Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020)
IEEE Conference Record # 48766; IEEE Xplore ISBN: 978-1-7281-5371-1

Face Recognition from Video using Deep Learning

Saibal Manna
Electrical Engineering Department
NIT Jamshedpur
Jharkhand, India
mannasaibal1994@gmail.com

Sushil Ghildiyal
Electrical Engineering Department
VIT Vellore
Tamil Nadu, India
info.sushil123@gmail.com

Kishankumar Bhimani
Information & Communication Technology
Marwadi University, Rajkot
Gujarat, India
info.bhimani@gmail.com

Abstract— Face recognition (FR) and verification is the immeasurable technology to encounter any criminal activities nowadays. With the remarkable applications extending from criminal ID, security, and observation to amusement sites. This system (recognition of faces) is exceptionally helpful in banks, air terminals, and different associations for screening customers. In deep learning, convolutional neural networks (CNN) have gained attention for face recognition but to train CNN requires more data, which is very difficult in case of applications like criminal activities (robbery, murder, etc). Therefore, this paper proposed a face recognition system that makes searching for criminals easy and quick with less time and hence efficiently helps police and administration. In this paper, a pretrained model i.e FaceNet (FN) is used for face recognition from video. FN modifies the face images into a close-packed Euclidean space where separations extent the face nearness.

Keywords— Face recognition, convolutional neural network, security, FaceNet

I. INTRODUCTION

For human identification, the face is important. It is the most recognizable item in an individual. FR is interesting and complicated and affects key applications in many fields, such as security access, personalized identity, law enforcement identification, and banking authentication. Face detection is very easy for humans but for a machine, it is distinct. To date, there is very little knowledge on how an image is autonomized and how the cerebrum encodes it and inner (nose, mouth, eyes) or external highlights (face shape, structure, hairline) used for the effective recognition of the face? Neurophysiologist Wiesel and Hubel have proved that our minds respond to certain nerve cells, such as boundaries, curves, motion, or angles, in specific circumstances. Because nooone are viewing it as dissipated pieces, one way or the other our visual cortex should add useful examples to the various data sources. Automatic FR involves the removal, placing, and implementing certain classifications of the purposeful character from an image. The most intuitive technique for human identification is probably the FR based on the geometric highlights of a facial. The total operation may be categorized into three significant steps in which the first part is to find a decent database of human faces, each with many images. The second stage is to identify and prepare the FR faces in the database images and the last step is to check the FR to see the face for which it has been trained. Face detection (FD) is currently used in different places like Picasa, Photo Bucket, and Facebook. The natural tag character brings a new aspect to the image-sharing between the people in the picture. IN this way, the FR algorithm is reseaerhced and implemented in this paper that was straightforward but very powerful. For FR (this is the person), verification (this is the same person), the FN system is used in this article. Our technology depends on the use of a deep coevolutionary network to learn Euclidean embedding per image. The framework is trained to explicitly correlate to face similarity with the square L2 distances in the embedding space: images of similar individuals have little distance and faces of other individuals have a large distance. This model may be used by the police or the examining division to detect criminals. The FR strategy used is quick, relatively straightforward, robust, and accurate, with algorithms and techniques that are relatively easy to understand. Firstly, the simpler problem for our face detection problem is resolved. A device has been developed to detect a single face of the human being and multiple faces from both photographs and videos and to return their rectangular coordinates with their names.

II. RELATED WORK

From the seventies, FR was an important trend for science. Provided an input image with different faces, FR first runs the FD for face separation. Individual faces are preprocessed and eventually, a low-dimensional embedding is achieved. For a professional classification, a low-dimensional integration is important. For an intrapersonal variety of images, such as style, appearance, and age, face portraying should be powerful, while recognizing relationship image variations among different individuals.

FR and verification are both problems, which are popular in computer vision research and image processing. The neural frames that can be used for the task have made the subject more and more popular. Generally, more time is required for the preparation of neural networks, training data, and computer power; hence, a good deal of research has been undertaken to reduce these factors.

For over two decades, the issue of FR has been considered. The methodologies suggested in writing up until now are mainly classifiable as model and appearance-based [1].

The author runs a deep "warp" network to the canonical front, then learn from the CNN that categorizes every face to be part of an established identity. For facial verification, PCA in combination with an ensemble of SVMs is used [2].

Introducing a multi-stage approach that adapts to a general model of the three-dimensional structure [3]. A multi-class network is designed for more than 4,000 characters to carry out FR tasks. The designers also analyze the L1 distance between two faces using the siamese network. Their best performance on LFW (97.35 percent) is a group of 3 systems that use distinctive arrangements and shading channels. The expected distances from these networks are coupled with a non-linear SVM (non-linear SMV prediction based on χ^2 kernel).

They are introducing a compact and fairly cheap network to measure [4]. They use 25 of these systems, each of which operates on a different face patch. The developers pooled 50 responses (regular and reverse), for their final execution on LFW (99.47 percent [5]).

A Joint Bayesian and PCA [6] are used, which correspond adequately to a linear transformation in the embedding space. Their approach is not explicitly arranged in 2D/3D. The systems are developed using a combination of loss of verification and classification. The verification loss is the same as the TL [7], [8] as it restricts the L2 gap between faces with identical characteristics and sets a margin between face distances with a distinct personality. A similar loss to the one used here was examined by semantic and visual closeness for ranking photographs [9].

The separation vector approach between facial features, eyes, and ear size was first used for FR [6]. The vector contained 21 subjective characters with each emphasizing faces recognizable. They adopted a similar approach in 1973, using template matching to suit the facial characteristics globally [7]. The author developed the fully automatic FR (1973) on a computer system that included geometric parameters that extracted sixteen facial characters [8]. The average positive detection was above 50 percent correct. In the 1980s, few blueprints were enhanced calculating more enhanced subjective facial highlights and algorithms that were developed based on ANN. In1986, the author provided its faces based on the PCA [9]. The core concept was to replicate images of the lower dimensions without any information loss. Finally, in 1922, New Algorithms for the proper classification of heads on one's faces were implemented [10].

Here the author implements their crime identification program by using FR and detection paper [11]. In this paper, the fast data boost training algorithm is implemented [12]. For face representation, they use method joint identification-verification to their document [13].

III. METHODOLOGY

FaceNet is the facial recognition and clustering function of Google, which has an accuracy of 99.63 percentage [14]. The goal of this proposed paper is to achieve FR with high precision. FN uses the architecture of a deep neural network. The FN model architecture [14] is displayed in Fig.1. This maps a face image from Euclidean space, in which the distances directly correspond to an approximation of the face's similarity. When space is generated, different tasks can be easily accomplished. Face verification, identification, and clustering use regular FN embedding methods as the function vectors. Triplets are used for training the framework. Triplets are the set of a single image of an anchor, a positive image of an anchor, and a negative image of an anchor.

Let's follow the steps for pre-trained model:

1. Collect photos of people one stage ahead to develop a model.

2. Place the faces using Open CV, multi-task Cascaded Convolution Neural Networks (MTCNN). They recognize, align, and detect faces.

3. Using the pre-trained FN model to represent or embed the faces of all persons on 128-dimensional Euclidean space.

4. Accumulate the embeddings on a disc with the names of the respective user.

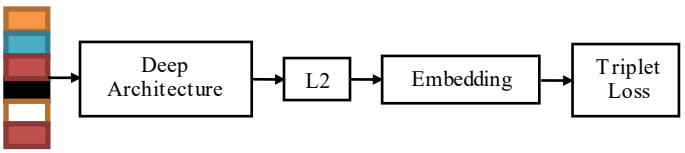

Fig. 1. Illustrating FaceNet Model architecture

Our framework includes a batch input layer and a deep CNN, preceded by standardization L2, leading to face embedding. This is trailed by the TL during preparation**.** The length of the segment that interfaces them is the L2 or Euclidean distance from two-point p and q. If p = $(p_1,p_2,\ldots,p_n)$ and q = $(q_1,q_2,\ldots,q_n)$ in Euclidean n-space are two points, then in "(1)" the distance(S) between p to q or q to p is provided.

$$S(p,q) = S(p,q) = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots.+(p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad (1)$$

Within Euclidean n Space, there is a Euclidean vector. The Euclidean length of a vector estimates the length of the vector. It represented as $\|S\|$. Our system of facial recognition is ready now. The 128-dimensional embeddings with the names of the person are available. Whenever a face is visible, the image is running through a previously trained network to create a 128-dimensional embedding, which is then compared by using Euclidean (L2) distance with stored embeddings. To this end, the triplet loss is employed to demonstrate and achieve the goal in FR and verification. In other words, a picture is integrated into a feature space, so that, irrespective of imaging conditions, the squared distance between every face, from different identities, is small while

there is a great squared distance between a couple of face pictures.

The main aim is to develop and constantly improve a comprehensive facial recognition system that works with any type of image. This change must be self-sufficient and allow citizens to be better recognized and included. Moreover, it is just a matter of time, as this identification must be made as close as possible to real-time. It is an extremely difficult problem to recognize faces, particularly outside regulated conditions. In reality, many methods have not succeeded in the course of history.

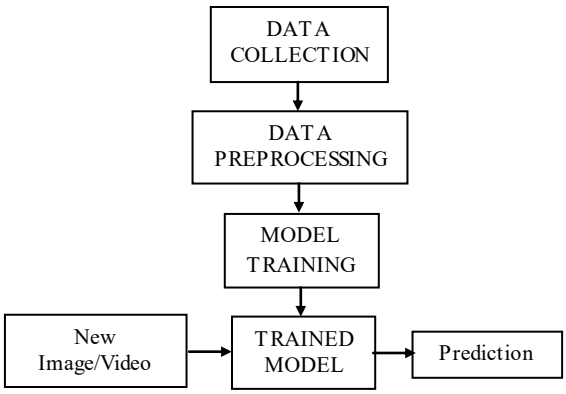

Fig. 2. Work Flow of the proposed methodology

Besides the difference between images of a similar face, like hair, lighting conditions, or expression, determining what makes the face visible is difficult. As a result, it should not start from scratch at the start of this project, but rather using certain existing work. This will allow us to accelerate the process and make it easier to produce results of quality. For this, a literature survey has been performed. Many effective ways are discovered and encouraged to address the issue. Finally, it has been chosen to focus on the model approach of FN. The principal causes are the good results obtained − the state of the art is indeed close − and the description's quality. The model FN gives good accuracy of 99.63 percent. Fig. 2 illustrate the workflow of the proposed methodology which involves various stages for face recognition.

A. Data collection

The database has been proposed. For data preparation, Indian actors' images are downloaded from Google. Our databases contain 8 subjects (persons) images. Each person has 100 images and among these 680 images were taken for training and 120 for testing.

The development of facial databases for benchmarking purposes has been a crucial element of the continuous advances made in the field of automated facial and appearance recognition. Since the 1990s, the tremendous developments in computer and sensor engineering have led to the development of new strategies for automatic FR [15-19]. There are currently several databases that are used to identify the face due to the size, articulations, position, condition of lights, obstacles, and image quality. The facial databases documented the variations in posture, lighting, imaging points, ethnicity, sexual orientation, and outward appearances from the year 2000 and beyond [15]. The absolute latest databases catch the varieties in picture sizes, pressure, impediments, and are assembled from shifted sources, for example, web and social media[20]. Probably the most recent facial databases are discussed below:

Labeled Wikipedia Faces (LWF) [21] has compiled photographs from more than 0.5 million biographical portions of the passages of the Wikipedia Living People and includes 8,500 appearances from 1,500 subjects.

YouTube Faces Database (YFD) [22] comprises 3425 recordings of 1595 distinct subjects (2.15 recordings per subject) with 48-6070 edge video cuts. A dataset was created to provide recognizable evidence from recordings and benchmarking video pair-coordinating procedures to an aggregation of recordings and names for the subject.

YouTube Makeup Dataset (YMD) [23] includes pictures from 151 subjects (Caucasian females) from YouTube cosmetics instructional exercises when unobtrusive to overwhelming cosmetics is connected. Four shots (2 previous shots and 2 after the cosmetic is connected) are taken for each subject. This database has a steady development, but due to cosmetic changes, it reveals issues with FR. The Indian movie face database (IMFD) [24] consists of 34512 photographs of 100 Indian acting artists, compiled from approximately 100 recordings and trimmed to integrate different forms of posture, mood, lighting, goals, impediments, and cosmetics.

B. FaceNet

Google scientists are presenting FN in 2015 [14]. It modifies the face like word embedding into 128D Euclidean space. FN is a one-shot model that can directly map facial photographs to a compact Euclidean environment where distances are directly balanced by face closeness measurements. Tasks such as facial control and recognition can be effectively done using the standard method with FN integrations as feature vectors when this space has been developed. In training, three times the identical / not identical facial patches were roughly matched. Triplets of approximately spaced identical/non-identical face patches were used to prepare. Illustration of the FN model shown in Fig. 3 [14]. In particular, f(z) has been integrated from an image z to a feature space, so that small square separation between all faces, free from image condition, is similar, while the square separation between a few face images from different characteristics is large. While it did not explicitly compare with other losses, such as those with positive and negative sets, as used in [14], it has been assumed that the triple loss is ideal for facial verification.

C. Triplet loss function

Triplet Loss (TL) is a particular type of loss used in FaceNet model, The TL limits the anchor-positive distance,

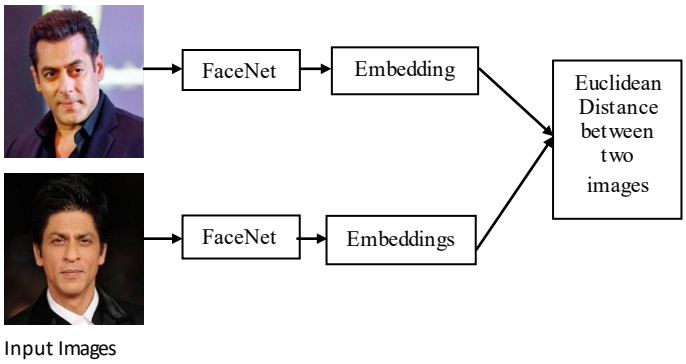

Fig. 3. FaceNet model illustration

both with a similar character, and enhances the anchor's distance from a negative individual. It has been preferred to see the anchor-positive distance less than the anchor-negative distance expressed in "(2)".

$$\left\|z_i^a - z_i^p\right\|_2^2 + \beta\left\langle\left\|z_i^a - z_i^n\right\|_2^2 \right. \quad (2)$$

Google introduces the TL in their FN paper [14] and can be used in networks that map data to a point in a d-dimensional space. It depends on calculating the distance between these points and gets its name by the use of triplets (three points), an anchor point (z_i^a), a positive point (z_i^p), and a negative point (z_i^n).

The anchor is in the same class with a positive point, while the negative point is another class. The overall loss aims to reduce the gap from the positive to the anchor point while maximizing the gap from the negative to the anchor. It is explained by "(3)" where β is an imposed margin between the positive and negative pairs.

$$Loss = \sum_{i=1}^{N}\left[\left\|f(z_i^a) - f(z_i^p)\right\|_2^2 - \left\|f(z_i^a) - f(z_i^n)\right\|_2^2 + \beta\right] \quad (3)$$

$f(z_i^a), f(z_i^p), f(z_i^n)$ indicates the output encoding of the anchor, positive and negative β is a constant used to ensure that the network isn't trying to optimize

$$f(z_i^a) - f(z_i^p) = f(z_i^a) - f(z_i^n) = 0$$

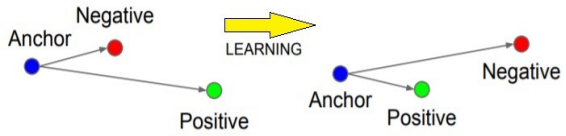

Fig. 4. Illustration of training with triplet loss

Fig. 4 reveals that negative and anchor points are close to one another than positive and anchor points. After training positive and anchor points are close to one another than negative and anchor points [14].

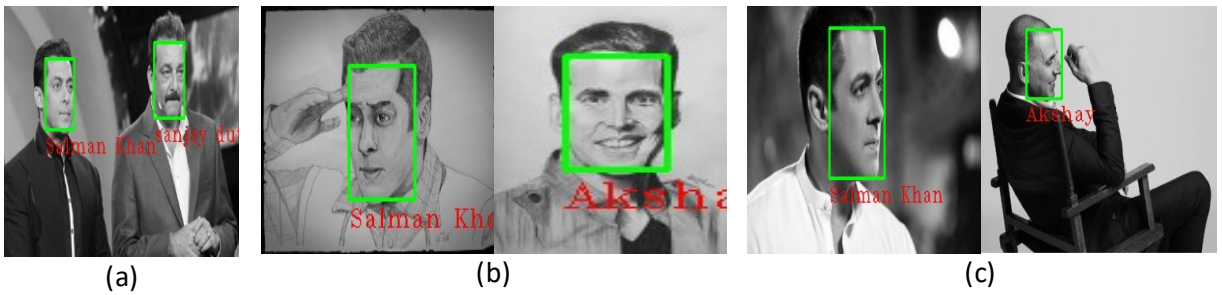

Fig. 5. Illustrate the face recognition using pretrained model i.e FaceNet. Fig. 5 (a) shows face recognition with real images, (b) shows with image drawings, (c) represents the recognition for side-face images, (d) shows recognition of face even with less light intensity and (e) recognizes from video even with dark background.

IV. RESULTS & DISCUSSION

A database has been built with the required people to recognize in the system. Pre-processing of data is a process that makes raw data understandable. Real data are often insufficient, incompatible, and/or missing in other habits or patterns and contains several errors. Each face will be cropped and each face will be labeled with the folder name. The model should be trained with a predefined model after the pre-processing of the data. Ultimately, the step is ready and it can use our video and image data to test this step. Python language is used to implement this process. The model is effectively applied and can recognize faces in painting, side faces, dark faces, still photos, and video. The result is shown below for different images.

Fig.5(a) shows that our system can detect human faces from the image. When the image is processed, the square box of the face is drawn and the name of the individual is composed at its lower portion. Fig.5(b) shows the hand-drawn picture of the Indian actor. Faces from hand-drawn pictures are recognized by this system. Fig.5(c) and Fig.5(d) show the side face image of the person and dark face image of one person. This model is recognized by faces above-mentioned figures. Fig.5(e) shows a video result picture. Given a video with a person in it, it was capable of following and recognizing them. After every video is handled, the square box of the face is drawn, and the name of the individual in it is composed at its bottom. Multiple faces from the video can be recognized by this system.

Table I. Comparison of the face recognition performance

Method	Number of networks	Accuracy
DeepFace	1	97.35%
DeepID1	60	97.20%
Face ++ v2014	-	97.30%
FaceNet	1	98.47%

Table I. While doing a literature survey it has been found that among all these models, the FaceNet model represents the highest accuracy, after training with a particular dataset. Realizing this fact the dataset ahs been collected and applied FaceNet on it and accounted for the accuracy of 90 %.

V. CONCLUSION

The suggested proposal is capable of correctly recognizing faces from videos as well as images. It can work with any sort of pictures and is sensibly strong to changes in face appearance or orientation, light conditions, and different variables. The benefit of this model is that it can recognize the side face and blurred image that other conventional models can't recognize. The acquired framework has been widely tried, and distinctive parameter combinations have been attempted. Given a video with a person in it, it was capable of following and recognizing them. After every video is handled, the square box of the face is drawn, and the name of the individual in it is composed at its bottom. Multiple faces from the video can be recognized by this system.

For preparing and testing, 800 face pictures dataset is collected. The facial recognition part has been tested and it has acquired steady outcomes with around 90% of accuracy. These outcomes are superior to anything expected, and they take into consideration some genuine use cases. In any case, there is still an opportunity to get better.

This can be applied in the future to recognize individuals using video capture that would help to identity from CCTV cameras that allow police to recognize the person within a fraction of time. It can also be incorporated in home security systems, visitor analysis systems as well. In the future, face detection as well as the time when faces are appearing in the video is implemented.

REFERENCES

[1] Fu Jie Huang Zhihua Zhou "Pose Invariant Face Recognition

[2] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. CoRR, abs/1404.3543, 2014. 2.

[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface:Closing the gap to human-level performance in face verification. In IEEE Conf. on CVPR, 2014. 1, 2, 5, 8.

[4] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. CoRR, abs/1406.4773, 2014. 1, 2, 3.

[5] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. CoRR, abs/1412.1265, 2014. 1, 2, 5, 8.

[6] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In Proc. ECCV, 2012. 2.

[7] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In S. Thrun, L. Saul, and B. Schölkopf, editors, NIPS, pages 41–48. MIT Press, 2004. 2.

[8] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In NIPS. MIT Press, 2006. 2, 3.

[9] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. CoRR, abs/1404.4661, 2014. 2.

[10] Turk, Matthew, and Alex P. Pentland. "Face recognition using eigenfaces." IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1991, pp. 586-591.

[11] Piyush Kakkar, Mr. Vibhor Sharma,"Criminal Identification System Using Face Detection and Recognition",International Journal of Advanced Research in Computer and Communication Engineering, Vol. 7, Issue 3, March 2018

[12] Hui-Xing, J., Yu-Jin, Z.: Fast Adaboost Training Algorithm by Dynamic Weight Trimming. Chinese Journal of Computers (2009).

[13] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. CoRR, abs/1406.4773, 2014. 1, 2, 3.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815-823

[15] Tian, Ying-li, Takeo Kanade, and Jeffrey F. Cohn. "Recognizing action units for facial expression analysis." IEEE Transactions on Pattern Analysis and Machine Intelligence, 23.2 (2001): 97-115.

[16] V. Bettadapura (2012). Face expression recognition and analysis: the state of the art. ArXiv preprint arXiv:1203.6722.

[17] National Science and Technology Council,"Face Recognition".

[18] University of Notre Dame, "CVRL Data Sets-Biometrics Data Sets," March2011.[Online].Available:http://www3.nd.edu/~cvrl/CVRL/Data_Sets.html.

[19] Lucey, Patrick, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression." In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 94-101.

[20] H.-W. Ng, S. Winkler. A data-driven approach to cleaning large face datasets. Proc. IEEE International Conference on Image Processing (ICIP), Paris, France, Oct. 27-30, 2014

[21] Hasan, Md Kamrul, and Christopher J. Pal. "Improving alignment of faces for recognition." In IEEE International Symposium on Robotic and Sensors Environments (ROSE), 2011, pp. 249-254.

[22] Wolf, L., Hassner, T., & Maoz, I. (2011, June). Face recognition in unconstrained videos with matched background similarity. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011 (pp. 529-534).[27] Chen, C., Dantcheva, A., & Ross, A. (2013, June). Automatic facial makeup detection with application in face recognition. In International Conference on Biometrics (ICB), 2013 (pp. 1-8).

[23] Chen, C., Dantcheva, A., & Ross, A. (2013, June). Automatic facial makeup detection with application in face recognition. In International Conference on Biometrics (ICB), 2013 (pp. 1-8).

[24] Setty, S., Husain, M., Beham, P., Gudavalli, J., Kandasamy, M., Vaddi, R. & Jawahar, C. V. (2013, December). Indian Movie Face Database: A benchmark for face recognition under wide variations. In IEEE 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), (pp. 1-5).