Real-time Age-Invariant Face Recognition in Videos Using the ScatterNet Inception Hybrid Network (SIHN)

Saurabh Bodhe[1,2], Prathamesh Kapse[1,2] and Amarjot Singh[1]

[1]Skylark Labs, San Francisco, USA
[2]National Institute of Technology, Warangal, India
{saurabh, prathamesh, amarjot}@skylarklabs.ai

Abstract

Face recognition has become a vital component of various safety and security systems with applications in safety and security systems, law enforcement applications, access control etc. Ageing makes face recognition challenging as the facial features evolve over time. In this paper, we propose a ScatterNet Inception Hybrid Network (SIHN) network that learns deep features for age-invariant face recognition. The trained system is evaluated on a separate dataset of 200 videos corresponding to 100 celebrities collected from public sources. These videos contain faces recorded at different locations, scales, rotations, illumination and ages. Experimental results evaluated over 27000 frames show that the proposed method can achieve state-of-the-art performance on both our video dataset as well as the other widely used datasets for age-invariant face datasets such as CACD and FG-NET. The system finds the individuals of interest from the videos in real-time at 18 fps. This research also introduces the Celebrities Video Aging (CVA) dataset used for evaluating the deep network which hopefully may encourage researchers interested in using deep learning for age-invariant face recognition.

1. Introduction

Sci-Fi and crime dramas have recently shown the use of extremely advanced facial recognition systems. Recent advances have made these seemingly unrealistic systems a reality [23, 27, 38, 21, 43]. These systems have proven to be very useful for numerous applications including face recognition in crowds [23], disguised face recognition [27, 38], kinship verification [21], and most recently age-invariant face recognition [43].

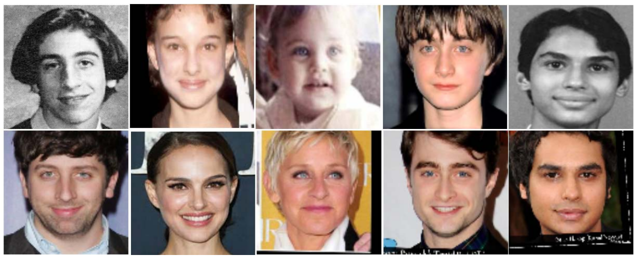

Figure 1. Examples of effect of aging on human faces. Top image is a childhood images of celebrities and the corresponding images below those are the adult images of same celebrities

Temporal invariance or age-based invariance in face recognition is of particular interest to law enforcement agencies as it can be used to search for missing children [2, 30], recognize individuals of interest over time [42], perform passport photo verification [16] etc.

This is a challenging task because human faces can vary significantly over time in many aspects, including facial texture (e.g. wrinkles), shape (e.g. weight gain), facial hair, the presence of glasses, etc. In addition, the image acquisition conditions and environment may also change, which can cause-uniform illumination and scale changes. Fig. 1 shows several such examples with different age gaps.

A number of systems have recently been proposed which can effectively perform age-invariant face recognition. These systems are primarily trained and evaluated on few face-frontal images of individuals recorded in a well-lit camera setting with minimal variations in position, rotation, scale or illumination. In this paper, we propose the ScatterNet Inception Hybrid Network (SIHN) that can, in real-time, solve the challenging task of age-invariant face recognition in real-world videos. The proposed SIHN network learns deep features for age-invariant face recognition rapidly from relatively fewer labelled examples. The network is trained and evaluated on a dataset of 200 real-world celebrities videos collected from public sources. These videos contain faces recorded at different locations, scales, rotations and illumination variations.

The novelties of the proposed system and the advantages over other techniques are detailed below:

- ***ScatterNet Inception Hybrid Network***: The proposed SIHN network for age-invariant face recognition consists of a hand-crafted ScatterNet front-end and an Inception ResNet-v1 (IR) [40] based backend. The SIHN network is constructed by replacing the first two convolutional, relu and pooling layers from stem of the IR network with the ScatterNet [34]. This improves learning speed of the Inception ResNet-v1 network as the ScatterNet front-end extracts invariant (translation, rotation, and scale) [31] edge features which can be directly used to learn more complex patterns from the start of learning. Processing speed is very important in this application to enable real-time recognition, SIHN complimenting this requirement.
- ***False Positive Removal***: The paper introduced three novel strategies for false positive removal for this application. They are described here: (i) A gender determination algorithm is used to ensure that the gender of the detection is same as that of query image. (ii) A detection is considered to be true only when the same detection appears in the previous frame as well as the next frame of the video. This eliminates the majority of false positives (iii) Certainty metrics are proposed (Section 4.4) to show how certain the system is about its detection. These metrics can be used to set a threshold. A recognition resulting in a certainty metric below the threshold is ignored.
- ***Proposed Datasets***: The paper presents the Celebrities Video Aging (CVA) dataset of 200 videos of 100 labeled individuals and about 557 unknown individuals. The dataset contains videos with faces recorded at different variations of scale, illumination, orientation, blurriness, etc. In addition, we propose the Modified Large Age Gap (MLAG) dataset having total of 5,045 images of 1,010 celebrities. This dataset was created by making additions to the original LAG dataset to take in account the variations in orientation, blurriness etc. which were not present in LAG. These datasets may encourage researchers interested in using deep learning for age-invariant face recognition.
- ***Real-time Identification***: In addition to the time boost provided by the SIHN, few more techniques are used to reduce processing time. Frame skipping is used in videos with high frames per second (fps) measure. This can be done as several intermediate frames contain redundant information. Video formats with least read-write delay are used at the cost of extra storage space, saving the time taken in compression and decompression.

The proposed Age-Invariant Face Recognition system can be used to identify faces in a running video stream which correspond to childhood images already in the database. The face recognition performance of the system is also compared with the state-of-the-art methods.

The paper is divided into the following sections. Section 2 summarizes the previous related work done on this topic. while Section 3 introduces the proposed CVA and MLAG datasets. Section 4 describes the Proposed System, Section 5 details the experimental results and Section 6 concludes this research.

2. Related Work

In recent years, many systems have been proposed on age-invariant face recognition or verification [12, 18, 26, 11, 5]. Generally, all these approaches can be categorized into two groups.

The first is based on the generative approaches [10, 17, 26, 11], which construct 2D or 3D generative models to compensate for the aging process and synthesize face images that match the age of query face images. These models heavily depend upon strong parametric assumptions, clean training data, as well as accurate age estimation and hence, are limited in unconstrained environments. The second set of approaches are based on discriminative models [18, 5, 19, 25, 8] which use robust facial features and discriminative learning methods to reduce the gap between face images captured at different ages. Probabilistic Linear Discriminant Analysis (PLDA) [14] was employed to establish a generative linear model whose latent space was iteratively derived by using the Expectation-Maximization (EM) [22] algorithm to solve the task of age-invariant face recognition.

The second set of approaches rely on discriminative methods. Li et al. [18] use multi-feature discriminant analysis for close-set face identification. Gong et al. [11] proposed to separate identity and age components using hidden factor analysis. Ling et al. [19] used gradient orientation pyramid with SVM for age-invariant face recognition. Otto et al. [25] used facial component localization to analyze variations due to age, Chen et al. [5] used the Cross Age Reference Coding (CARC) method. CARC tries to use a set of reference faces in order to accomplish the task of age-invariant face recognition. It encodes the low level facial features with the representation in the reference space.

These reported systems have been reasonably successful in matching the picture of the child to that of its adult self. However, these systems perform matching only when the images (in the database and to be matched) contain frontal faces, recorded from close proximity. This limits the applicability of these systems in real-world scenarios as the image in the database needs to be matched to a video or video frames which may contain numerous faces that can appear at different positions, orientations, and scales. These

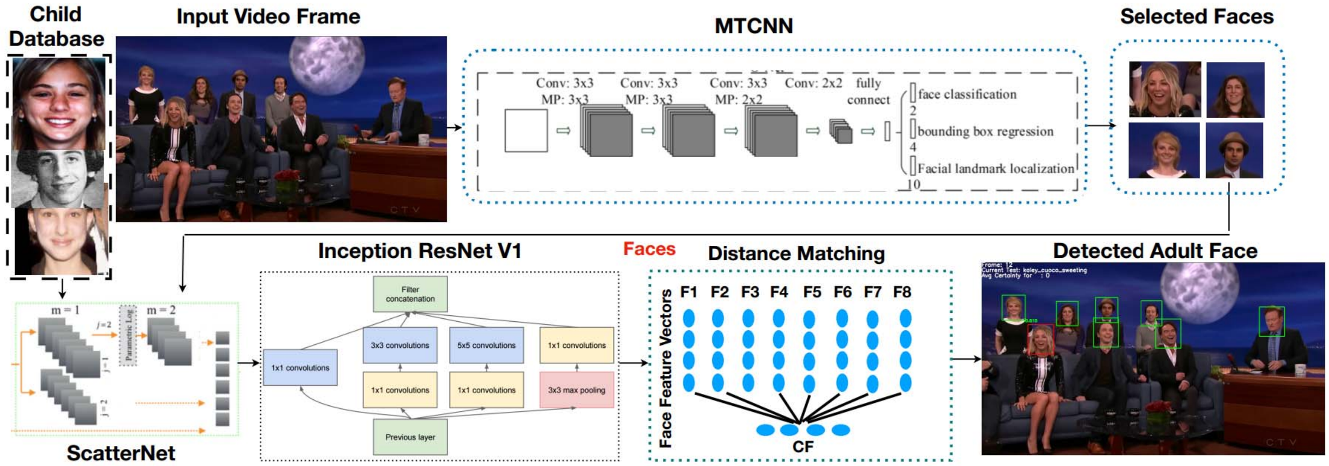

Figure 2. Pipeline of proposed age-invariant recognition system. The MTCNN is used to detect all the faces in input video frame to get the bounding boxes. The cropped faces with the target childhood image are sent to the SIHN for embedding generation. The low level features extracted by ScatterNet at L0, L1, and L2 are concatenated and given as input to the Inception ResNet (IR), which in turn outputs the embedding. By distance matching, the most similar face is determined and certainty metrics are calculated. In addition, false positive checks are done, if passed, the similar face is marked in red box and others are marked in green.

videos can also be affected by noise and illumination variations which further complicates this problem. In addition, there are no computation optimizations presented by these methods. This makes these methods ineffective in real time scenarios where fast processing is a must.

Chen et al. [5] has done extensive literature review on the previous related work. The same matches up well for our work, and hence this section is a derivative from same section of that work.

3. Datasets

Many cross age datasets have been presented over recent years. Most of these do not provide sufficient variations for effective training in real world conditions like blurriness etc. as shown in Table 1. We present two custom datasets below, the Celebrities Video Aging (CVA) dataset in unconstrained video format and Modified Large Age Gap (MLAG) dataset which is our extension to the original LAG dataset [3] by adding additional faces to account for more drastic variations.

3.1. Modified Large Age Gap (MLAG) Dataset

LAG [3] is a smaller dataset characterized by images of each person at different wide spaced ages, it contains 3,828 images of 1,010 celebrities. For each identity at least one child image and one adult image are present. This dataset was modified and extended by adding more cropped faces from few videos retrieved from public sources similar to CVA dataset. 1,217 more images were added to the dataset to make the new dataset having total of 5,045 images of 1,010 celebrities. This addition of images was done to take in account the variations in illumination, orientation, blur levels etc. These variations were more or less absent in the original LAG dataset. The meaning of large age gap is twofold: one way it refers to photos with extreme difference in age, e.g. a child face of 5 years old may be matched with adult face of 65 years; on the other hand it also refers to large difference in appearance due to the aging process: for example, 5 years old to 20 years old is numerically not a very large difference in age but may be very large difference in appearance due to rapid changes during early years of a person. The MLAG dataset covers both aspect. The dataset contains images ranging from 2 years old to about 90 years old. This modified LAG dataset was used to fine tune the pretrained model to learn age progression.

3.2. Celebrities Video Aging (CVA) Dataset

This research proposes an annotated Celebrities Video Aging (AVI) dataset which is used by the proposed SIHN network to evaluate age-invariant facial recognition performance. The dataset is composed of 200 videos where each video contains 2 to 20 different human faces. The complete dataset consist of 657 individuals out of which 100 (15.22%) are labeled and rest 557 are unknown. The videos are retrieved from public sources like YouTube. Each video ranges from 4-14 seconds at 30fps and are at varied resolutions. This is primarily used as a validation set to mimic real world scenarios. The videos are directly given as input to the system without any prepossessing for validation.

The age-invariant face recognition task on these videos is an extremely challenging problem as these videos can be affected by illumination changes, shadows, poor resolution, and incessant motion blurring common in most videos recorded at low fps. In addition to these, the hu-

Table 1. Comparison of different cross-age datasets

Dataset	LAG [3]	**Modified LAG**	**CVA**	CACD [5]	FG-NET [1]
Images	3,828	5,045	27k	163k	1k
Subjects	1,010	1,010	100	2k	82
Noise-free	Yes	Yes	No	Yes	Yes
Frontal Faces only	Yes	No	No	Yes	Yes
Blurred Faces present	No	Yes	Yes	No	No

man faces can appear at different locations, orientations, and scales. The proposed dataset includes images with the above-detailed variations as these can significantly alter the appearance of the humans and affect the performance of the surveillance systems. Such dataset enables a largely fair evaluation of the system. The SIHN network, when trained on the VGGFace2 and Modified LAG (MLAG) dataset, can learn to recognize humans despite these variations.

4. Proposed System

This section introduces the age-invariant recognition system which first uses MTCNN to detect faces in video frame followed by the proposed ScatterNet Inception Hybrid Network (SIHN) to extract age-invariant features. The system uses cloud computation to achieve the identification in real-time. Each part of the age-invariant recognition system is explained in the following sub-sections.

4.1. Face Detection and Alignment using MTCNN

We use MTCNN to detect faces in a given video frame. MTCNN [47] is a deep-cascaded framework which inherits the correlation between face detection and alignment to boost up their performance. It produces state-of-the-art performance with the use of a Proposal Network (P-Net), a Refine Network (R-Net), and O-Net which is similar to R-Net. The detected faces are used by the SIHN network to learn high level age-invariant representations.

4.2. ScatterNet Inception Hybrid Network (SIHN)

The SIHN is used to encode the faces detected, using MTCNN detailed above, into age-invariant high-level representation. This section details the proposed ScatterNet Inception Hybrid Network (SIHN), inspired from Singh et al.'s work in [35, 36, 32, 37], composed by combining the hand-crafted (front-end) two-layer parametric log ScatterNet [34] with the Inception ResNet (IR) (back-end) shown in Fig. 2. The ScatterNet accelerates the learning of the SIHN network by extracting invariant edge-based features which allow the SIHN network to learn complex features from the start of the learning [35]. The ScatterNet (front-end) part of the proposed SIHN network are presented below.

ScatterNet (front-end): The parametric log based DTCWT ScatterNet [34] is an improved numerous version of the hand-crafted multi-layer Scattering Networks [33, 24] proposed over the years. The parametric log ScatterNet extracts relatively symmetric translation invariant representations using the *dual-tree complex wavelet transform* (DTCWT) [29] and parametric log transformation layer. The ScatterNet features are denser over scale as they are extracted from multi-resolution images at 1.5 times and twice the size of the input image. Below we present the formulation of the parametric DTCWT ScatterNet for a single input image which may then be applied to each of the multi-resolution images.

The parametric log ScatterNet is a hand-crafted two-layer network which extracts translation invariant feature representation from an input image or signal. The invariant features are obtained at the first layer by filtering the input signal x with dual-tree complex wavelets (better than cosine transforms [15]) $\psi_{j,r}$ at different scales (j) and six predefined orientations (r) fixed to $15°, 45°, 75°, 105°, 135°$ and $165°$. To build a more translation invariant representation, a point-wise L_2 non-linearity (complex modulus) is applied to the real and imaginary part of the filtered signal:

$$U[\lambda_{m=1}] = |x \star \psi_{\lambda_1}| = \sqrt{|x \star \psi^a_{\lambda_1}|^2 + |x \star \psi^b_{\lambda_1}|^2} \quad (1)$$

The parametric log transformation layer is then applied to all the oriented representations extracted at the first scale $j = 1$ with a parameter $k_{j=1}$, to reduce the effect of outliers by introducing relative symmetry of pdf [34], as shown below:

$$U1[j] = \log(U[j] + k_j), \quad U[j] = |x \star \psi_j|, \quad (2)$$

Next, a local average is computed on the envelope $|U1[\lambda_{m=1}]|$ that aggregates the coefficients to build the desired translation-invariant representation:

$$S_1[\lambda_{m=1}] = |U1[\lambda_{m=1}]| \star \phi_{2^J} \quad (3)$$

The high frequency components lost due to smoothing are retrieved by cascaded wavelet filtering performed at the second layer. Translation invarinace is introduced in these features by applying the L2 non-linearity with averaging as explained above for the first layer [34].

The scattering coefficients at L0, L1, and L2 are:

$$S = \big(x \star \phi_{2^J}, S_1[\lambda_{m=1}], S_2[\lambda_{m=1}, \lambda_{m=2}] \star \phi_{2^J}\big) \quad (4)$$

The rotation and scale invariance are next obtained by filtering jointly across the position (u), rotation (θ) and scale(j) variables as detailed in [31].

The features extracted from L0, L1, and L2 are concatenated and given as input to the Inception ResNet (IR), to learn high-level features for age-invariant face recognition. The ScatterNet features help the proposed SIHN to converge faster as the convolutional layers of the Inception network can learn more complex patterns from the start of learning as it is not necessary to wait for the first layer to learn invariant edges as the ScatterNet already extracts them.

The ScatterNet Inception Hybrid Network (SIHN) is trained on the modified LAG dataset using triplets loss. A triplet consists of one anchor image, one matching image to anchor and one non-matching image to anchor [40]. A triplet loss function is designed to minimize the the distance value between the anchor and the match which have the same identity and maximize the distance value between anchor and the non-matching image which are different identities. The network generates an embedding of 512-dimensions which capture the age-invariant facial features.

4.3. Face Comparison

The faces in the queue frames obtained by decomposing the test video are compared with the database of child images. The matching is performed by extracting the embedding of the images in the databases as well the faces extracted from the test video. The aim is to find a embedding for one of the faces in the test video which matches closes to one of the faces in the database. These embeddings are compared to the stored embeddings of all childhood faces using Euclidean (L2) distance. A distance sufficiently below a distance threshold denotes a potential match further verified by the three false positive removal filters given above. This approach is inspired by the FaceNet Unified Embedding [28] system.

4.4. Uncertainty Estimates

The uncertainty estimates are generated on the embeddings produced by the network using dropout at test time as proposed by Yarin Gal et al. [9]. We generate 50 embeddings at test time for a single test image with dropout enabled. These embeddings are used to perform face comparison. This is particularly important in this application as in the majority of the instances, the childhood pictures in the database don't have a match in the current video frame. In such cases, it is crucial to know the confidence of the closest recognition. Low confidence recognition is simply ignored.

4.5. Elimination of False Positives

We propose a new algorithm for eliminating the false positives to improve the efficiency and performance of the system. False positives refer to normal data being falsely judged as alerts, which, to a certain degree, reduces the percentage of authentic predictions by the system. Due to various unpredictable factors, in real life scenarios it is evident that the system gets perplexed while making predictions for a frame or two.

To obviate such false positives, an inter-frame comparison process is invoked which works in parallel with the prognosis algorithm. The inputs, videos from the CCTV cameras are split into a certain number of frames and then, a frame by frame analysis is done on the target individual. The system identifies the target only when it is positively detected in three consecutive frames. This makes it more certain for us to believe that the target individual is indeed caught in the video. Certainty measurements are also used as a threshold to eliminate additional false positives.

In detection with near-threshold L2 distances, additional checks are necessary. A gender identification script is invoked and it is ensured that the gender of detected face is same as gender of the younger face in database to which it was matched.

5. Experimental Results

This section presents the results of the experiments performed on the introduced datasets using the proposed SIHN network for age-invariant face recognition. The system uses the MTCNN first to detect and align faces in the current frame. The detected, cropped and aligned faces are passed through SIHN to construct a 512 dimensional vector representation for each face and finally the proximity of these vectors is compared with the target's vector to identify the person of interest in the video frame. The consequent sections details the performance of each part of the system. We also compare this system with the state-of-the-art with regards to performance of age-invariant face recognition.

5.1. MTCNN for Face Detection

The faces in the frames are detected using the MTCNN. MTCNN is composed of three sub-networks namely: P-Net, R-Net and O-Net as detailed in Section 4.1. P-Net, R-Net and O-Net takes a 12x12, 24x24 and 48x48 image as input respectively and outputs a matrix showing whether or not a there is a face. And if there is, the coordinates of the bounding boxes and facial landmarks for each face are presented. The dataset WIDER-FACE [46] was used to train all of these three networks. WIDER-FACE has 32,203 images with 393,703 faces in different situations and locations. This dataset was pre-processed for the respective networks of MTCNN. Training parameter set were: Adam Optimizer, learning rate of 0.001, weight decay 0.0005, batch size 256, iteration count 135,000 (31.5 epochs). The detection performance of the MTCNN network is presented in Table 2 below:

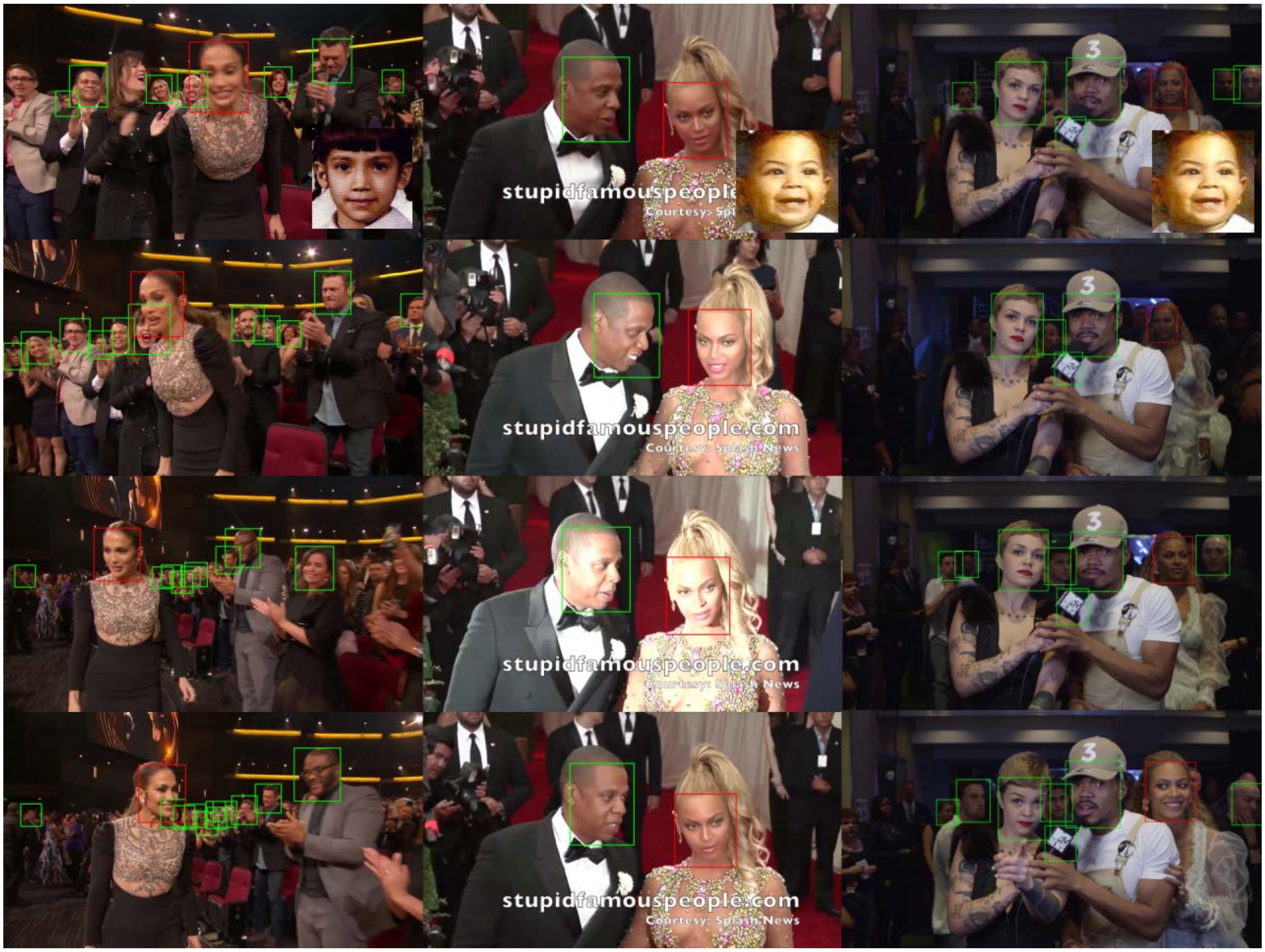

Figure 3. Still frames from CVA showing red box around the detected faces under different variations by matching from child photo (inset). These frames are extracted from 3 videos of CVA, processed using the proposed system. 1st column show the robustness against orientation and blurriness. 2nd column against extremely variant illumination conditions. 3rd column against low illumination and varying scale.

Table 2. Accuracy of P-Net, R-Net and O-Net

N/W.	12-Net	24-Net	48-Net
MTCNN	**94.24%**	**95.1%**	**95.1%**

5.2. Fine tuning on Modified LAG (MLAG) Dataset

This section details the training of Inception ResNet model on the modified LAG dataset. An Inception ResNet model pretrained on the VGGFace2 [4] is used for this task.

VGGFace2 [4] is a popular large scale image dataset for face recognition. It contains 3.31 million images of 9131 subjects having an average of 362.6 images per subject. Images are retrieved from Google Image Search and have large variations in illumination, pose, age, ethnicity and profession. Identity overlap with LFW [13] were not removed and only the training set was used. A pre-trained model using VGGFace2 was used as a base model characterizing the general face modalities. The accuracy of this pre-trained model is 99.65%

A pre-trained Inception ResNet model trained on VGGFace2 is loaded. A set of layers representing high level features are chosen by trial and error to be retrained. The Modified LAG dataset is pre-aligned using MTCNN and cropped to standardized sizes. The chosen layers are retrained using these face images from LAG. A subset of still frames extracted from CVA dataset are used to perform unit tests on the resultant model. The extent of fine-tuning is decided by observing the performance of intermediate models in the unit test.

The Inception ResNet-v1 model contains a total of 490 trainable variables. From these last 95 variables responsible for high level features are being trained and rest are frozen. Following are the training parameters: imageSize 160x160, optimizer RMSprop, learningRate 0.01, weightDecay 0.0001, number of epochs 2, embedding_size 512, gpuFraction 0.5

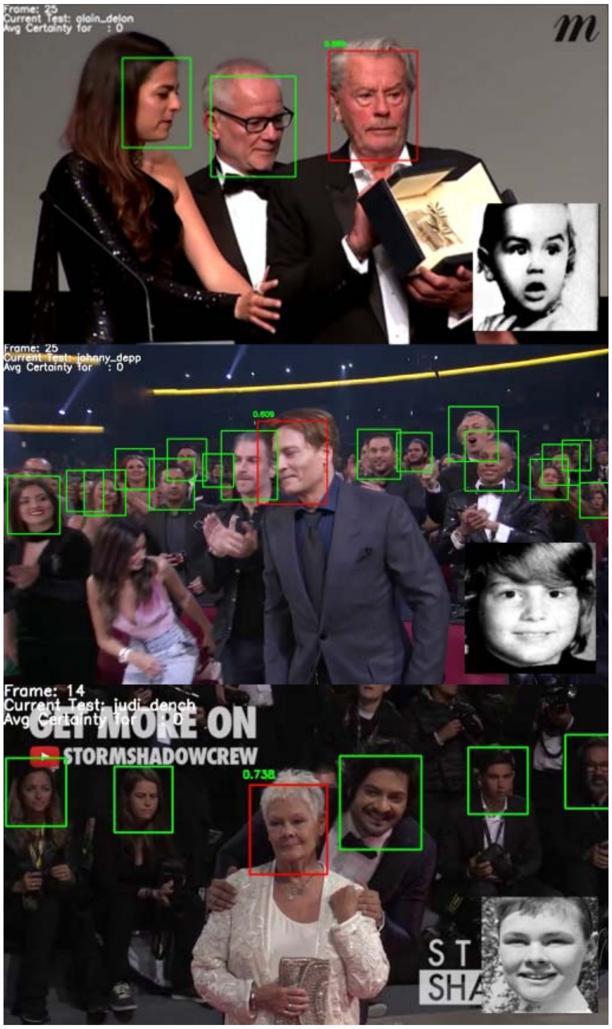

Figure 4. Examples of large age gap recognition. The childhood image (bottom right) of the celebrity is used to find a match from all faces in the video frame

5.3. Experiments on Datasets

The performance of the proposed system is measured on two standard datasets (CACD-VS and LFW) and one custom dataset (CVA) presented in this paper.

The performance on the proposed datasets is presented with True Positive (TP), False Positive (FP), True Negative(TN) and False Negative (FN) measures as well as the final classification accuracy. As for the standard datasets, on only the final classification accuracy is presented.

5.3.1 Experiments on CVA Dataset

CVA dataset contains a total of 200 videos (27k frames) with 100 labeled individuals and 557 unlabeled. Each frame may contain 2 to 20 different faces. For better accuracy analysis on CVA dataset, we chose to perform manual counting on a randomly chosen subset of 540 frames from CVA dataset. The frames in FP and FN sets are not mutually exclusive (contain repeated frames).

Table 3. Confusion Matrix

**	Predicted NO	Predicted YES
Actual NO	TN=**92**	FP=**37**
Actual YES	FN=**44**	TP=**378**

Calculations made from Table 3: Total = 551 (includes repeated frames)

Accuracy 0.8529; Mis-classification 0.147; Precision 0.911; Sensitivity 0.896; Specificity 0.713

Table 4. Comparison of different methods on CVA

Method	Accuracy
CARC [5]	74.23%
LF-CNNs [44]	84.19%
SIHN	**85.29%**

5.3.2 Experiments on CACD-VS

CACD [5] dataset consists of 163,446 images of 2,000 individuals. The ages of the images range from 10 years to 62 years. The dataset presents these images in different illumination, orientation and scales. CACD-VS is a verification subset of CACD such that there are 2,000 pairs of positive samples and 2,000 pairs of negative samples to give a total of 4,000 image pairs. We pass these through MTCNN and then the SIHN system. We then compare the prediction with actual.

The results are shown in Table 5. We can observe that SIHN significantly outperforms other methods. It is also seen that SIHN is better than human performance, further reinforcing the robustness of this work.

Table 5. Comparison of different methods on CACD-VS

Method	Accuracy
High-Dimensional LBP [7]	81.6%
HFA [11]	84.4%
CARC [5]	87.6%
LF-CNNs [44]	98.5%
Human, Average [6]	85.7%
Human, Voting [6]	94.2%
SIHN	**96.2%**

5.3.3 Experiments on LFW

LFW [13] is a popular benchmark for general face recognition. It has 13,233 images of 5,749 individuals obtained in-the-wild (unconstrained). The model is tested on 4,000 face pairs randomly selected from LFW. We follow the ”Unrestricted, Labeled Outside Data” protocol of LFW. This test set is disjoint from the training set. The results given in Table 6 show that in addition to cross-age recognition, SIHN is also as effective in general face recognition.

Table 6. Comparison of general methods on LFW

Method	Images	Accuracy
DeepFace [41]	4M	97.35%
FaceNet [28]	200M	99.65%
DeepID2+ [39]	-	99.47%
Center-Loss [45]	0.7M	99.28%
SphereFace [20]	0.5M	99.42%
SIHN	3.3M	**99.36%**

5.4. Runtime Performance

The runtime performance of age-invariant face recognition system was computed on cloud. It consists of three parts: i) Face detection using MTCNN, ii) Obtaining embedding using SIHN iii) Calculation of L2 distances and false positive removal. The server was equipped with Intel Xeon family CPU and 1x NVIDIA Tesla GPU. The deep learning framework used was Tensorflow, accelerated using cuDNN framework. The system performs age-invariant face recognition on videos at a speed of 12fps to 18fps depending on number of faces in frame and other factors like system load. In comparison, under similar conditions, LF-CNNs operates at 6-7 fps.

6. Conclusion

The paper proposed the Real-time Age-Invariant Face Recognition System for videos. The system first uses the MTCNN network to detect faces in the input video which are then used by the SIHN to extract the embeddings. Similar embeddings are obtained for the images in the child database. The aim was to find an embedding for one of the faces in the test video which matches closes to one of the faces in the database. The proposed system is able to solve this task effectively. The proposed system outperforms the state-of-the-art technique on the cross-age dataset CACD-VS as well as gives similar performance on general face recognition dataset LFW when compared with popular general face recognition frameworks.

The proposed SIHN network uses ScatterNet features which allows them to learn useful representations rapidly using relatively fewer labelled examples. The utilization of fewer labelled examples for age-invariant recognition is beneficial as it is expensive to collect annotated video examples.

The paper also introduced the Celebrities Video Aging Dataset which can be used by other researchers to use deep learning for aerial surveillance applications. This system is highly application oriented and deployable in real-world scenarios.

References

[1] Fg-net aging database. https://web.archive.org/web/20070217193535/http://www.fgnet.rsunit.com/, 2004.

[2] S. Bachhety, R. Singhal, K. Rawat, K. Joshi, and R. Jain. Crime detection using text recognition and face recognition.

[3] S. Bianco. Large age-gap face verification by feature injection in deep networks. *Pattern Recognition Letters*, 90:36–42, 2017.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[5] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV*, pages 768–783. Springer, 2014.

[6] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015.

[7] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3025–3032, 2013.

[8] L. Du and H. Ling. Cross-age face verification by coordinating with cross-face age verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2329–2338, 2015.

[9] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[10] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 29(12):2234–2240, 2007.

[11] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *Proceedings of the IEEE ICCV*, pages 2872–2879, 2013.

[12] D. Gong, Z. Li, D. Tao, J. Liu, and X. Li. A maximum entropy feature descriptor for age invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5289–5297, 2015.

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[14] S. Ioffe. Probabilistic linear discriminant analysis. In *European Conference on Computer Vision*, pages 531–542. Springer, 2006.

[15] V. Jeengar, S. Omkar, A. Singh, M. K. Yadav, and S. Keshri. A review comparison of wavelet and cosine image transforms. *International Journal of Image, Graphics and Signal Processing*, 4(11):16, 2012.

[16] K. Kim. Intelligent immigration control system by using passport recognition and face verification. In *International Symposium on Neural Networks*, pages 147–156. Springer, 2005.

[17] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[18] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *IEEE transactions on information forensics and security*, 6(3):1028–1037, 2011.

[19] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs. Face verification across age progression using discriminative methods. *IEEE Transactions on Information Forensics and security*, 5(1):82–91, 2010.

[20] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[21] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):331–345, 2014.

[22] T. K. Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[23] Y. M. Mustafah et al. An automated face recognition system for intelligence surveillance: Smart camera recognizing faces in the crowd. In *Distributed Smart Cameras, 2007. ICDSC'07. First ACM/IEEE International Conference on*, pages 147–152. IEEE, 2007.

[24] S. Nadella, A. Singh, and S. Omkar. Aerial scene understanding using deep wavelet scattering network and conditional random field. In *ECCV*, pages 205–214, 2016.

[25] C. Otto, H. Han, and A. Jain. How does aging affect facial components? In *ECCV*, pages 189–198. Springer, 2012.

[26] U. Park, Y. Tong, and A. K. Jain. Age-invariant face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):947–954, 2010.

[27] S. V. Peri and A. Dhall. Disguisenet: A contrastive approach for disguised face verification in the wild. In *CVPR Workshop on Disguised Faces in the Wild*, volume 4, 2018.

[28] 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*, pages 815–823, 2015.

[29] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury. The dual-tree complex wavelet transform. *IEEE signal processing magazine*, 22(6):123–151, 2005.

[30] S. Siddiqui, M. Vatsa, and R. Singh. Face recognition for newborns, toddlers, and pre-school children: A deep learning approach.

[31] L. Sifre and S. Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *CVPR, 2013*, pages 1233–1240, 2013.

[32] A. Singh, D. Hazarika, and A. Bhattacharya. Texture and structure incorporated scatternet hybrid deep learning network (ts-shdl) for brain matter segmentation. *ICCV Workshop*, 2017.

[33] A. Singh and N. Kingsbury. Multi-resolution dual-tree wavelet scattering network for signal classification. In *International Conference on Mathematics in Signal Processing*, 2016.

[34] A. Singh and N. Kingsbury. Dual-tree wavelet scattering network with parametric log transformation for object classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[35] A. Singh and N. Kingsbury. Efficient convolutional network learning using parametric log based dual-tree wavelet scatternet. *IEEE ICCV Workshop*, 2017.

[36] A. Singh and N. Kingsbury. Scatternet hybrid deep learning (shdl) network for object classification. *International Workshop on Machine Learning for Signal Processing*, 2017.

[37] A. Singh and N. Kingsbury. Generative scatternet hybrid deep learning (g-shdl) network with structural priors for semantic image segmentation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[38] A. Singh, D. Patil, G. M. Reddy, and S. Omkar. Disguised face identification (dfi) with facial keypoints using spatial fusion convolutional network. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, pages 1648–1655. IEEE, 2017.

[39] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.

[40] C. Szegedy et al. Going deeper with convolutions, corr abs/1409.4842. *URL http://arxiv. org/abs/1409.4842*, 2014.

[41] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE CVPR*, pages 1701–1708, 2014.

[42] X. Wang, M. S. Gerber, and D. E. Brown. Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer, 2012.

[43] Y. Wang et al. Orthogonal deep features decomposition for age-invariant face recognition. In *ECCV*, volume 3, page 7, 2018.

[44] Y. Wen, Z. Li, and Y. Qiao. Latent factor guided convolutional neural networks for age-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4893–4901, 2016.

[45] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[46] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[47] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.