Deep learning for pose-invariant face detection in unconstrained environment

Shivkaran Ravidas, M. A. Ansari
Department of Electrical Engineering, School of Engineering, Gautam Buddha University, India

Article Info

Article history:

Received Mar 5, 2018
Revised Aug 6, 2018
Accepted Aug 20, 2018

Keywords:

Computer vision
Convolutional neural network (CNN)
Deep learning
Face detection
Multi-view face detection
Pose invariant faces

ABSTRACT

In the recent past, convolutional neural networks (CNNs) have seen resurgence and have performed extremely well on vision tasks. Visually the model resembles a series of layers each of which is processed by a function to form a next layer. It is argued that CNN first models the low level features such as edges and joints and then expresses higher level features as a composition of these low level features. The aim of this paper is to detect multi-view faces using deep convolutional neural network (DCNN). Implementation, detection and retrieval of faces will be obtained with the help of direct visual matching technology. Further, the probabilistic measure of the similarity of the face images will be done using Bayesian analysis. Experiment detects faces with ±90 degree out of plane rotations. Fine tuned AlexNet is used to detect pose invariant faces. For this work, we extracted examples of training from AFLW (Annotated Facial Landmarks in the Wild) dataset that involve 21K images with 24K annotations of the face.

Corresponding Author:

Shivkaran Ravidas,
Department of Electrical Engineering, School of Engineering,
Gautam Buddha University,
Gautam Buddha Nagar, Uttar Pradesh, India.
Email: mailmekaran@gmail.com

1. INTRODUCTION

We can define face detection as the process of extracting faces from the given images. Hence, the system should positively identify a certain region as a face. According to Yang *et al.* and Erik Hjelmas *et al.*, face detection is a process of finding regions of the input image where the faces are present [1-2]. A lot of work has been done in detecting faces in still and frontal faces in-plane as well as complex background [3]. With the advancement in the field of information technology and computational power, computers are more interactive with humans. This human-computer interface (HCI) is done mostly via traditional devices like mouse, keyboard, and display. One of the most important medium is the face and facial expression [4], [5].

There are several algorithms that address frontal face detection [6] but only a small number of techniques exist that addresses non-frontal or multi-view face detection [7]. Most of the techniques uses scanning the image with sub window and then classify the sub window as a face and non-face pattern. The statistical learning methods are used for classification. This is because the pixels on faces are highly correlated while in non-face sub window they have less regularity. Hence use of nonlinear classifier is necessary due to huge variations in lightning and illumination, face expression, pose or appearance variations. Examples of such techniques are neural networks [8] or Support Vector Machines [9]. They used two neural network classifiers, first one for pose estimation and second for conventional face detection. Schneiderman *et al.* [10] proposed a technique that detects faces with out-of-plane rotation. In [11], Jones and Viola [12] extend this framework. Convolutional Neural Networks (CNN) [13], are the most recent cascade framework with quick rejection of background regions. The amount of research works on multi-view face detection making use of CNNs is exploding [14, 15], success of CNNs in many computer vision problem.

The CNNs can be visualized as a series of layers. The initial set of layers respond to discriminative low level patterns. The next set of layers respond to intermediate patterns which are composed of low level patterns and so on. The inspiration for CNNs and neural networks in general has been the biological understanding of the brain. It has been known for quite some time that the brain is made of over 100 billions neurons and these neurons are densely connected. The CNNs mimic neurons and their connections. A layer in CNN is made up of m × n neurons and neurons of the neighbouring layers are connected. In this section, we will describe about neurons, the connections and various types of layers that the modern CNNs have. Zhang *et al.* studied about enhancing multi-view detection of a face with multi-task deep CNN [16]. Farfade et al. [17] conducted a research to examine multi-view detection of the face using deep CNN. According to Parkhi *et al.* the recognition of the face from either a set of faces or single photograph tracked in a video [18]. Li *et al*. analyzed about CNN cascade for detecting the face [19].

Detecting face is a well-studied problem in the vision of computer. Contemporary detectors of the face can effortlessly identify near front faces. Complexities in detecting the face come from two aspects such as large space for searching of probable face sizes, positions and large visual differences of human faces in a chaotic environment. Former one imposes a requirement for the efficiency of time while later one needs a detector for a face to perfectly addressing a binary issue in classification. It was noted that uncontrolled issue in detecting face are extreme illuminations and exaggerated expressions can lead to large differences in visual in the appearance of the face and affect the face detector robustness [20]. This is significant to develop a method to properly detecting the faces as pointed out in [21]. Therefore, this particular research intends to concentrate on detecting the face with the help of multi-view face using deep convolution neural network.

In this work, we have presented a novel architecture of deep convolutional neural network (DCNN) for multi-view face detection. In most of the previous work feature selection was manual, that is handcrafted but in convolutional neural network feature selection is automatically, even in complex visual variations. As we know that CNNs need huge computational power because it requires exhaustively scanning of the entire image in multiple scales which is a bit difficult. Hence to speed up the detection, we proposed a CNN cascade structure which rejects false detection very quickly in early stage. The most prominent contribution of our work is as follows:

1. We designed a CNN cascade for fast face detection.
2. Our designed architecture is able to detection pose invariant faces in changing environment.
3. Our design is able to handle multi resolution images.
4. We improve the state-of-the-art performance on the face detection data set and benchmark.

2. IMPLEMENTAION METHOD

In the implementation, detection of the face and retrieval of the image will be attained with the help of direct visual matching technology. A probabilistic computation of resemblance among the images of the face will be conducted on the basis of the Bayesian analysis for achieving various detection of the face. After this, a neural network will be developed and trained in order to enhance the outcome of the Bayesian analysis. Next, to that, training and verification will be adapted to test other images which involve similar face features. Deep learning can be performed by supervisory signals.

$$\mathrm{Ident}(f,\mathrm{t},\emptyset_{\mathrm{id}}) = -\int \sum_{\mathrm{i=1}}^{\mathrm{n}} \log \hat{p}_i \quad = -\log \hat{p}_t \qquad (1)$$

Where, f is the feature vector, t represents target class and $\emptyset_{\mathrm{id}}$ is softmax layer parameter, p_i is the target probability distribution (p_i=0 for all i except p_t=1). $\hat{p}_i$=1 is the predicted probability distribution. The verification signal regularize feature and reduces intra personal variations given by Hadsell et al. [22].

$$Verif(f_i,f_j,y_{ij},\emptyset_{ve}) = \begin{cases} \frac{1}{2}\left\|f_i - f_j\right\|_2^2 & if\ \ y_{ij} = 1 \\ \frac{1}{2}\max(0,m - \ \left(\left\|f_i - f_j\right\|_2\right)^2 & if\ \ y_{ij} = -1 \end{cases} \qquad (2)$$

$$Verif(f_i,f_j,y_{ij},\emptyset_{ve}) = \frac{1}{2}\left(y_{ij} - \sigma(wd + b)\right)^2 \qquad (3)$$

Where, $\emptyset_{ve} = \{w, b\}$; are denote shifting parameters and learning scaling, $\boldsymbol{\sigma}$ represented as sigmoid function and y_{ij} is denoted as binary target of two compared facial images relate to same identity. Further operation of the convolution is represented as:

$$y^{j(r)} = \max(0, b^{j(r)} + \sum_i k^{ij(r)} * x^{i(r)} \tag{4}$$

Where, x^i is input map and y^j is output map, k^{ij} is the convolution between input and output. Maxpooling is given by:

$$y^i_{j,k} = \max_{0<m,n<s} \{x^i_{j.s+m,k.s+n}\} \tag{5}$$

Where, output map pools over $s \times s$ non-overlapping region.

$$y_j = max\left(0, \sum_i x_i^1 . w_{i,j}^1 + \sum x_i^2 . w_{i,j}^i + b_j\right) \tag{6}$$

Where,x^1,w^1,x^2, w^2 represent the neurons and weights in 3rd and 4th convolutional layers. Output of ConvNet is n-way software to predict the distribution of probability over n-unique identities.

$$y_i = \frac{\exp(y_i')}{\sum_{j=1}^{n} \exp(y_j')} \tag{7}$$

DCNN is mostly adopted for classification and also adopted for detection and recognizing the face. Most of them consider the cascade strategy as well as consider batches with various locations and scales as inputs.

2.1. Proposed algorithm for deep convolutional neural network (DCNN)

This particular work develops an algorithm for detecting the face using multi-view with the help of deep convolution neural network. The steps of implementation are described below:

Step 1: In the implementation, detection of face and retrieval of the image will be attained with the help of direct *visual matching technology* which matches the face directly. This technology makes use of similarity metrics of an image which can either be normalized correlation or it can be Euclidean distance, which corresponds to the approach called *template matching*. The similarity between the two images is measured through *similarity measure*, denoted by $S(I_a, I_b)$, Where, I_a and I_b are the two images between which the similarity is being measured.

Step 2: The next step is measuring probabilistic similarity or Δ (the measure of intensity difference between the two images) given by Probabilistic similarity or Δ= $(I_a - I_b)$. This calculation of resemblance among the images of face will be conducted on the basis on the Bayesian analysis for achieving various detection of face.

Step 3: The probabilistic calculation of resemblance also supports multiple face detection. In order to characterize the various types of image variations were used for statistical analysis. Under this the similarity measure S (I_a, I_b) between the pair of images I_a and I_b is given in terms of posteriori probability (interpersonal variation) is provided by:

$$S(I_a, I_b) = P(\Omega_I)P(\Omega_I|\Delta)/\{P(\Omega_I)P(\Omega_I|\Delta) + P(\Omega_E)P(\Omega_E|\Delta)\} \tag{8}$$

If the multi-view face detection is done for a single person then $P(\Omega_I|\Delta) > P(\Omega_E|\Delta)$ or it can be said that $S(I_a, I_b) > ½$.

Step 4: Further a neural network will be developed and trained in order to enhance the outcome from the Bayesian analysis.

Step 5: Next to that, training and verification will be adopted to test other images which involve similar face features. Implementation of the code is done step by step as follows:

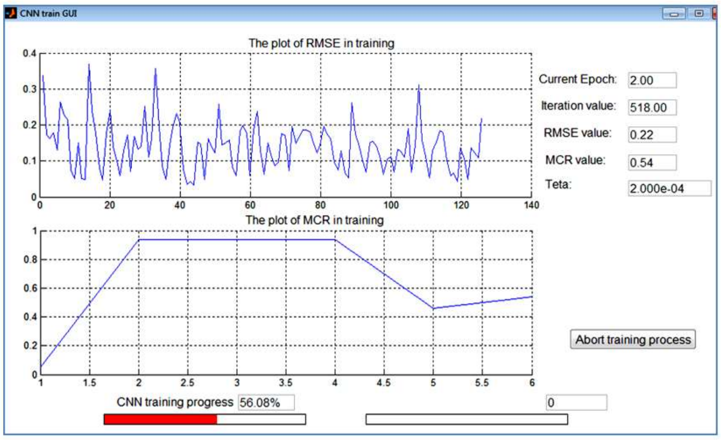
a. First, the DCNN object is created.
b. Second, after this Graphical user interface is initialized.
c. Then MCR (Misclassification rate) calculation is initialized and plot of MCR id created defining the current epoch, iteration, RMSE (Root Mean Square Error), MCR value of the image data.
d. Training data is being loaded.
e. Training data is pre-processed, errors are deleted and then image data is simulated.

f. After the simulation, the multi-faces are detected in the image shown in the red rectangular boxes.

The screenshot for CNN training progress is shown in Figure 1. The plot of RMSE in training and plot of MCR is also shown in CNN training progress. The below equation is the CNN which is trained to minimize the risk of soft max loss function.

$$R = \sum_{x_i \in \beta} \log | prob(x_i|y_i) | \tag{9}$$

Here ‘β’ represents the batch used in iteration of stochastic gradient descent and label is $'x_i'$ and $'y_i'$. Hessian calculation progress is started. Current epoch used for this is 3.00. Iteration value used for this research is 759.00. RMSE value used for this research is 0.18. MCR value used for this research is 0.90. Here ‘theta’ used is 8.000e^{-05}. Plot of RMSE in training is showed in zigzag lines. Plot of MCR in training is showed in curved lines.



Current Epoch: 2.00

Iteration value: 518.00

RMSE value: 0.22

MCR value: 0.54

Teta: 2.000e-04

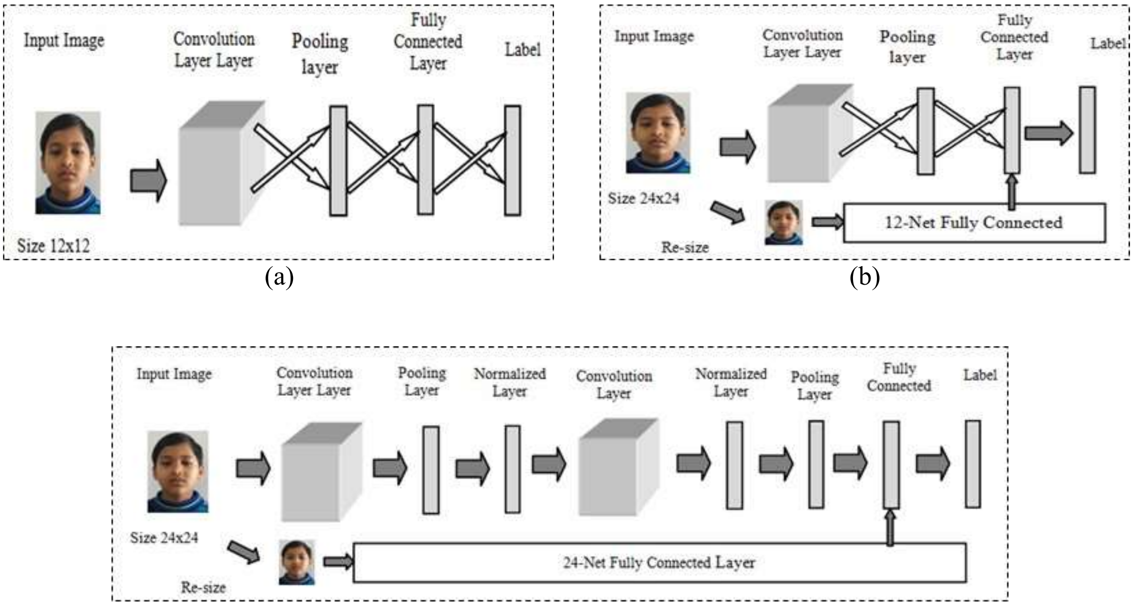
CNN training progress 56.08%

0

Figure 1. DCNN training process

2.2. CNN Structure

The CNN structure which is adopted in the present study is shown in Figure 2 which consists of 12-net CNN, 24-net and 48-net structure.



(a)

(b)

(c)

Figure 2. CNN structures of the (a) 12-net, (b) 24-net and (c) 48-net

a. 12-net CNN

It is the first CNN that scans or tests the image quickly in the test pipeline. An image having the dimensions of $w * h$ having the pixel spacing of 4 with 12x12 detection windows for such type of image 12-net CNN is suitable to apply. This would result a map of:

$$\left(\left(\frac{W-12}{4}+1\right) \times \left(\frac{H-12}{4}+1\right)\right) \tag{10}$$

A point on the image map defines detection window of 12x12 onto the testing image. The minimum size of the face acceptable for testing an image is 'T'. Firstly an image pyramid is built through the test image in order to cover the face from varied scales. At each level an image pyramid is created, it is resized by $12/T$ which would serve as an input image for 12-net CNN. Under this structure, 2500 detection windows are created as shown in Figure 1.

b. 12-Calibration-net

For bounding box calibration, 12-calibration-net is used. Under this the dimension of the detection window is $(x,\ y,\ w,\ h)$ where $'x'$ and $'y'$ are the axis, $'x'$ and $'h'$ are the width and height respectively. The calibration pattern adjusts itself according to the size of the window is:

$$(x - \frac{x_n w}{s_n}, y - \frac{y_n h}{s_n}, \frac{w}{s_n}, \frac{h}{s_n}) \tag{11}$$

In the present study number of patterns i.e. N=45. Such that:

$$s_n \quad \in \{0.87, 0.95, 1.2, 1.13, 1.25\}$$

$$x_n \quad \in \{-0.19, 0, 0.19\}$$

$$y_n \quad \in \{-0.19, 0, 0.19\}$$

The image is cropped according to the size of detection window that is 12*12 which would serve as an input image to 12-calibration-net. Under this CNN average result of the patterns are taken because the patterns obtained as an output are not orthogonal. A threshold value is taken i.e. t in order to remove the patterns which are not the confidence patterns

c. 24-net CNN

In order to further lower down the number of the detection windows used, a binary classification of CNN called 24-net CNN is used. The detection window which remained under the 12-calibration net are taken and then resized to 24*24 image and then this image is re-evaluated using 24-net. Also under this CNN, a multi-resolution structure is adopted, through this, the overall overhead of the 12-net CNN structure got reduced and hence the structure becomes discriminative.

d. 24-Calibration-net CNN

It is another calibration CNN similar to that of 12-calibrationnet. Also under this number of calibration patterns are N. the process of calibration is similar to that of 12-calibration-net.

e. 48-net CNN

It is the most effective CNN used after 24-calibration-net but is quite slower. It follows the same procedure as in 24-net. This procedure used in this CNN is very complicated as compared to rest of the CNN substructures. It also adopts the multi-resolution technique as in case of 24-net.

f. 48-calibration-net CNN

It is the last stage or sub-structure of CNN. The number of calibration patterns used is same as in case of 12-calibration-net i.e. N=45. In order to have more accurate calibration, pooling layer is used under this CNN sub-structure.

3. RESULT AND DISCUSSION

Examples of the input images for two different identities with generated pose invariant output results are illustrated in Figure 3. In this figure, detected face for the various angle and poses for left and right profile faces including the frontal face are shown. Our detector gives results for images with varying poses with resolution. The modern face detection solutions performance on multi-view face set of data is unsatisfactory. Under this it was observed that in the presence of multi-resolution in CNN which is shown in Figure 5, number of false detection comes to halt (at the 10000 number of falsely detected faces) and the face is detected or the detection rate is achieved.

However, without the use of multi-resolution in CNN, more number of faces are detected falsely as compared to that of multi-resolution shown in Figure 4. Examples of the input images for two different identities with generated pose invariant output results are illustrated in Figure 3. In this figure, detected face for the various angle and poses for left and right profile faces including frontal face are shown.

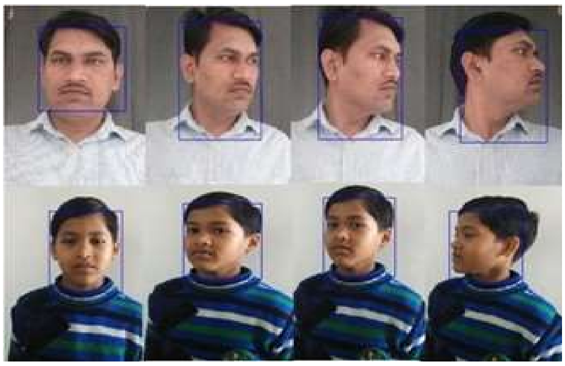


Figure 3. Pose invariant face detected images;
(a), (b), (c) and (d) are right profile faces; (e) Frontal Face; (f) Left up profile face and
(g), (h) Right profile faces

3.1. Comparison of Face Detectors

Effectiveness of the developed method is compared and contrasted with existing methods and techniques. It was noted that proposed method performs well in terms of accuracy and the recognition rate. We compare our method with other approaches including EdgeBox [23], Faceness [24], and DeepBox [25] on AFLW data set. Our method detects the input image at low resolution by rejecting quickly non-face regions for accurate detection. The use of Calibrated nets in the cascade improves the quality of bounding box. Meanwhile, we show that our detector can be easily tuned to be a faster version with minor performance decrease. The use of multi-resolution in CNN, more number of faces is detected falsely as compared to that of multi-resolution shown in Figure 4.

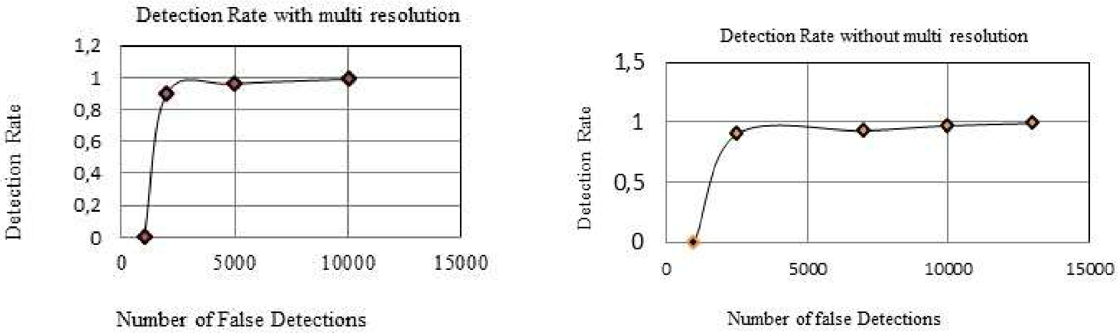


Figure 4. Detection rate with multi-resolution in 24-net CNN

The overall test sequence is shown in Figure 5. First of all test image is applied to the system, a 12 net structure will scan the whole image and quickly rejects about 90% of detection windows. Remaining detected window will be processed by 24 calibrated CNNs. In next subsequent stages, the highly overlapped window will be removed. Then a 48 net will take detected windows and evaluate the window with calibrated boundary box and produces as output as detected boundary box. Figure 6 shows all detection stages with different structure stages.

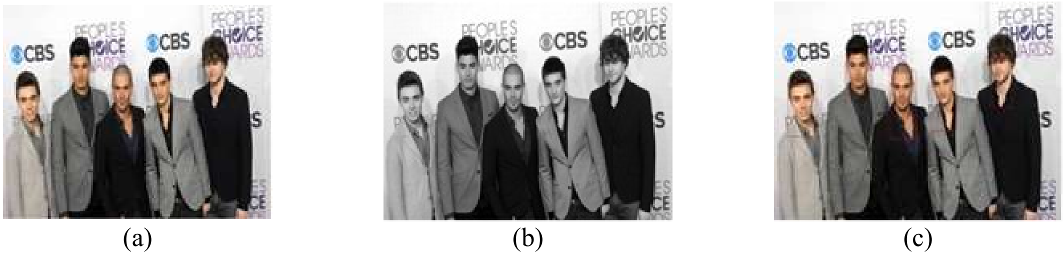


Figure 5. Detection results: (a) Original image given for detection, (b) Image at preprocessing stage (c) Detected face position with CNN

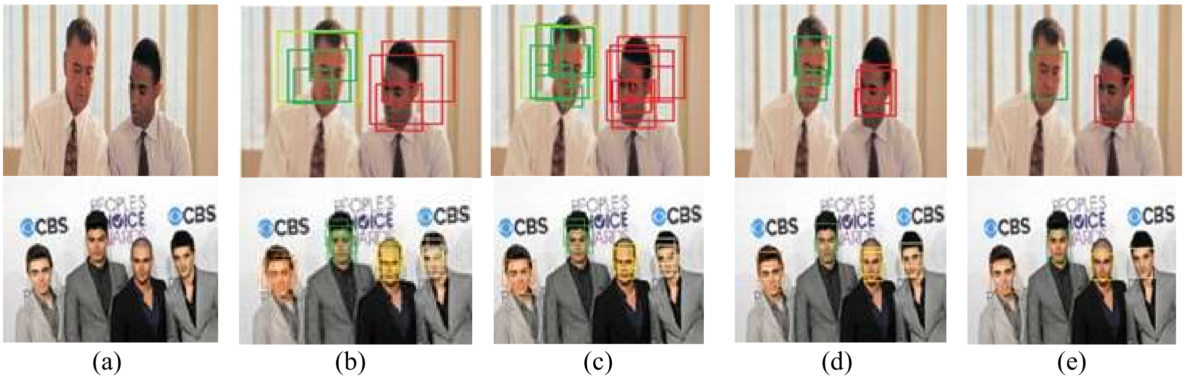


Figure 6. Detection results for different CNN structure: (a) Input/test image, (b) Image after 12-net CNN, (c) Image after 24-net CNN, (d) Image after 48-net CNN, (e) Output face detected image

4. CONCLUSION

In this work, we develop an algorithm for detecting multi-view faces using deep convolution neural network. A major contributions were made in this particular research is that we have developed a procedure which can assemble a wide range of dataset, with the small noise of label while reducing the quantity of manual annotation included. The main concept of the algorithm is to influence the high ability of DCNN to classify and extract the feature. To learn the single classifier for detecting faces from different views and reduce the computational difficulty to simplify the detector architecture. For this work, we first transformed the completely linked layers into the convolutional one to reshape the parameters of the layer. By exploring a few key features of the network structure, we achieve high performance convolutional networks with a relatively small scale. Our detector gives results for images with varying poses and resolutions.

REFERENCE

[1] Yang, Ming-Hsuan, David J. Kriegman, and Narendra Ahuja, “Detecting Faces in Images: A Survey,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24.1 (2002): 34-58.

[2] Hjelmås, Erik, and Boon Kee Low, “Face Detection: A Survey. Computer Vision and Image Understanding,” 83.3 (2001): 236-274.

[3] Sheikh Amanur Rahman M.A. Ansari and Santosh Kumar Upadhyay, “An Efficient Architecture for Face Detection in Complex Images,” *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, Issue 12, 2012.

[4] Li, Wei. "An Adaptive Detection Method of Multiple Faces." *Indonesian Journal of Electrical Engineering and Computer Science* 12.4 (2014): 2743-2752.

[5] Lin, Chuan, et al., "Face Detection Algorithm Based On Multi-Orientation Gabor Filters and Feature Fusion," *Indonesian Journal of Electrical Engineering and Computer Science* 11.10 (2013): 5986-5994.

[6] M.-H. Yang, D. Kriegman, and N. Ahuja, “Detecting Faces in Images: A Survey,” *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, Jan 2002.

[7] H. A. Rowley, S. Baluja, and T. Kanade, “Neural Network-Based Face Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1098

[8] K.-K. Sung and T. Poggio, “Example-Based Learning For View-Based Human Face Detection,” *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39–51, 1998

[9] Y. M. Li, S. G. *Gong, and H. Liddell, "*Support Vector Regression and Classification Based Multi-View Face Detection And Recognition," Proc Intl Conf Automatic Face and Gesture Recognition, FG, 2000.

[10] H. Schneiderman and T. Kanade, "A Statistical Method for 3D Object Detection Applied to Faces and Cars," *Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, 2000

[11] Jones, M.; Viola, P., "Fast Multi-view Face Detection", MITSUBISHI ELECTRIC RESEARCH LABORATORIES, TR2003-96 August 2003

[12] P.Viola and M. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, 57(2) 2004.

[13] Masi, Iacopo, et al. "Learning Pose-Aware Models for Pose-Invariant Face Recognition in the Wild." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[14] 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, Oct 2016.

[15] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A Convolutional Neural Network Cascade For Face Detection," *In Proc Intl Conf on Computer Vision and Pattern Recognition*, CVPR, Jun 2015, pp. 5325–5334.

[16] Cha Zhang and Zhengyou Zhang, "Improving Multiview Face Detection With Multi-Task Deep Convolutional Neural Network," *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM*, 2015.

[17] Farfade, Sachin Sudhakar, Mohammad J. Saberian, and Li-Jia Li, "Multi-View Face Detection Using Deep Convolutional Neural Networks," *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval.* ACM, 2015.

[18] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman, "*Deep face recognition. British Machine Vision Conference*. Vol. 1. No. 3. 2015.

[19] Li, Haoxiang, et al., "A Convolutional Neural Network Cascade for Face Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[20] Cha Zhang and Zhengyou Zhang, "Improving Multiview face detection with Multi-Task deep Convolutional Neural Network," *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015.

[21] Vasanth, P. C., and K. R. Nataraj, "Facial Expression Recognition Using SVM Classifier," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)* 3.1 (2015): 16-20.

[22] Hadsell, Raia, Sumit Chopra, and Yann LeCun, "Dimensionality Reduction by Learning an Invariant Mapping." In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2, pp. 1735-1742. IEEE, 2006.

[23] P. Doll´ar and C. L. Zitnick, "Fast Edge Detection Using Structured Forests," *IEEE Trans. Pattern Anal. Mach. Intell.*,37(8):1558–1570, 2015.

[24] S. Yang, P. Luo, C. C. Loy, and X, "Tang. From Facial Parts Responses To Face Detection: A Deep Learning Approach. In ICCV, pages 3676–3684, 2015.

[25] Jain, Vidit, and Erik Learned-Miller, "Fddb: A Benchmark for Face Detection In Unconstrained Settings," Vol. 88. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

BIOGRAPHIES OF AUTHORS

Shivkaran Ravidas has received B.E.in Electronics Engineering from Dr. Baba Saheb Ambedkar Marathwada University, Aurangabad, India, in 2002 and M.E .in Electronics Engineering from SGGS College of Engg. and Technology, Nanded, India, in 2006. Currently he is working as research scholar in the Department of Electrical Engineering, Gautam Buddha University, Greater Noida, India. He has published several papers in reputed journals & conferences. His main areas of research interest are image processing, machine learning and Multi-view face Detection.

M.A.Ansari received B.Tech. (Electrical Engineering) in 1998 from AMU, Aligarh, India, M.Tech. & Ph.D. (Electrical Engineering) in 2001 & 2009 respectively from Indian Institute of Technology, Roorkee, India. The author is associated with Gautam Buddha University and is currently working in the Department of Electrical Engineering. He has wide national and international experience of teaching and has visited several countries.He has published several papers in reputed national and international journals and conferences. His research interest includes medical image& signal processing, Biomedical Instrumentation, Softcomputingand wavelet applications. He is senior member of IEEE and ISIAM.