International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

BSF-RCNN-VFR: Background Subtracted Faster RCNN for Video based Face Recognition

Seshaiah M
Research Scholar, VTU - Regional Center
Bengaluru, India.

Dr. Shrishail Math
Professor, Dept. of CSE, SKIT
Bengaluru, India.

***Abstract*— Surveillance systems are widely deployed in various organization and public palaces to monitor suspicious activities and reduce the crime rate. Recently, visual surveillance systems has gained huge attraction from research community due to their significant impact on monitoring application. Several techniques have been developed which are based on the still image which do not provide efficient solution for real-time application. Hence, video based face recognition is considered as a tedious task. Recently, deep learning based schemes have been adopted widely for video face recognition but these techniques suffer from well-known challenges such as pose and illumination variation. Hence, we present a Convolutional Neural Network (CNN) based approach for video face recognition. According to the proposed approach, we employ background subtraction scheme which helps to reduce the scene complexity and improves the feature extraction process. Later CNN based face detection scheme is developed which uses region proposal generation networks. In this phase, we incorporate bounding box regression model to reduce the face detection error. Finally, RCNN based learning model is applied which uses Joint Bayesian learning to discriminate the classes of detected faces. Based on these stages, the proposed model is named as Background subtracted Faster RCNN for Video based Face Recognition (BSF-RCNN-VFR). An experimental study is carried out based on the proposed method where we use publically available datasets such as YouTube celebrity dataset, Buffy dataset and YouTube face dataset. The experimental study shows a significant improvement in the detection and recognition accuracy process.**

Keywords*— Face recognition, object detection, deep learning, video clips, CNN

I. INTRODUCTION

The demand of surveillance applications have gained huge attraction and the surveillance systems are widely deployed in various real-time applications. Generally, these security systems are based on the face identification which are based on the computer vision based systems and some applications are based on the biometric verification but this type of security systems are not considered as surveillance systems. In this work, we focus on the image and video based surveillance systems. The main aim of these surveillance systems is to detect, track and identify the human face for various applications such as crime investigation. Face recognition using digital images has been adopted widely for several offline and online applications such as Compact Binary Face Descriptor (CBFD) in [1], where binary patterns are extracted and these patterns are learned using unsupervised learning approach, Liu et al. [2] presented an open source framework known as VIPLFaceNetfor face recognition. Similarly, dictionary based approaches also widely introduced such as Jing et al. [3] presented Multi-spectral low-rank structured dictionary learning scheme for face recognition, kernel extended learning [4], Multi-feature kernel discriminant dictionary learning [5] and cost-sensitive dictionary learning [6].

In this field, pose and illumination are considered as a challenging task which degrades the recognition performance. In order to mitigate these issues, various pose and illumination based schemes of face recognition are developed, such as Tran et al. [8] presented a learning scheme for pose variation face recognition, Kakadiaris et al. [9] presented 3D-2D face recognition by considering both face and pose illumination variations, Hu et al. [10] presented singular value decomposition based approach for face recognition.

On the other hand, deep learning based schemes have been adopted widely in this field. Several deep learning methods are introduced during last decade such as Sphereface [7],Facenet2expnet: a deep learning model for face and expression recognition, Cosface [12], andArcface [13]. Similarly, convolutional networks are also considered as a promising solution for this types of applications. Based on this assumption, several approaches are introduced such as Sun et al. [14] presented improved faster RCNN for face recognition, faster RCNN [15], Wasserstein CNN [17] and CMS-RCNN [18]. The deep learning based models provide promising solution for pose and illumination variation for face detection and recognition.

Currently, face detection and recognition in video sequence is considered as a challenging task because of complexity caused due to the poor quality sequence and scene complexities. Compared to single image based recognition, the video based recognition systems provide useful information to improve the surveillance system. However, videos contains several information about the object and also it poses several challenges such as blur, out of focus, motion blur, occlusion, illumination variation and pose variations. Hence, designing a novel approach for feature extraction which can provide significant feature extraction for entire frames. Several techniques have been introduced for face detection, tracking and recognition from video sequences. Moreover, each subject has different number of face images which causes complexity in recognition due to the illumination and pose variation. Hence, computing these features and combining them across the different frames is also considered a challenging task. However, this type of method will require more memory and time hence it becomes a tedious task to detect and recognize the human faces in the video sequence.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

Sohn et al. [19] presented unsupervised scheme for face recognition from video dataset. Similarly, in [20-21] authors presented metric learning based scheme for face recognition from videos. Deep learning plays important role in this field of video based recognition. Various researches have been carried out for video based face recognition [23] such as Parchami et al. [22] presented haar-like deep convolutional neural networks based approach, deep-learning based emotion recognition [24], MDLFace using deep learning scheme, and deep CNN features in [26].

The aforementioned techniques of face recognition from video are based on the deep learning schemes which shows that the deep learning approach can significantly improve the performance of the system. There exists several challenges in video face recognition systems. Some of these challenges are discussed in next-subsection.

A. Issues and challenges

There exist several challenges in both still-image and video based face recognition system. Some of the issues are listed here which are as follows

(a) Pose variations: for each user, the pose may vary which causes complexity in detection and feature extraction hence less significant features are extracted from the varied pose face image which may lead towards the misclassification.
(b) Occlusions: sometimes occlusion may degrade the feature extraction process in still-image as well as video face recognition systems.
(c) Expression change: in these recognition system, there is a probability that user may change his or her expression at any time in the captured sequence which can cause poor feature extraction.
(d) Image resolution and illumination variation: these two are also considered as challenging task because extraction of features from poor quality images is a tedious task. Moreover, the varied illumination also affects the feature extraction process.

These issues affects the feature extraction process which degrades the learning process by inducing training error and class error. Due to these issue, the face recognition system may achieve poor performance in terms of detection and recognition accuracy.

Based on the study presented in this section, we introduce a novel approach for face recognition using CNN model. The main contribution of this work are as follows:

- First of all, we present a model for background subtraction which can help to improve the feature extraction process.
- In the next phase, we present faster RCNN based model for face detection along with the bounding box regression.
- CNN based model for face recognition and bounding

Rest of the article is organized as follows: section presents a brief survey about recent techniques of video based face detection and recognition. Section III presents the proposed solution using CNN model, experimental study is presented in section IV and finally, concluding remarks are given in section V.

II. LITERATURE SURVEY

This section presents a brief discussion about recent techniques of face recognition from video sequence. As discussed before that the video based face detection and recognition is considered as the most challenging task in visual surveillance system. Du et al. [27] presented video based face recognition system in multi-view videos. According to this study, the traditional approaches suffer from the pose variations hence in this work authors presented a novel feature extraction method by diffusion lighting and pose variations. In order to develop the feature extraction model, facial data is mapped onto a sphere using spherical harmonic representation. This texture map helps to generate the multi-view video data by back projecting the texture maps.

Huanget al. [28] presented a subspace representation based approach for video face detection. According to this study, the linear representation of videos shows a significant improvement in video analysis for human face detection. This linear representation lies in special type of non-Euclidean space known as Grassmann manifold. This technique presents a Fisher-like framework which helps to learn the projection metric. This process of learning is performed by mapping the data from the obtained Grassmann manifold to the new discriminant data.

Recently, the neural network and deep learning based approaches are being adopted widely in various computer vision applications. The deep learning models are capable to solve the complex architecture of pattern learning and provide the desired solution for classification. Based on this assumption, Herrmann et al. [29] presented convolutional neural network based model video face recognition. This work uses manifold based track comparison model to address the issues of low-resolution faces. The complete process is as follows: the image spaces are provided as input to the CNN architecture which generates the descriptor space for entire dataset. Later, local mean method (LM) is applied to tack the distance based on the similarity in the dataset. Moreover, a novel loss function is also incorporated to improve the detection accuracy. Fan et al. [30] focused on emotion recognition system and developed a novel approach using combined 3D-convolutional neural network and recurrent neural networks. These two models help to represent the encode the appearance and motion in different manner such as RNN processes the CNN features initially and later encodes the features of motion whereas the 3D CNN model uses appearance and motion features simultaneously.

The CNN based model use a loss function which is helpful for training the deep learning model. Generally, the soft-max loss function is widely adopted. Wen et al. [31] suggested that the learning performance can be improve and introduced a new function called as center loss. This function learns deep features for considered category of faces. Authors presented a mathematical modeling to show that the proposed loss function is easily trainable and also it can be used for optimizing the CNN model. This combined model of supervision of soft-max loss and center loss is helpful to achieve the deep features. Yang et al. [32] introduced neural network aggregation model for video face recognition. According to this process, a face image is processed with multiple images as input to the network and generates the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

compact and fixed-dimension features. The complete process is divided into two phases as feature embedding and feature aggregation. The feature embedding module uses deep convolutional network to map the face image to the feature vector whereas aggregation module aggregates the feature from the face image with the help of convex hull. Wu et al. [33] discussed that during training phase, the data label may be inaccurate and ambiguous which can lead towards the noisy labeled training data resulting in poor classification. Thus, authors presented a light CNN model for learning the pattern from noisy labels. According to this process, the max-feature map (MFM) activation function is applied into each convolutional layer which helps to separate the noisy data. This work also presents the three architecture of light CNN based on the concept of AlexNet, VGG andResNet. Bashbaghi et al. [23] presented a study on CNN architecture for face recognition from video sequence. In this work, authors have focused on the existing issues such as computational complexity, single training reference sample and domain adaption. This article studied about the deep learning architectures which are based on the triplet-loss function and autoencoder CNNs. This study shows that the Triplet-based loss optimization method is promising solution which can improve the learning process of the system.

Li et al. [35] presented a recurrent regression neural network (RRNN) for face recognition from still images and videos under varied pose conditions.

III. PROPOSED MODEL

As discussed in previous section, the video face recognition is a challenging task due to pose and illumination variations. Recent advancements in CNN shows that it can provide promising solution for this task. Several techniques have been introduced for face recognition using CNN but computational complexity and desired accuracy is still remains a challenging task. Hence, wepresent a novel approach for video face recognition where background subtraction, faster RCNN with bounding box regression for detection and CNN for recognition. The proposed approach is named as Background Subtracted Faster RCNN for Video based Face Recognition (BSF-RCNN-VFR).

A. Background Subtraction/Removal

The complex and continuously varying background can be responsible for generating false positive bounding boxes for detection. Hence, background removal can be considered as a promising task to improve the detection accuracy. In order to perform this task, we consider any video (video is converted into grayscale frame) or any still image. Let us consider that the video frame is presented in a single column vector as $\mathrm{f} \in \mathbb{F}^{\mathrm{N}}$where $\mathrm{N} = \mathrm{R} \times \mathrm{C}$. Let us consider that video frames are given as $\{\mathrm{f_j}\}, \mathrm{j} = 1,2, \ldots \mathrm{K}$. This can be represented vertically as:

$$\mathrm{y} = [\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_\mathrm{k}]^\mathrm{T} \in \mathbb{F}^{\mathrm{N}\times\mathrm{k}} \quad (1)$$

These frames are acquired from the same video hence frames are correlated to each other. Each frame can be represented as the sum of the common component which represents the background and foreground component as:

$$\mathrm{f_j} = \mathrm{z^c} + \mathrm{z_j^i} \quad (2)$$

z^cdenotes background component and z_j^i denotes foreground component of the considered f frame. Let $\gamma \in \mathbb{F}^{N\times N}$ is a matrix which is represented as an orthonormal basis which is used for representing the input image in the sparse form and the coefficients can be represented as $\alpha_j = \gamma f_j \in \mathbb{F}^N$ of the given input signal f_j as:

$$\alpha_j = \beta^c + \beta_j^i = \gamma z^c + \gamma z_i^c \quad (3)$$

Where β^cremains unchanged if the similar background is present in the frame and β_j^i is continuously varying according to the $j = 1,2, \ldots k$. The common component as given in (2) are represented as:

$$w = \left[\theta^c \theta_1^i \theta_2^i \ldots \theta_k^i\right]^T \in \mathbb{F}^{N(k+1)} \quad (4)$$

Eq. (2) can have multiple solutions to satisfy the conditions but here our main aim is to capture maximum solution about the foreground data and representing the w in a sparse manner of $\{f_j\}$, thus the frames can be represented in sparse manner as:

$$y = \bar{\gamma} w \quad (5)$$

Where $\bar{\gamma}$ is obtained by concatenating the background and foreground metrices as $\bar{\gamma} = [I_1 I_2]$ where I_1 is denoted as $[\gamma^T \gamma^T \ldots \gamma^T]^T \in \mathbb{F}^{(kN)\times N}$ and I_2 is denoted as $diag(I_1) \in \mathbb{F}^{(kN)\times kN}$. The optimal solution for foreground/common component can be obtained by performing the l_1 minimization problem. This can be given as:

$$min_w \frac{1}{2}\|y - \bar{\gamma} w\|_2^2 + \lambda \|w\|_1 \quad (6)$$

At this stage, the common and innovative components can be recovered by applying inverse transform as:

$$z = \Lambda w \quad (7)$$

Where $\Lambda = diag([\gamma^T \gamma^T \ldots \gamma^T]) \in \mathbb{F}^{k(N+1)\times k(N+1)}$. The above given minimization problem can be resolved by applying Separable Surrogate method and z can be obtained by applying inverse transformation.

B. CNN for Face Detection

In this section we present the proposed solution for face detection in any given video sequence. According to this process, we presented a cascaded convolutional neural network based approach. The proposed approach is divided into two main stages, first of all, we apply proposal generation to obtain the initial bounding boxes and then a bounding box regression model is applied. In the next stage, we apply the Refine network which helps to reduce the false positive bounding boxes which is performed by calibrating the refine network with bounding box regression.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

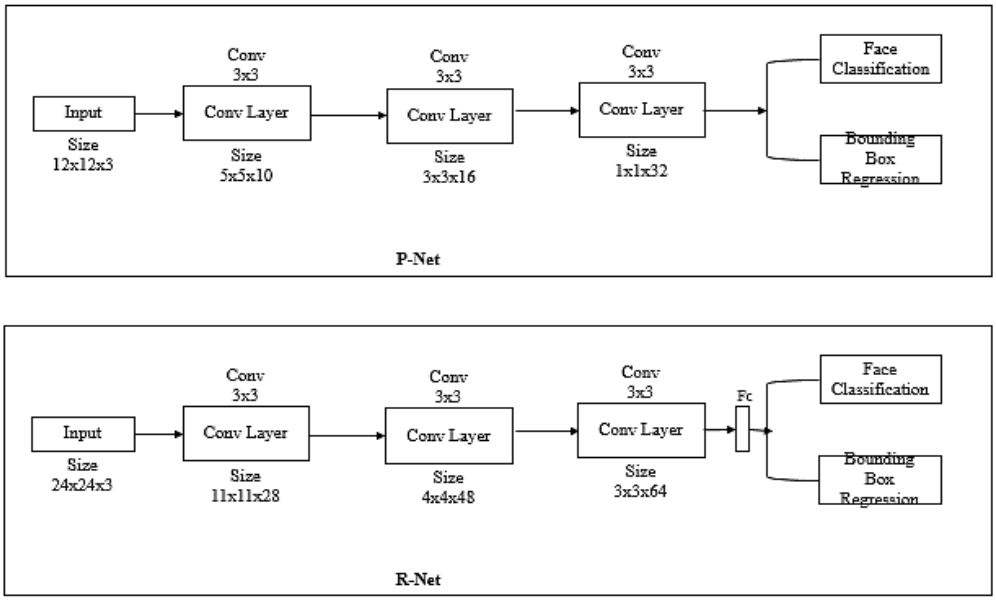


Fig. 1. Face detection and Bounding Box Regression Architecture

Above given figure shows a complete architecture of two stages where faces is classified and bounding box regression is applied. Finally, obtained faces are considered for further processing. This network architecture is trained using following models of CNN which are: face/non-face classification, bounding box regression.

In order to classify the face/non-face models, learning problem is formulated as two-class problem. Let us consider that for each sample x_i we define a cross entropy loss as:

$$L_i^{det} = -\left((1 - y_i^{det})(1 - \log(p_i)) + (y_i^{det} \log(p_i)\right) \quad (8)$$

Where p_i denotes the probability of sample to be a face image.

After identifying the face, we focus on the bounding box regression. In this phase, we consider a candidate window and predict the bounding box coordinates which are similar or near to the ground truth labels. This is formulated as a regression problem and Euclidean loss function is applied for each sample. This loss function is expressed as:

$$L_i^{box} = \left\|\hat{y}_i^{box} - y_i^{box}\right\|_2^2 \quad (9)$$

$\hat{y}_i^{box}$ denotes the coordinated obtained using CNN and y_i^{box} denotes the groundtruth coordinates of face. Similarly, landmark detection process is also implemented where a regression problem is formulated and Euclidean loss is minimized to obtain the landmark points, given as:

$$L_i^{box} = \left\|\hat{y}_i^{landmark} - y_i^{landmark}\right\|_2^2 \quad (10)$$

In this process, we assign different tasks to the CNN and also different types of training images are present in this learning process. The overall learning target can be formulated as:

$$\min \sum_{l=1}^{N} \sum_{j \in \{det, box\}} \alpha_j \beta_i^j L_i^j \quad (11)$$

Where N denotes the total number of training samples used

C. CNN for Face Verification

This section presents CNN model for face recognition where we consider deep feature learning. The feature learning is performed on the identified coordinates. According to the proposed architecture, the input layer is dimension is considered as 100x100x1 for any given input image. In this network, we have considered 10 convolutional layer, 1 fully connected layer and 5 pooling layers and each convolutional layer is provided a rectified linear unit (ReLU) except the last convolutional layer. In order to improve the performance we incorporate parametric ReLu and also two extra convolutional layers also included as Conv12 and Conv22 which is helpful to mitigate the effects of illuminations variations. In this work, we have used a 3x3 filter where first four pooling layer utilize max pooling operator and last layers of this network uses average pooling. To obtain the effective classification these low-dimensional features contain strong discriminative information of the training and testing face images from which are obtained from the face detection module. Here, pool5 layer features are used for representing the faces and CNN face features are optimized using L2-normalization before processing feature learning. However, in video sequence there exists several frames hence we use average feature of the pool5 features for representing the feature of entire sequence. In order to learn the feature, we use a Bayesian learning process where i^{th} and j^{th} images are modeled directly in the form of joint distribution using Gaussian distribution. Let us consider that the joint distribution of these images is represented as $P(x_i, x_j|H_I) \sim N(0, \Sigma_I)$ for the condition where input images x_i and x_j belongs to the same class, similarly, if these images belong to the different class, then the Gaussian distribution is expressed as, $P(x_i, x_j|H_E) \sim N(0, \Sigma_E)$. With the help of Gaussian distribution, the log-likelihood ratio between inter and intra classes can be computed as:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

$$r(x_i, x_j) = \log \frac{P(x_i, x_j|H_I)}{P(x_i, x_j|H_E)} = x_i^T G x_i + x_i^T G x_i - 2x_i^T R x_i \quad (12)$$

Where G and R and the two semi-definite matrices. In this learning process, the face vector can be modeled as $x = \mu + \xi$ where μ denotes the identity and ξ denotes the pose, and illumination variation. These both parameters are summed up into a zero-mean Gaussian distribution as $N(0, S_\mu)$and $N(0, S_\xi)$. Here, we focus on the estimation of S_μ and S_ξ by optimizing the distance as:

$$\underset{G,B,b}{\operatorname{argmin}} \sum_{i,j} \max\left[1 - y_{i,j}\left(b - (x_i - x_j)^T G(x_i - x_j) + 2x_i^T B x_j\right), 0\right] \quad (13)$$

$y_{i,j}$is the label pair such as $y_{i,j} = 1$ if x_i and x_j are the same person, otherwise $y_{i,j} = -1$..

IV. Results and Discussion

In this section we present the experimental analysis using proposed approach for face detection and recognition from video datasets. The proposed method is evaluated for using publically available dataset which are known as YouTube Faces [36], YouTube Celebrities, Buffy. Below given figure shows some sample images of the YouTube celebrity dataset.

The performance of proposed approach is compared with the existing techniques. In this work we also compare the performance of proposed approach in terms of tracking. The tracking performance analysis is presented in below given next sub-section.

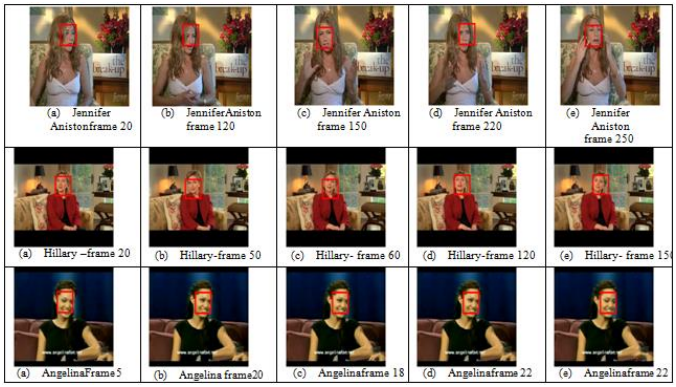
A. *Video Face Tracking Performance*

In this work we have considered the experimental setup as presented in [37]. We consider five movie trailer from the dataset which are Killer Inside', 'My Name is Khan', 'Biutiful', 'Eat, Pray, Love', and 'The Dry Land'. In order to measure the performance we use object tracking accuracy and object tracing precision. The obtained performance is presented in table 1.

TABLE I. FACE TRACKING PERFORMANCE

Name of the Sequence	Scale	KLT [38]	MSSRC [37]	BSF-RCNN-VFR
'The Killer Inside'	Pre.	68.93	69.35	86.33
	Acc.	42.88	42.16	92.28
'My Name is Khan'	Pre.	65.63	65.77	90.28
	Acc.	44.26	48.24	91.24
'Biutiful'	Pre.	61.58	61.34	88.56
	Acc.	39.28	43.96	92.36
'Eat Pray Love'	Pre.	56.98	56.77	85.21
	Acc.	34.33	35.6	89.66
'The Dry Land'	Pre.	64.11	62.7	93.52
	Acc.	27.9	30.15	91.2
Pre.: Precision (%), Acc.: Accuracy (%)				

The obtained tracking results are depicted in below given figure. Each row of the figure shows the tracking results for different frames of 5 videos.



B. *Video Face Recognition Performance*

In this section we present the face recognition performance using proposed approach. This experiment is carried out using YouTube Faces Dataset, YouTube Celebrities Dataset and Buffy Dataset

The YouTube face dataset is a huge dataset which contains total 3,425 videos which are acquired from 1,595 different people. These videos are obtained from the YouTube. The shortest clip duration is 48 frames, the longest clip is 6,070 frames, and the average length of a video clip is 181.3 frames. Similarly, YouTube celebrity data is obtained which contains total 1910 videos of 47 different people. The minimum frames are 8 and the maximum frame in a video are 400 in this dataset.

The buffy dataset contains total 639 face tracks which are extracted from the TV series "Buffy the Vampire Slayer", this dataset is obtained from the episodes 9, 21 and 45.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

The recognition results for buffy video sequence are presented in below given figure where the correctly detected faces are presented in white bounding box and incorrect recognition is depicted in red bounding box.

The performance of Buffy dataset is measured in terms of average precision and compared with the existing techniques. The obtained performance comparison is presented in table 2.

TABLE II. AVERAGE PRECISION PERFORMANCE COMPARISON FOR "BUFFY DATASET"

Buffy Episode	[39]	[40]	[40] with LP	[40] Raw	BSF-RCNN-VFR
1	90	98	99	92	99
2	83	96	96	95	98
3	73	95	95	91	98
4	86	97	96	92	99
5	85	97	97	93	98

Similarly, we presented an experimental study using YouTube celebrity dataset and the accuracy is obtained as 95.52% which shows a better improvement when compared with the existing techniques. The obtained performance is compared as presented in table 3.

TABLE III. RECOGNITION ACCURACY COMPARISON FOR YOUTUBE CELEBRITY DATASET

Techniques	Recognition Accuracy (%)
HMM [41]	71.24
MDA [42]	67.20
SANP [43]	65.03
COV+PLS [44]	70.10
UISA [45]	74.60
MSSSRC [37]	80.75
Proposed	95.52

These comparative studies show that the proposed CNN based architecture achieves better performance in terms of face detection, recognition and tracking.

V. CONCLUSION

In this work, we have focused on the face detection, tracking and recognition using convolutional neural network based approach. According to the proposed approach, first of all we present background removal model which helps to extract and learn the features significantly. Later, we developed a CNN based architecture which helps to detect the face region and also a bounding box regression method is also developed. In order to learn the features we apply CNN based feature learning model with log-likelihood ratio computation between inter and intra features,and based on those features the face recognition is performed. An extensive experimental is carried out which shows improved performance of proposed approach. The proposed approach attains precision of 93.52% on 'The Dry Land' movie, 98-99% precision on Buffy Dataset and 95.52% accuracy on YouTube celebrity dataset that outperforms recent standard works.

REFERENCES

[1] Lu, J., Liong, V. E., Zhou, X., & Zhou, J. (2015). Learning Compact Binary Face Descriptor for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(10), 2041–2056. doi:10.1109/tpami.2015.2408359.

[2] Liu, X., Kan, M., Wu, W., Shan, S., & Chen, X. (2017). VIPLFaceNet: an open source deep face recognition SDK. Frontiers of Computer Science, 11(2), 208-218.

[3] Jing, X. Y., Wu, F., Zhu, X., Dong, X., Ma, F., & Li, Z. (2016). Multi-spectral low-rank structured dictionary learning for face recognition. Pattern Recognition, 59, 14-25.

[4] Huang, K. K., Dai, D. Q., Ren, C. X., & Lai, Z. R. (2016). Learning kernel extended dictionary for face recognition. IEEE transactions on neural networks and learning systems, 28(5), 1082-1094.

[5] Wu, X., Li, Q., Xu, L., Chen, K., & Yao, L. (2017). Multi-feature kernel discriminant dictionary learning for face recognition. Pattern Recognition, 66, 404-411.

[6] Zhang, G., Sun, H., Ji, Z., Yuan, Y. H., & Sun, Q. (2016). Cost-sensitive dictionary learning for face recognition. Pattern Recognition, 60, 613-629.

[7] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 212-220).

[8] Tran, L., Yin, X., & Liu, X. (2017). Disentangled representation learning for pose-invariant face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1415-1424).

[9] Kakadiaris, I. A., Toderici, G., Evangelopoulos, G., Passalis, G., Chu, D., Zhao, X., ...&Theoharis, T. (2017). 3D-2D face recognition with pose and illumination normalization. Computer Vision and Image Understanding, 154, 137-151.

[10] Hu, C., Lu, X., Ye, M., & Zeng, W. (2017). Singular value decomposition and local near neighbors for face recognition under varying illumination. Pattern Recognition, 64, 60-83.

[11] Ding, H., Zhou, S. K., &Chellappa, R. (2017, May). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 118-126). IEEE.

[12] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., ...& Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5265-5274).

[13] Deng, J., Guo, J., Xue, N., &Zafeiriou, S. (2018). Arcface: Additive angular margin loss for deep face recognition. arXiv preprint arXiv:1801.07698.

[14] Sun, X., Wu, P., & Hoi, S. C. (2018). Face detection using deep learning: An improved faster RCNN approach. Neurocomputing, 299, 42-50.

[15] Jiang, H., & Learned-Miller, E. (2017, May). Face detection with the faster R-CNN. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 650-657). IEEE.

[16] Yin, X., & Liu, X. (2017). Multi-task convolutional neural network for pose-invariant face recognition. IEEE Transactions on Image Processing, 27(2), 964-975.

[17] He, R., Wu, X., Sun, Z., & Tan, T. (2018). Wasserstein cnn: Learning invariant features for nir-vis face recognition. IEEE transactions on pattern analysis and machine intelligence.

[18] Zhu, C., Zheng, Y., Luu, K., &Savvides, M. (2017). Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection. In Deep learning for biometrics (pp. 57-79). Springer, Cham.

[19] Sohn, K., Liu, S., Zhong, G., Yu, X., Yang, M. H., &Chandraker, M. (2017). Unsupervised domain adaptation for face recognition in unlabeled videos. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3210-3218).

[20] Huang, Z., Wang, R., Shan, S., Van Gool, L., & Chen, X. (2018). Cross euclidean-to-riemannian metric learning with application to face

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 06, June-2019

recognition from video. IEEE transactions on pattern analysis and machine intelligence, 40(12), 2827-2840.
[21] Huang, Z., Wang, R., Shan, S., & Chen, X. (2015). Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning. Pattern Recognition, 48(10), 3113-3124.
[22] Parchami, M., Bashbaghi, S., & Granger, E. (2017, May). Video-based face recognition using ensemble of haar-like deep convolutional neural networks. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 4625-4632). IEEE.
[23] Bashbaghi, S., Granger, E., Sabourin, R., &Parchami, M. (2018). Deep Learning Architectures for Face Recognition in Video Surveillance. arXiv preprint arXiv:1802.09990.
[24] Kaya, H., Gürpınar, F., & Salah, A. A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image and Vision Computing, 65, 66-75.
[25] Goswami, G., Bhardwaj, R., Singh, R., &Vatsa, M. (2014, September). MDLFace: Memorability augmented deep learning for video face recognition. In IEEE International Joint Conference on Biometrics (pp. 1-7). IEEE.
[26] 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.
[27] Du, M., Sankaranarayanan, A. C., &Chellappa, R. (2014). Robust face recognition from multi-view videos. IEEE transactions on image processing, 23(3), 1105-1117.
[28] Huang, Z., Wang, R., Shan, S., & Chen, X. (2015). Projection metric learning on Grassmann manifold with application to video based face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 140-149).
[29] Herrmann, C., Willersinn, D., &Beyerer, J. (2016, August). Low-resolution convolutional neural networks for video face recognition. In 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 221-227). IEEE.
[30] Fan, Y., Lu, X., Li, D., & Liu, Y. (2016, October). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 445-450). ACM.
[31] Wen, Y., Zhang, K., Li, Z., &Qiao, Y. (2016, October). A discriminative feature learning approach for deep face recognition. In European conference on computer vision (pp. 499-515). Springer, Cham.
[32] 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).
[33] Wu, X., He, R., Sun, Z., & Tan, T. (2018). A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security, 13(11), 2884-2896.
[34] Kim, K., Yang, Z., Masi, I., Nevatia, R., &Medioni, G. (2018, March). Face and body association for video-based face recognition. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 39-48). IEEE.
[35] Li, Y., Zheng, W., Cui, Z., & Zhang, T. (2018). Face recognition based on recurrent regression neural network. Neurocomputing, 297, 50-58.
[36] https://www.cs.tau.ac.il/~wolf/ytfaces/index.html#download
[37] Ortiz, E. G., Wright, A., & Shah, M. (2013). Face recognition in movie trailers via mean sequence sparse representation-based classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3531-3538).
[38] M. Everingham and J. Sivic. Taking the bite out of automatednaming of characters in TV video. CVIU, 2009. 3, 6, 7
[39] J. Sivic, M. Everingham, and A. Zisserman. “Who are you?” – learning person specific classifiers from video.In Proc. CVPR, 2009.
[40] Parkhi, O., Rahtu, E., Cao, Q., &Zisserman, A. (2018). Automated video face labelling for films and TV material. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1. doi:10.1109/tpami.2018.2889831.
[41] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-worldvideos. In CVPR, 2008.
[42] R. Wang and X. Chen. Manifold Discriminant Analysis. InCVPR, 2009
[43] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In CVPR, 2011.
[44] L. Wolf, T. Hassner, and Y. Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. TPAMI, 2011.
[45] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for Video-Based Face Recognition. In CVPR,2012.