some object. The score does not say anything about what kind of object is int that box, it describes the shape of the box Higher the score thicker is the bounding boxes which tells that something significant in there. Yolo was trained on PASCAL VOC dataset ,which can detect 20 different classes such as Bicycle,car,cat,dog,boat,person etc.Yolo has 845 total bounding boxes,so from that we have to keep the boxes which gives us the best result.Yolo takes an input image,the image goes through a convolution neural network through one pass,and comes out as an 13x13x125 tensor describing the bounding boxes for the cells.Then we have to compute the final scores for the bounding boxes and then throw away the one scoring less than30%.Prior detection systems repurpose classifiers or localizers to perform detection.

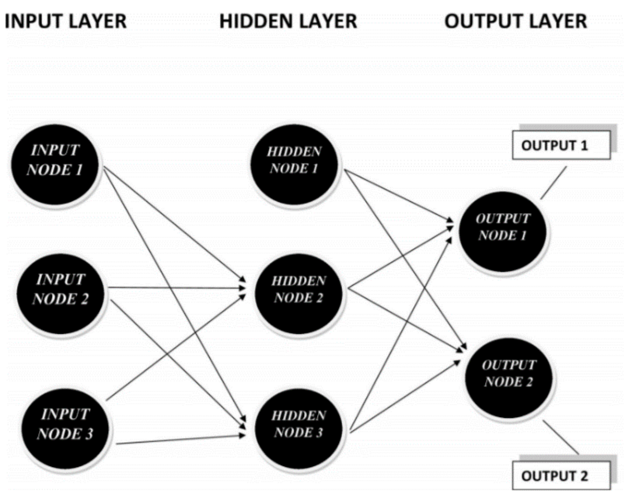

Fig.1 Neural network layers

They apply the model to an image at multiple locations and scales. High scoring regions of the image are considered detections. We use a totally different approach. We apply a single neural network to the full image. This network divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities.

Fig.2 R-CNN Network Architecture

Region-CNN (R-CNN) is one of the state-of-the-art CNN-based deep learning object detection approaches.

To have object detection, we need to know the class of object and also the bounding box size and location.Conventionally, for each image, there is a sliding window to search every position within the image as below.It is a simple solution. However, different objects or even same kind of objects can have different aspect ratios and sizes depending on the object size and distance from the camera.And different image sizes also affect the effective window size. This process will be extremely slow if we use deep learning CNN for image classification at each location.First, R-CNN uses selective search by [2] to generate about 2K region proposals, i.e. bounding boxes for image classification.Then, for each bounding box, image classification is done through CNN.Finally, each bounding box can be refined using regression.R-CNN is slow because it performs a ConvNet forward pass for each object proposal, without sharing computation. And that’s why Faster R-cnn came in existence.

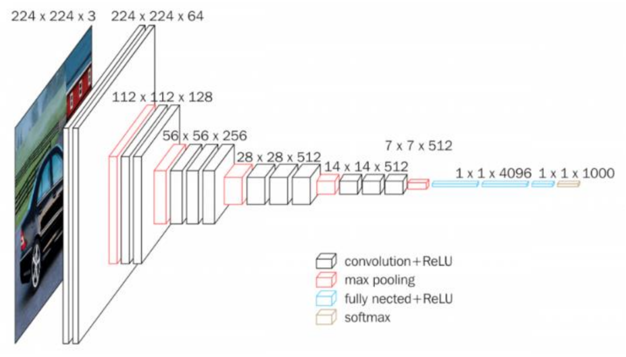

Fig 2. Architecture of YOLO technique

Some of the drawbacks of the R-CNN was solved to build a faster object detection algorithm and it was called Fast R-CNN. The approach is similar to the R-CNN algorithm. But, instead of feeding the region proposals to the CNN, we feed the input image to the CNN to generate a convolutional feature map. From the convolutional feature map, we identify the region of proposals and warp them into squares and by using a RoI pooling layer we reshape them into a fixed size so that it can be fed into a fully connected layer.The reason “Fast R-CNN” is faster than R-CNN is because you don’t have to feed 2000 region proposals to the convolutional neural network every time. Instead, the convolution operation is done only once per image and a feature map is generated from it.

IV. DATASET ANALYSIS AND PERFORMANCE

The proposed dataset contains 460 videos of 252 subjects captured during the day (visible spectrum) and night (NIR spectrum). All videos capture the subjects (individually or in groups) in unconstrained settings, at a varying distance of 1ft to 36ft.From the available datasets for training and testing the machine for face detection. It can be observed that only CrowdFaceDB [7], SN-Flip [2], CSCRV [2], and IJB-S [13] datasets have multiple subjects in video.Videos in SN-Flip dataset contain subjects having less movement, whereas in CrowdFaceDb are captured using hand-held devices such cell phones.Also almost all the datasets contain videos captured in visible spectrum only, without much variation of the subject distance from the camera. To the best of our knowledge, only the CSCRV dataset [9] contains videos captured across the two spectra (visible and NIR [Near Infrared Spectroscopy]) having multiple subjects per video.

Each of the videos contains 1 to 4 subjects walking from a distance of 36f t towards the camera. The subjects were asked to walk freely in an unconstrained manner, without any restriction on pose, or body movement. Videos have been captured at four locations namely:3 outdoor and 1 indoor; during the day and night time, across multiple sessions.

Table 2: Analysis of datasets

Database	Spectrum	Subjects per Video	Number of Subjects per video	
Face in Action [4]	VIS	Single	180	6470
YouTube Faces [5]	VIS	Single	1595	3425
ChokePoint [6]	VIS	Single	54	48
PaSC [7]	VIS	Single	262	2802
SN-Flip [2]	VIS	Single	190	28
McGillFaces [6]	VIS	Multiple	60	60
CrowdFaceDB [7]	VIS	Single	257	385
CSCRV [9]	VIS&NIR	Multiple	160	193
IJB-S [8]	VIS	Multiple	202	350
Proposed Dataset	VIS&NIR	Multiple	252	460

Performance:
Considering frames per second as a performance measure for different techniques such as R-CNN, SDD, Faster R-CNN, YOLO and YOLO 9000(YOLOv2) it can be seen in fig.4 YOLOv2 works faster and gives more accurate results as compared to other techniques.

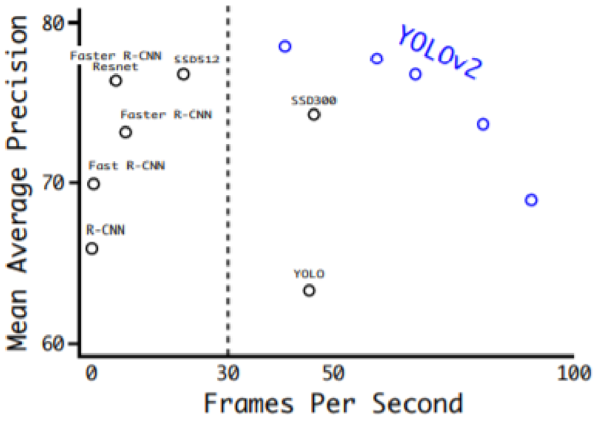

Fig.4 Performance Graph

V. CONCLUSION

In this study we have analyzed different techniques for face detection and pattern recognition techniques on different databases such as FIA,Chokepoint etc. and have thus concluded that Yolo algorithm has one of the most presentable architecture and easy to understand and provides one of most precise and efficient results.Since,it provides the desired output of the provided input image only in one pass of the convolution network. Moreover, different techniques such as LDA, HOG, CNN can be used for recognition purposes.

VI. REFERENCES

[1] Jumani, S.Z., Ali, F., Guriro, S., Kandhro, I.A., Khan, A. and Zaidi, A., 2019. Facial Expression Recognition with Histogram of Oriented Gradients using CNN. *Indian Journal of Science and Technology*, *12*, p.24.

[2] J. R. Barr, L. A. Cament, K. W. Bowyer, and P. J. Flynn. Active clustering with ensembles for social structure extraction. In Winter Conference on Applications of Computer Vision, pages 969–976, 2014

[3] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).

[4] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).R. Goh, L. Liu, X. Liu, and T. Chen. The CMU face in action (FIA) database. In International Conference on Analysis and Modelling of Faces and Gestures, pages 255–263. 2005.

[5] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In Computer Vision and Pattern Recognition, pages 529–534, 2011

[6] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patchbased probabilistic image quality assessment for face selection and improved video-based face recognition. In Computer Vision and Pattern Recognition Workshops, pages 74–81, 2011.

[7] B. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In Biometrics: Theory Applications and Systems, pages 1–8, 2013

[8] N. D. Kalka, B. Maze, J. A. Duncan, K. A. O Connor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. IJB–S: IARPA Janus Surveillance Video Benchmark. In IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2018.

[9] M. Singh, S. Nagpal, N. Gupta, S. Gupta, S. Ghosh, R. Singh, and M. Vatsa. Cross-spectral cross-resolution video database for face recognition. In IEEE International Conference on Biometrics Theory, Applications and Systems, 2016

[10] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2879–2886.

[11] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, “Joint cascade face detection and alignment,” in Proc. Eur. Conf. Comput. Vis., 2014, vol. 8694, pp. 109–122.

[12] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 2650–2658

[13] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[15] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[17] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580–587.

[18] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, International Conference on Learning Representations.

[19] DESivaram, M., Porkodi, V., Mohammed, A.S. and Manikandan, V., 2019. Detection Of Accurate Facial Detection Using Hybrid Deep Convolutional Recurrent Neural Network. *ICTACT Journal on Soft Computing*, *9*(2).

[20] Sivaram, M., Porkodi, V., Mohammed, A.S. and Manikandan, V., 2019. Detection of Accurate Facial Detection Using Hybrid Deep Convolutional Recurrent Neural Network. *ICTACT Journal on Soft Computing*, *9*(2).

[21] Shafiee, M.J., Chywl, B., Li, F. and Wong, A., 2017. Fast YOLO: a fast you only look once system for real-time embedded object detection in video. *arXiv preprint arXiv:1709.05943*.

[22] Ravidas, S. and Ansari, M.A., 2019. An Efficient Scheme of Deep Convolution Neural Network for Multi View Face Detection. *International Journal of Intelligent Systems and Applications*, *11*(3), p.53.

[23] Gupta, S., Gupta, N., Ghosh, S., Singh, M., Nagpal, S., Vatsa, M. and Singh, R., 2019. FaceSurv: A benchmark video dataset for face detection and recognition across spectra and resolutions. *challenge*, *14*, p.18.

[24] Guo, G., Wang, H., Yan, Y., Zheng, J. and Li, B., 2019. A Fast Face Detection Method via Convolutional Neural Network. *Neurocomputing*.