Face Recognition Using CNN Trained With Histogram Equalization Based Image Enhancement Scheme

Seshaiah Merikapudi[1] ,Dr. Shrishail Math[2], Dr. C.Nandini[3], Dr.Mahammed Rafi[4]

[1]Research Scholar, [2]Professor, [3]Vice Principal &HOD, 4Professor
[1] VTU RRC,Dept. of CSE, SJCIT,Chickballapur-562101, India.
[2]Professor,Dept. of CSE, Sri Krishna Institute of Technology,Bangalore-560090, India.
[3]Vice Principal &HOD,Dept. of CSE, DSATM,Bangalore-560082, India.
4Professor,Dept. of CSE, UBDTE,Davangere-577004, India.

[1] merikapudi@gmail.com,

Abstract
Face recognition is considered as a promising solution for video surveillance systems. Currently, the still image-based face recognition techniques have obtained promising accuracy but detection and recognition of faces in real-time videos has becomes a challenging task. Moreover, the demand of CCTV (Closed-Circuit Television) based surveillance has increased rapidly where the quality of videos is very low. Thus, the poor-quality video, occlusion, and other conditions creates various complexities in face recognition. Currently, CNN (Convolutional Neural Network) based techniques have gained attraction from research community because these techniques have good learning capability and provide better accuracy. In this work, we have introduced CNN (Convolutional Neural Network) based scheme which uses feature extraction and feature embedding modules along with Google Net architecture to improve the learning of CNN (Convolutional Neural Network). We have incorporated histogram equalization-based image enhancement approach to improvise the quality of video frames. The proposed approach is implemented using Python 3.7. The experimental analysis shows that proposed approach achieves the accuracy as 98.55% and AUC(Area under the Curve) as 99.10% for open source datasets whereas for real-time scenarios without occlusion it achieves accuracy as 99.12%, for occlusion scenario it achieves 98.87% classification accuracy.

Keywords: Face recognition, CNN, histogram equalization, CCTV, surveillance

1. INTRODUCTION

Recently, use of CCTV (Closed-Circuit Television) cameras has increased dramatically in various real-time surveillance systems. CCTV surveillance gained attraction after 2005 London bombings (caused around 56 deaths and 784 injuries) and New York terrorist attacks in 2001 (caused around 3000 deaths and 25000 injuries). A study presented in [1] revealed that the Metropolitan Police in London has installed more that million cameras for surveillance and to identify the criminal activities. To improve the performance of these surveillance systems, continuous monitoring is required which becomes a tedious task and also it increases the cost of surveillance systems. Moreover, human monitoring has several restrictions such as human can monitor limited number of videos whereas number of captured videos are more thus more persons are required to monitor the large number of videos which is infeasible due to increased cost. Another issue in this field is reliability of monitoring. To overcome these issues, automated analysis of these cameras is highly recommended.

In surveillance system, human face detection and recognition plays important role to identify the suspicious activities and criminals. Video-based surveillance systems suffer from various challenging factors such as motion, nature of subject. Lighting conditions and poor quality of CCTV video data. Moreover, low resolution, low contrast, blurriness, and pose variation creates additional complexities for face recognition. Conventional face recognition systems are based on pattern recognition where training and testing sets are created. These training and testing images are captured from various users and have varied pose, illumination, expression, and resolution. Figure 1 shows a machine learning based model to recognize the face where image pre-processing, extracting the features, and feature classification are the main modules. [19]. Generally, existing methods work on the feature extraction and dictionary learning such as small-scale illumination invariant features [2], Gabor-features [3], 3D-DWT [4] and many more [5]. Similarly, dictionary-based learning also has achieved promising performance because these techniques are suitable for encoding of the feature of posture variation, brightness change and varied expression in learned dictionaries [7-8]. However, these methods are unsupervised and achieve poor performance for different face sets. Moreover, these models utilizes original pixel values that might hold noisy information which is unnecessary for the dictionary learning. Several methods were presented in the past to handle these issues and to improvise accuracy of face recognition system. Similarly, video-based recognition is considered quite challenging job in the field of real-time surveillance [9]. Video based recognition system captures person’s face from different angles and provides various vital data of the individual face. However, techniques of video-based face detection and recognition suffer from uncontrolled pose and illumination variations which leads to increase the misclassification error and intra-class distance. To overcome these performance related issues, fixed size feature representation for the entire video is a possible solution which would perform distance and similarity matching without considering each frame for evaluation. Feature extraction and performing feature pooling is the solution to deal with these issues. Recently, deep learning-based techniques have gained huge attraction in this field of video face detection and recognition due to feature pooling and robust feature learning [10]. Deep learning approaches like Deep Neural Network (DNN), Convolution Neural Network (CNN)and Recurrent Neural network (RNN) [11-12], use cascaded multiple layers to extract the features and transformation. These layers help to learn the features for different levels and formulate hierarchal structure to demonstrate the relationship for varied pose, lighting and expression conditions. In many face recognition techniques, the multiple face images are presented as set of face descriptors which are extracted using deep neural network and matching features are fused. However, this technique consumes more time and memory hence cannot be implemented for large-scale recognition tasks. Feature aggregation is a promising solution for these issues. The max-pooling and average pooling techniques are the widely adapted for most of the aggregation techniques. In [13] Liu et al. reported that incorporating feature aggregation strategy can help to improve the performance of video face recognition. Unlike, the conventional method, this work also considers low-quality frames and creates feature aggregation network using feature embedding and aggregation module. Gong et al. [14] presented focused on feature aggregation and introduced component-wise feature aggregation model for face recognition from videos. However, the first phase of this work is to train the base CNN for still images and later, in the next phase, aggregation model is adopted to aggregate the deep features in a single feature vector.

However, the increasing demand of face recognition system urge for higher accuracy which makes it more challenging task. Moreover, the similarity of the same person may vary due to image capturing variations such as lighting, view point, head pose, occlusion and facial expression. Current face recognition systems have reported promising performance of classification for the facial images which are captured in the controlled conditions. Similarly, the video face recognition suffers from various challenges. Compared to other video data, the CCTV video data create more difficulty for face recognition due to poor quality of videos. Due to significant advantages of deep learning-based schemes, we adopt deep learning-based scheme for video face identification. The proposed scheme uses feature aggregation and a novel architecture is introduced. The major objectives accomplished in this researchare defined as:

- Incorporating image enhancement technique
- Development of feature extraction and embedding module
- Development of attention block and aggregation technique to generate the significant features and their weights
- CNN training and testing for video face recognition.

Remaining article is arranged in following order: section II presents the review of literature of recent techniques of face recognition and presents their drawbacks and challenges. Section III presents a proposed solution for these techniques using deep learning method, section IV presents a comparative experimental analysis where we make the comparison of proposed methods performance against existing techniques. Finally, section V summarizes the work and future scope in this study.

2. Literature Survey

In this section, we present a brief literature review about the current techniques of video face recognition which includes feature based and deep learning-based techniques along with CCTV video data. Several techniques have been introduced in the past thattake the video as an array of multiple frames and apply image matching for face recognition such as Yang et al. presented neural aggregation framework for processing the image set for face recognition. According to Rao et al. [15], video face identificationtechniques can be classified in2types as instance-based and manifold techniques. According to the first process, the video frames are showed as set of multiple instances to obtain the associationamongst different frames. Similarly, for second category, each set of images is exhibited as manifold,also the resemblance distance among each manifold of image is measured by computing the distance between obtained sets. Wang et al. [16] discussed that existing techniques use Covariance Matrix (CM), Linear Subspace (LS), or Gaussian Distribution (GD) to model the original image manifold. These techniques adopt single geometric model to describe the given manifold. However, individual geometric model may missvaluabledata which can affect the classification performance. To overcome these issue, multi-geometry model is developed where CM, LS, and GD are used. Next phase performs Riemannian manifold valued feature extraction which are later mapped as amulti-dimensional spaces. Finally, a multi-kernel learning model is presented to set the kernels for classification. In [15] authors presented a technique which can use information from both side such as manifold and instances. Similarly, RNN withLSTM (Long Short-Term Memory) networks are widely adopted but these systems suffer from the overfitting issue. Hence, to overcome the issues of these networks, authors introduced Recurrent Embedding Aggregation Network (REAN) for face identification in surveillance videos. This method overcomes the overfitting issue by aggregating the pre-trained embedding’s rather than learning the parameters from initial step. Liu et al. [18] presented image set based verification approach for face verification system. In these schemes, establishing the relationship between images which are not arranged in a proper order, is considered as a challenging task. This problem is solved by using Markov Decision Process (MDP). This approach develops DAC (Dependency-aware Attention Control) network thatutilizes actor-critic reinforcement learning for sequential attention findings to achieve correlation between unordered images. The low-resolution images create additional complexities in the face recognition systems. Zangeneh et al. [20] proposedDeep CNN (DCNNs) model for face identification from low-resolution images. This approach combined poor andHD images in shared space. The HD image section network consists of 14 layers and low-resolution network consist of 5 layers. This network is formulated by connecting these layers. The HD and low-resolution image features are back propagatedfor training the network. Li et al. [21] used recurrent regression NN for cross-pose face identification in image and video datasets. For learning purpose, a potential dependencies network is constructed to regularize the final learning model. For still image scenario, sequential poses are processed to obtain the useful information, similarly, for videos, it considers entire sequence as input to the regression network. Chen et al. [22] presented unconstrained face identification system that includes detection of face, orientation, and face verification. This model uses DCNN model to train the data. Lin et al. [23] focused on facial landmark-based feature extraction. Authors suggested that these features provide significant information of human face for face recognition systems. However, in video frame-based feature extraction, these landmarks are completely redundant which need to be aligned. Hence, authors present a model to integrate the feature data to the shared coordinate frame, finally, AdaBoost function is applied for classification. In [24], Masi et al focused on the pose variance issue and presented a pose-invariant method for face recognition. According to this study, the existing techniques use a single method to learn all the parameters of huge data of different pose. These methods align all poses in a single frontal pose. However, these methods suffer from accuracy and complexity issues, hence authors in [24] introduced Pose-aware Models using deep convolution neural networks. Further, a 3D rendering model is also applied which helps to combine the multiple face poses to improve the robustness to pose variations. Zhang et al. [25] introduced a Thermal-to-Visible Generative Adversarial Network (TV-GAN) which transforms the thermal image of the face to Visible Light Domain (VLD).

3. Proposed Model

This section presents the proposed solution for video face recognition for low-illumination and poor quality of data using deep learning techniques. Generally, the quality of video frames is much lower than the still images which are captured under constrained conditions. Currently, several image enhancement techniques are present such as histogram equalization, wiener filter, linear contrast adjustment, unsharp mask filtering and Contrast-limited adaptive histogram equalization (CLAHE). Moreover, the video frames may suffer from various issues such as motion blur, out-of-focus blur, jitter, occlusions, etc. In order to identify the face from video frames, features aggregation of obtained features from extracted frames is the simplest solution to generate the face representation for the simple template. The face embedding’s are obtained by using a facial representation model which is expressed as:

$$p(\boldsymbol{f}|\ \mathbb{I}^*, \mathbb{F}^*) = \int p(\boldsymbol{f}|\boldsymbol{i}, \mathbb{I}^*, \mathbb{F}^*)p(\boldsymbol{f}|\boldsymbol{i}, \mathbb{I}^*, \mathbb{F}^*)\boldsymbol{di} \quad (1)$$

Here,$\mathbb{I}^* = \{\boldsymbol{i_1}, \boldsymbol{i_2}, \dots \boldsymbol{i}_{\mathbb{M}}\}$ denotes the set of training images, $\mathbb{F}^* = \{\boldsymbol{f_1}, \boldsymbol{f_2}, \dots \boldsymbol{f_M}\}$denotes the features extracted from the training data. These features are also known as noisy embedding’s of training data. $p(\boldsymbol{f}|\boldsymbol{i}, \mathbb{I}^*, \mathbb{F}^*)$ represents vagueness of feature embedding estimate and $p(\boldsymbol{f}|\ \mathbb{I}^*, \mathbb{F}^*)$ denotes the probability density of face image in the clean embedding. Here, we use a deterministic function $\boldsymbol{\varphi}$ which helps to map the face images to the corresponding embedding. Let us consider a temple $T = \{i_1, i_2, \dots, i_N\}$ which has N number of images in the template of one identity. The count of N can be high in videos where clean embedding of attributes are estimated using expectation function$\hat{E}(\boldsymbol{F}^T)$:

$$\boldsymbol{\varphi} \approx \hat{E}(\boldsymbol{F}^T) = \sum_{i=1}^{N} p(\boldsymbol{f}|\mathbb{I}^*, \mathbb{F}^*)\boldsymbol{f}_i \quad (2)$$

Here,$\boldsymbol{F}^T$ denotes the set of noisy attributes of the template$T$. In order to estimate the probabilistic density of feature embeddings, we linearly combine adaptive scalar weights and approximated template embedding based on the weights as:

$$\boldsymbol{r}^T = \sum_{i=1}^{N} g(\boldsymbol{f_i})\boldsymbol{f}_i \quad (3)$$

Here,r^T represents the templates and $g(\boldsymbol{f_i})$ denotes the predicted weights for attributes of i^{th} image in the considered template$T$. This method helps to decrease the noise in attributes. However, the CCTV videos have low-quality frames and the quality of frames is highly unbalanced hence the recognition performance degrades. To overcome these issues, we present a novel feature aggregation model. In this work, we adopt the deep convolution neural network based embedding module where each frame of video is embedded for feature representation. Further, we implement the $GoogleNet$ with BN (Batch Normalization) technique to improve the performance of CNN [26]. The $GoogleNet$ produces multi-dimensional image features which are firstly normalized in the form of unit vector then processed through the aggregation module. Let us consider that we have n number of faces from the video data as $(\mathcal{X}^i, y_i)_{i=1}^n$ where $\mathcal{X}^i$ denotes the video sequence with varying number of images K_i as $\mathcal{X}^i = \{x_1^i, x_2^i, \dots x_{K_i}^i\}$ where x_k^i denotes the k^{th} frame in the current video and y_i denotes the label. Prior to this, we extract the feature for each frame x_i^k from feature embedding module denoted as $\boldsymbol{f}_k^i$. Here, our main aim is to produce a set of linear weights $\{a_k\}_{k=1}^K$ for each video using features which are extracted from feature embedding module. Thus, the aggregated features can be denoted as:

$$r = \sum_k a_k \boldsymbol{f}_k \tag{4}$$

This helps to generate a feature vector which consists of similar equivalentsize as an individual image of face. The proposed work considers three main concepts to design the feature aggregation module. First of all, the model should process different number of images to ensure that it can handle varied data in video files. Next, the aggregation model process should not get affected due to the change in the image sequence. This helps to generate the robust feature aggregation module.

We present the modeling of attention block which is capable to process the features obtained from feature embedding module and generates the linear weights. Let $\{\boldsymbol{f}_k\}$ denotes the face feature vector which need to be processed through attention block. The attention block applies feature filtering with kernel $\boldsymbol{q}$with the help of dot product. This generates a set of significance values $\{s_k\}$ for each feature. Later, these significance features are processed through the softmax operator which generates the positive weights. These operators can be expressed as:

$$s_k = \boldsymbol{q}^T \boldsymbol{f}_k$$

$$a_k = \frac{\exp(s_k)}{\sum_j \exp(s_j)} \tag{5}$$

Further, we improve the aggregation module by cascading two attention blocks. Let us consider that $\boldsymbol{q}^1$ denotes the kernel of 1st attention block,$\boldsymbol{r}^0$ denotes the accumulated features. Here, we calculate the $\boldsymbol{q}^2$ kernel of 2ndattention block using transfer layer using aggregated features$\boldsymbol{r}^0$. This can be expressed as:

$$\boldsymbol{q}^2 = \tanh(\boldsymbol{r}^{\mathbf{0}} \boldsymbol{W} + \boldsymbol{b}) \tag{6}$$

Where $\boldsymbol{W}$signifies the weight matrix and $\mathbf{b}$denotes the bias array of neurons and $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ applies a hyperbolic tangent nonlinearity. Based on these functions, we share the weights and measure the loss. The main aim of training the network is to minimize the loss as:

$$l_{i,j} = y_{i,j} \left\| \boldsymbol{r}_i^1 - \boldsymbol{r}_j^1 \right\|_2^2 + (1 - y_{i,j}) \max\left(0, m - \left\| \boldsymbol{r}_i^1 - \boldsymbol{r}_j^1 \right\|_2^2\right) \tag{7}$$

Where $y_{i,j} = 1$ and m is a constant set as 2. During training phase, we can train these models simultaneously in an end-to-end fashion and one-by-one. In this work, we select, one-by-one training and trained the network on single images, later, we trained the aggregation module. This network is trained using 3M face images 50K identities obtained from internet. Face detection and alignment is also performed prior to feature extraction. The input image size is considered as 224x224.

4. Results and Discussions

This part of the article presents the experimental evaluation of proposed scheme of face detection and identification in still images, videos and real-time videos. The projected approach is implemented on Python3.7 running on windows platform with NVIDIA GPU. For analysis, we consider YouTube face data and IJB-A data. A short narrative of these datasets is presented as:

- YouTube Face Dataset: initially, this dataset was released in 2011[27]. In this data total 3425 number of videos are present which are recorded from 1595 different users. Theframe count in the YTF face videos variesin between 48 to 6, 070. Also, on an average it has 181 frames. In contrast to IJB-S, the YTF's has greater photojournalistic media [12]. In trials, we have followed the 1:1 face verification practicethat has5,000 pair of videos.
- IJB-A: IJB-A [28] is basically a template basedunrestricted face identificationstandard. Though,the images existent in it has alike challenges as it is with IJB-S, the templates exist in IJB-A consist of images from assortedsource of medias, havingbetter resolution of imagescompared to IJB-S. The standardoffers template-based 1:1 verification and 1: N identification procedures. IJB-A comprisesof 500 subjects having overall images size as 25,813. Also, it is extensivelybeing utilized by a no. of both still image and video-based face identification algorithms

Results for IJB-A dataset:

Here, we demonstrate the face detection and recognition analysis for IJB-A dataset which contain videos and face images from different environment. This data has multiple variations and conditions thus it becomes a challenging task of face recognition. Below given table 1 shows the comparative analysis in terms of the true accept rates (TAR) vs. false positive rates (FAR) where we compared the efficiency of proposed approach with existing techniques.

The above given Table 1 and Bar Graph 1shows the proposed approach achieves better performance when compared with existing techniques. We have adopted some comparative techniques from Yang et al. [9] where experimental study is extended by incorporating L2 distance measurements with CNN architecture such as CNN + Max L2, CNN + Min L2, $CNN + Mean\,L2$, and $CNN + SoftMin\,L2$ along with max and average pooling such as $CNN + MaxPool$, and $CNN + AvePool$. These techniques also achieve better performance as 0.978±0.004 but proposed aggregation module helps to reduce the noisy features resulting in improving the accuracy of the system.

YouTube Face dataset:

Here, we present the experimental assessment for YouTube face database. This data contains 3425 number of videos which are of 1595 different peoples. The number of frames in these vides varies from 48 to 6070 frames. When it comes to a classification problem, we can count on an AUC Curve. When we need to check or visualize the performance of the multi - class classification problem, we use AUC. It is one of the most important evaluation metrics for checking any classification model's performance.

Prior to processing the video faces for recognition, we detect the faces, extract the features and align these features to generate the feature vector. Table 2 shows a comparative performance for video face recognition in the parameters of recognition accuracy and area under curve. We also consider base line methods such as CNN + Max L2, CNN + Min L2, CNN + Mean L2, CNN + Soft Min L2, CNN + Max Pool and CNN +Avg Pool adopted from [9]. The comparative analysis shows that proposed approach achieves accuracy of 98.23% which demonstrates a significant improvisation in contrast to existing standard methods.

Real-time experiment :

This section presents the experimental analysis for real-time dataset where we have setup a webcam as CCTV camera to capture the videos and recognize the person from live video streams. We have captured training images from 3 users as depicted in below given figure. To show the efficiency of the proposed scheme, we include different occlusions as presented in figure 2(Case 1), 2 (Case 2) and 2(Case 3) which contains partial occlusions such as covering the lips, and one eye. For each scenario, we present an experimental study, detection and recognition outcome is presented in figure 2(Case 4).

The above given experiment shows a significant performance for real-time face recognition where we have tested different cases. Without any occlusion the proposed model achieves reliable accuracy of 99.12%, for occlusion scenario it achieves 98.87% classification accuracy.

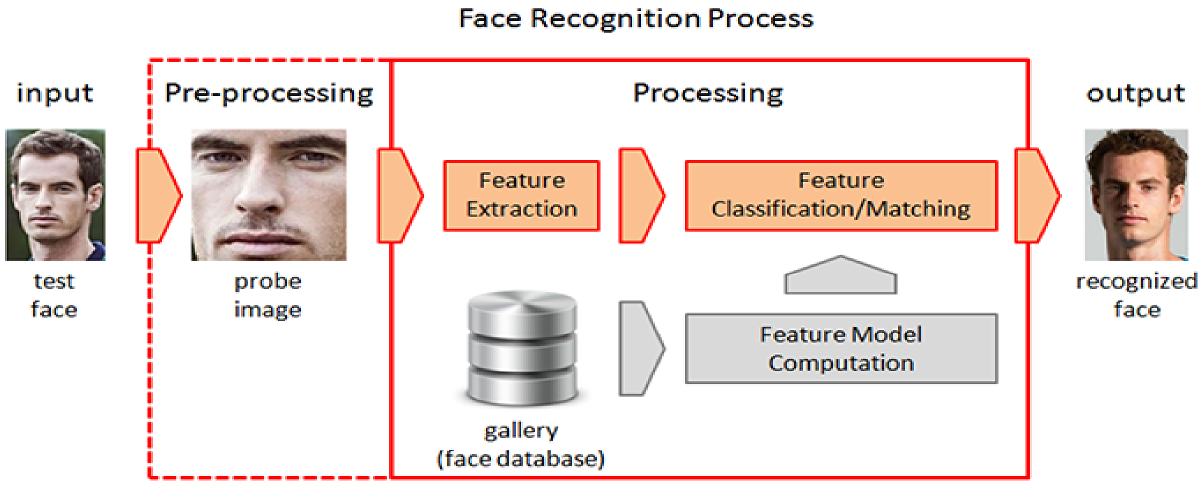
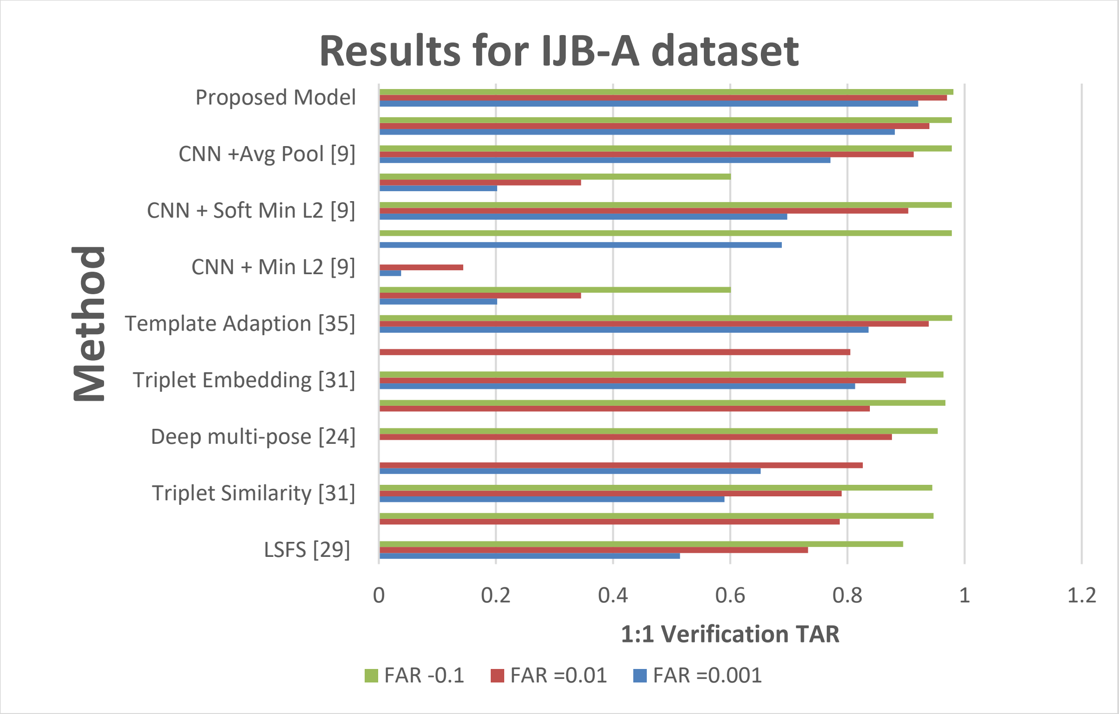


Figure 1. Automated face recognition system



Bar

Figure 2: comparative analysis in terms of the true accept rates (TAR) vs. false positive rates (FAR)

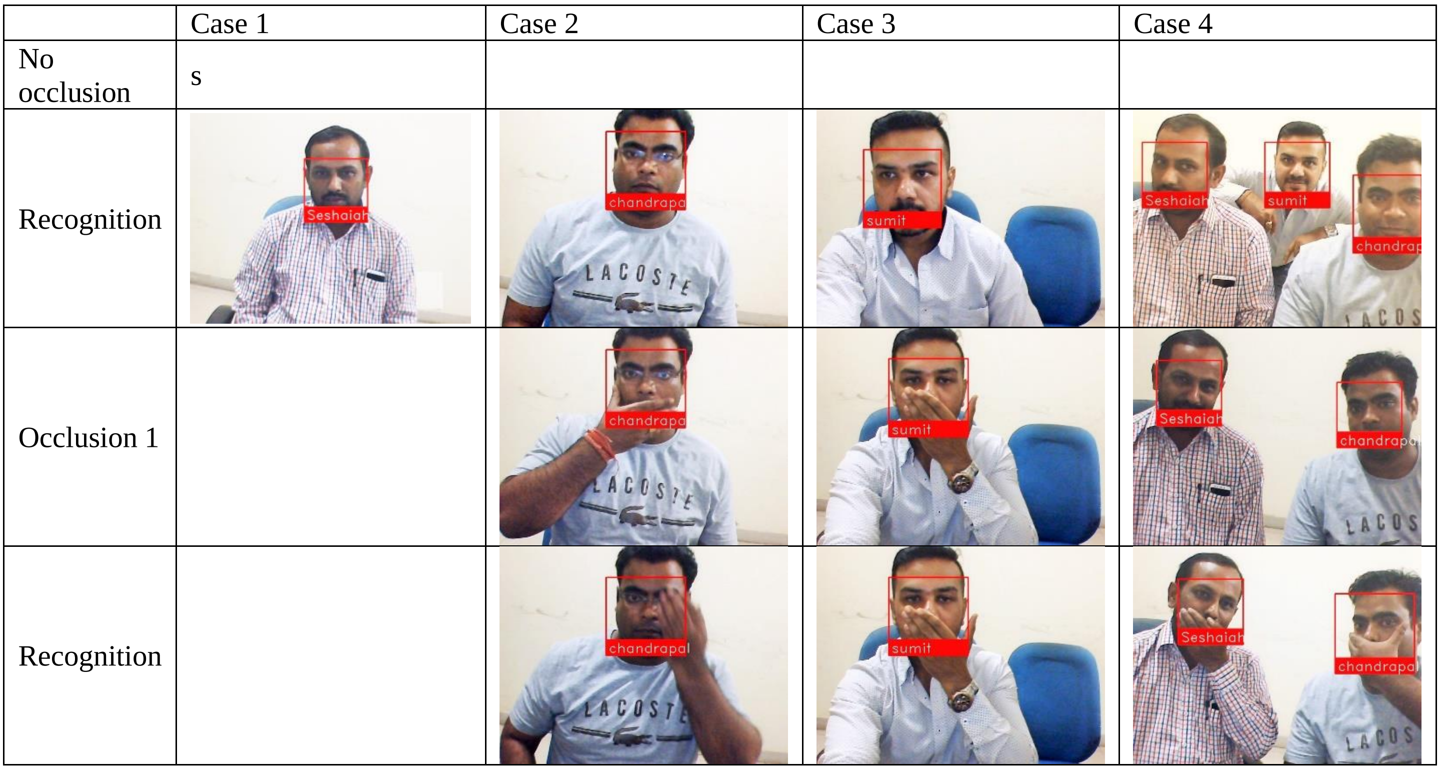


Figure 2: Experimental analysis for real-time dataset

Table 1:comparative analysis in terms of the true accept rates (TAR) vs. false positive rates (FAR)

	1:1 Verification TAR		
Method	**FAR =0.001**	**FAR =0.01**	**FAR -0.1**
LSFS [29]	0.514±0.060	0.733±0.034	0.895±0.013
DCNN [30]	-	0.787±0.043	0.947±0.011
Triplet Similarity [31]	0.590±0.050	0.790±0.030	0.945±0.002
Pose-aware models [32]	0.652±0.037	0.826±0.018	-
Deep multi-pose [24]	-	0.876	0.954
DCNN Fusion [33]	-	0.838±0.042	0.967±0.009
Triplet Embedding [31]	0.813±0.02	0.90±0.01	0.964±0.005
VGG-Face [34]	-	0.805±0.030	-
Template Adaption [35]	0.836±0.020	0.939±0.013	0.979±0.004
CNN + Max L2 [9]	0.202±0.029	0.345±0.025	0.601±0.024
CNN + Min L2 [9]	0.038±0.008	0.144±0.073	0.972±0.006
CNN + Mean L2 [9]	0.688±0.080	0.895±0.016	0.978±0.004

CNN + Soft Min L2 [9]	0.697±0.085	0.904±0.015	0.978±0.004
CNN + Max Pool [9]	0.202±0.029	0.345±0.025	0.601±0.024
CNN +Avg Pool [9]	0.771±0.064	0.913±0.014	0.978±0.004
NAN [9]	0.881±0.011	0.94±0.008	0.978±0.003
Proposed Model	0.921±0.012	0.97±0.011	0.981±0.001

Table 2:Experimental assessment for YouTube face database

Method	Accuracy (%)	AUC
LM3L [37]	81.3±1.2	89.3
DDML [38]	82.3±1.5	90.1
EigenPEP [39]	84.8±1.4	92.6
DeepFace-single [40]	91.4±1.1	96.3
FaceNet [36]	95.12±0.39	-
VGG-Face [34]	97.3	-
CNN + Max L2 [9]	91.96±1.1	-
CNN + Min L2 [9]	94.96±0.79	98.5
CNN + Mean L2 [9]	95.30±0.74	98.7
CNN + Soft Min L2 [9]	95.30±0.77	98.7
CNN + Max Pool [9]	88.36±1.4	95
CNN +Avg Pool [9]	95.20±0.76	98.7
NAN [9]	95.72±0.64	98.8
Proposed Model [9]	98.55±0.10	99.10

5. CONCLUSION

Face detection and recognition plays a significant role in the field of visual surveillance systems. In this work, we present a novel solution for detection and recognition of face for the real-time videos which are captured using CCTV surveillance cameras. According to this work, we extract the features in the form of feature embedding’s and feature weights are generated. Further, we adopt the *GoogleNet* architecture to improve the CNN. Later, feature attention and aggreation models are incorprated and significant features are processed through softmax operator. Finally, the network is trained in one-by-one training method. The comparative experimental analysis shows that proposed approach achieves better accuracy when compared with existing techniques.

Acknowledgments:

The authors gratefully acknowledge-edge support from the Northeast Forestry University (NEFU) as well as the Heilongjiang Institute of Tech-nology (HLJIT). This work is supported by the Spe-cial Scientific Research Funds for Forest Non-profit Industry (Grant No. 201504508) and the Youth Science

REFERENCES

[1] Davis, M., Popov, S., &Surlea, C. (2011). Real-time face recognition from surveillance video. In Intelligent Video Event Analysis and Understanding (pp. 155-194). Springer, Berlin, Heidelberg.

[2] Yu, Y. F., Dai, D. Q., Ren, C. X., & Huang, K. K. (2017). Discriminative multi-layer illumination-robust feature extraction for face recognition. Pattern Recognition, 67, 201-212.

[3] Li, L., Ge, H., Tong, Y., & Zhang, Y. (2018). Face recognition using Gabor-based feature extraction and feature space transformation fusion method for single image per person problem. Neural Processing Letters, 47(3), 1197-1217.

[4] Patil, C. M., &Ruikar, S. D. (2020). 3D-DWT and CNN Based Face Recognition with Feature Extraction Using Depth Information and Contour Map. In Techno-Societal 2018 (pp. 13-23). Springer, Cham.

[5] Makhija, Y., & Sharma, R. S. (2019). Face recognition: Novel comparison of various feature extraction techniques. In Harmony Search and Nature Inspired Optimization Algorithms (pp. 1189-1198). Springer, Singapore.

[6] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[7] Lu, J., Wang, G., & Zhou, J. (2017). Simultaneous feature and dictionary learning for image set based face recognition. IEEE Transactions on Image Processing, 26(8), 4042-4054.

[8] Xu, Y., Li, Z., Yang, J., & Zhang, D. (2017). A survey of dictionary learning algorithms for face recognition. IEEE access, 5, 8502-8514.

[9] Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., & Hua, G. (2017). Neural aggregation network for video face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4362-4371).

[10] Masi, I., Wu, Y., Hassner, T., & Natarajan, P. (2018, October). Deep face recognition: A survey. In 2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI) (pp. 471-478). IEEE.

[11] Wang, Z., He, K., Fu, Y., Feng, R., Jiang, Y. G., &Xue, X. (2017, June). Multi-task deep neural network for joint face recognition and facial attribute prediction. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (pp. 365-374).

[12] Phillips, P. J. (2017, May). A cross benchmark assessment of a deep convolutional neural network for face recognition. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 705-710). IEEE.

[13] Liu, Z., Hu, H., Bai, J., Li, S., &Lian, S. (2019). Feature Aggregation Network for Video Face Recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops.

[14] Gong, S., Shi, Y., & Jain, A. K. (2019). Video face recognition: Component-wise feature aggregation network (c-fan). arXiv preprint arXiv:1902.07327.

[15] Rao, Y., Lu, J., & Zhou, J. (2017). Attention-aware deep reinforcement learning for video face recognition. In Proceedings of the IEEE international conference on computer vision (pp. 3931-3940).

[16] Wang, R., Wu, X., & Kittler, J. (2019). Multiple Riemannian Manifold-valued Descriptors based Image Set Classification with Multi-Kernel Metric Learning. arXiv preprint arXiv:1908.01950.

[17] Gong, S., Shi, Y., Jain, A. K., & Kalka, N. D. (2019). Recurrent embedding aggregation network for video face recognition. arXiv preprint arXiv:1904.12019.

[18] Liu, X., Guo, Z., You, J., & Kumar, B. V. (2019). Dependency-Aware Attention Control for Image Set-Based Face Recognition. IEEE Transactions on Information Forensics and Security, 15, 1501-1512.

[19] Olszewska, J. I. (2016). Automated face recognition: challenges and solutions. Pattern Recognition-Analysis and Applications, 59-79.

[20] Zangeneh, E., Rahmati, M., &Mohsenzadeh, Y. (2020). Low resolution face recognition using a two-branch deep convolutional neural network architecture. Expert Systems with Applications, 139, 112854.

[21] Li, Y., Zheng, W., Cui, Z., & Zhang, T. (2018). Face recognition based on recurrent regression neural network. Neurocomputing, 297, 50-58.

[22] Chen, J. C., Ranjan, R., Sankaranarayanan, S., Kumar, A., Chen, C. H., Patel, V. M., ...&Chellappa, R. (2018). Unconstrained still/video-based face verification with deep convolutional neural networks. International Journal of Computer Vision, 126(2-4), 272-291.

[23] Lin, J., Xiao, L., & Wu, T. (2018). Face recognition for video surveillance with aligned facial landmarks learning. Technology and Health Care, 26(S1), 169-178.

[24] Masi, I., Chang, F. J., Choi, J., Harel, S., Kim, J., Kim, K., ...&AbdAlmageed, W. (2018). Learning pose-aware models for pose-invariant face recognition in the wild. IEEE transactions on pattern analysis and machine intelligence, 41(2), 379-393.

[25] Zhang, T., Wiliem, A., Yang, S., & Lovell, B. (2018, February). Tv-gan: Generative adversarial network based thermal to visible face recognition. In 2018 international conference on biometrics (ICB) (pp. 174-181). IEEE.

[26] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ...&Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[27] Wolf, L., Hassner, T., &Maoz, I. (2011, June). Face recognition in unconstrained videos with matched background similarity. In CVPR 2011 (pp. 529-534). IEEE.

[28] Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., ...& Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: Iarpajanus benchmark a. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1931-1939).

[29] Wang, D., Otto, C., & Jain, A. K. (2015). Face search at scale: 80 million gallery. arXiv preprint arXiv:1507.07242.

[30] Chen, J. C., Ranjan, R., Kumar, A., Chen, C. H., Patel, V. M., &Chellappa, R. (2015). An end-to-end system for unconstrained face verification with deep convolutional neural networks. In Proceedings of the IEEE international conference on computer vision workshops (pp. 118-126).

[31] Sankaranarayanan, S., Alavi, A., Castillo, C. D., &Chellappa, R. (2016, September). Triplet probabilistic embedding for face verification and clustering. In 2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS) (pp. 1-8). IEEE.

[32] Masi, I., Rawls, S., Medioni, G., & Natarajan, P. (2016). Pose-aware face recognition in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4838-4846).

[33] Chen, J. C., Patel, V. M., &Chellappa, R. (2016, March). Unconstrained face verification using deep cnn features. In 2016 IEEE winter conference on applications of computer vision (WACV) (pp. 1-9). IEEE.

[34] Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

[35] Crosswhite, N., Byrne, J., Stauffer, C., Parkhi, O., Cao, Q., & Zisserman, A. (2018). Template adaptation for face verification and identification. Image and Vision Computing, 79, 35-48.

[36] Schroff, F., Kalenichenko, D., &Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

[37] Hu, J., Lu, J., Yuan, J., & Tan, Y. P. (2014, November). Large margin multi-metric learning for face and kinship verification in the wild. In Asian conference on computer vision (pp. 252-267). Springer, Cham.

[38] Hu, J., Lu, J., & Tan, Y. P. (2014). Discriminative deep metric learning for face verification in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1875-1882).

[39] Li, H., Hua, G., Shen, X., Lin, Z., & Brandt, J. (2014, November). Eigen-pep for video face recognition. In Asian conference on computer vision (pp. 17-33). Springer, Cham.

[40] Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1701-1708).