ChildrEN SafEty and Rescue (CENSER) System for Trafficked Children from Brothels in India

Raghu Vamshi Hemadri[1], **Amarjot Singh**[1], **Ajeet Singh**[2]
[1] Skylark Labs
[2] Guria, India
vamshi.hemadri@skylarklabs.ai, amarjot@skylarklabs.ai, director@guriaindia.org

Abstract

Human child trafficking has become a global epidemic with over 10 million children forced into labor or prostitution. In this paper, we propose the ChildrEN SafEty and Rescue (CENSER) system used by the Guria non-profit organization to retrieve trafficked children from brothels in India. The CENSER system is formed of the proposed Memory Augmented ScatterNet ResNet Hybrid (MSRHN) network trained on three databases containing images of trafficked children at different ages, their kins, and their sketches. The CENSER system encodes the input image of a child using the proposed Memory Augmented ScatterNet ResNet Hybrid (MSRHN) network and queries the encoding with the (i) Age, (ii) Kinship, and (iii) Sketch databases to establish the child's identity. The CENSER system can also predict if a child is a minor, which is used along with their identity to convince law enforcement to initiate the rescue operation. The MSRHN network is pre-trained on the KinFace database and then fine-tuned on the three databases. The performance of the proposed model is compared with several state-of-the-art methods.

Introduction

Child trafficking violates the core human rights of children under the age of 18 by forcing them into forced labor, under-age marriage, prostitution etc (General-Assembly 1989). According to the United Nations Children's Fund (UNICEF), nearly 2 million children with an average age range from 11 to 14 are subjected to prostitution in the global sex trade. The percentage of children (between the ages of 2 and 18) being trafficked has risen by about 25% from 2009 to 2012. (**?**).

In India alone, $20,000$ women and children were victims of human trafficking in 2016. This was a nearly 25 percent more over the previous year [1]. The latest available statistics on abduction in India indicates it is a growing crime, with the reported abductions of children rising from $15,284$ in 2011 to $41,893$ in 2015 (Guardian 2017).

Different states of India have suffered the brunt of this epidemic. According to the statistics provided by the Punjab Police to Dawn newspaper, a total of 767 children have been reported missing in 2016 from which 722 have been recovered (Guardian 2017). According to Kolkata's *Child in Need Institute*, $1,628$ trafficked children, in the age group of 4 to 15 years, were retrieved from a single railway station; among these, 134 were girls and the youngest was only four years old (Guardian 2017). Of course, these are official statistics and do not necessarily reflect the true numbers of child trafficking in a population of around 1.2 billion in India.

To trace missing children, face recognition is perhaps the primary biometric modality since parents and relatives are more likely to have a lost child's photographs(s) as opposed to, say, fingerprint or iris. Face recognition is an impressive and useful tool which effectively functions under the conditions of facial pose, illumination, and expression (Grother, Ngan, and Hanaoka 2017), (Flanagan 2017), (Huang et al. 2007), (Liao et al. 2014) but fails for aged faces (Yoon and Jain 2015) for fingerprint study and Grother (Grother et al. 2013) for iris.

In this work, we present the ChildrEN SafEty and Rescue (CENSER) System used by the Guria (Guria NGO) non-profit organization to retrieve trafficked children from brothels located at Varanasi in Uttar Pradesh, India. Volunteers of the non-profit visit the brothels as customers while wearing hidden cameras to record the children's faces. The recorded video is given as input to the CENSER system to establish a match with the missing children's database. Since these children's facial features may have changed significantly since being trafficked, the CENSER system can perform age-invariant face recognition. The system also predicts the children's age, which is further used to establish if they are a minor. This information is finally used to convince the local law enforcement agencies to raid the brothels leading to these children's rescue. The system is also used as a stand-alone mobile application to establish children's identity found abandoned at railway stations or bus stops. The Children's identity is used to rescue the children from railway stations and bus stops, whereas in brothels identifying the child as a minor is sufficient to convince law enforcement to initiate the rescue operation.

The missing child's picture is often not available at the age he/she was abducted as many families never took a photograph due to their weak financial condition. For those cases, the CENSER system can establish the identity of children

[1] According to the Ministry of Women and Child Development, Government of India, 2016

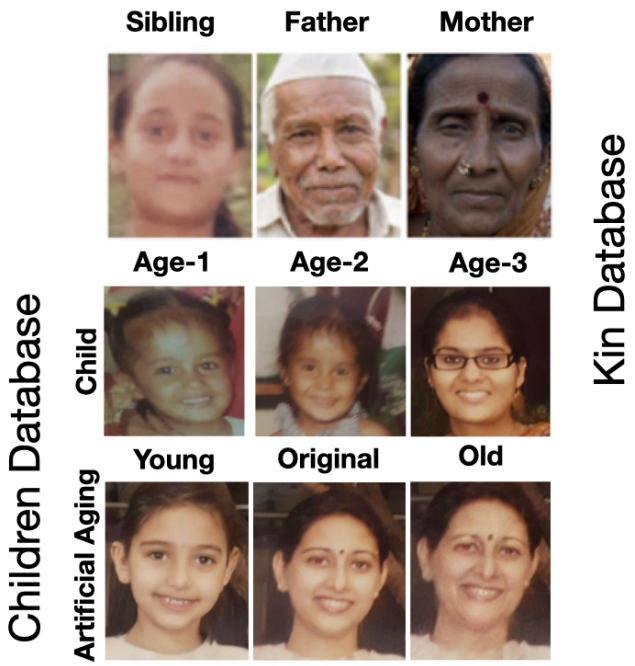

Figure 1: Figure illustrates image examples for children at different ages both from the child and the kin database.

using kinship verification. We also use sketches of the children made by the parents at the time of the abduction to aid the system's recognition performance further.

The CENSER system uses the proposed Memory Augmented ScatterNet ResNet Hybrid (MSRHN) Network to learn high-level features to perform age-invariant classification for all three age, kin, and sketch cases. The ScatterNet architecture allows the network to learn useful features rapidly from relatively fewer labeled examples (Singh and Kingsbury 2017a). The network is trained and evaluated on three datasets of trafficked children containing age, kin, and sketch images provided by the Guria Non-Profit. The dataset contains faces recorded at different locations, scales, rotations, illumination variations and ages.

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.

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 CENSER dataset. Section 4 describes the proposed CENSER System, Section 5 details the experimental results and Section 6 concludes this research.

Figure 2: Figure illustrates original and artificially aged sketches for children which forms the sketch database.

Related Work

Several attempts have been made in the past to develop an algorithm that can solve the individual tasks of (i) age-invariant face recognition (ii) kin recognition (iii) recognition using sketch images. We detail the existing literature on each of aforementioned tasks in this section.

Many methodologies that can perform age-invariant face recognition have been proposed in recent years (Gong et al. 2015; Li, Park, and Jain 2011; Park, Tong, and Jain 2010; Gong et al. 2013; Chen, Chen, and Hsu 2014a). These approaches can be broadly divided into generative and discriminative approaches.

Generative approaches (Geng, Zhou, and Smith-Miles 2007; Lanitis, Taylor, and Cootes 2002; Park, Tong, and Jain 2010; Gong et al. 2013), construct 2D or 3D generative models to compensate for the aging process and synthesize facial 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 (Li, Park, and Jain 2011; Chen, Chen, and Hsu 2014a; Ling et al. 2010; Otto, Han, and Jain 2012; Du and Ling 2015) which use robust facial features and discriminative learning methods to reduce the gap between face images captured at different ages. These methods incrementally adapt the facial features to derive the latent space suitable for the age-invariant face recognition. Y. Wen (Wen, Li, and Qiao 2016) achieves an accuracy of 84.19% on CVA dataset consisting of 27k frames having 2 to 20 different faces, and 98.5% on CACD-VS(Chen, Chen, and Hsu 2014b) dataset consisting of 163,446 images of 2,000 individuals. Best-Rowden studied face recognition performance of newborns, infants, and toddlers (ages 0 to 4 years) on 314 subjects acquired over a maximum time lapse of only one year (Best-Rowden, Hoole, and Jain 2016). Their results show that state-of-the-art face recognition technology has a very low True Accept Rate (TAR) of 47.93% at 0.1%

The most recent kin verification approaches used deep networks to extract relevant high-level features to solve this task (Zhang et al. 2015; Dehghan et al. 2014; Xiong et al. 2016). Zhang (Zhang et al. 2015) first attempted to find

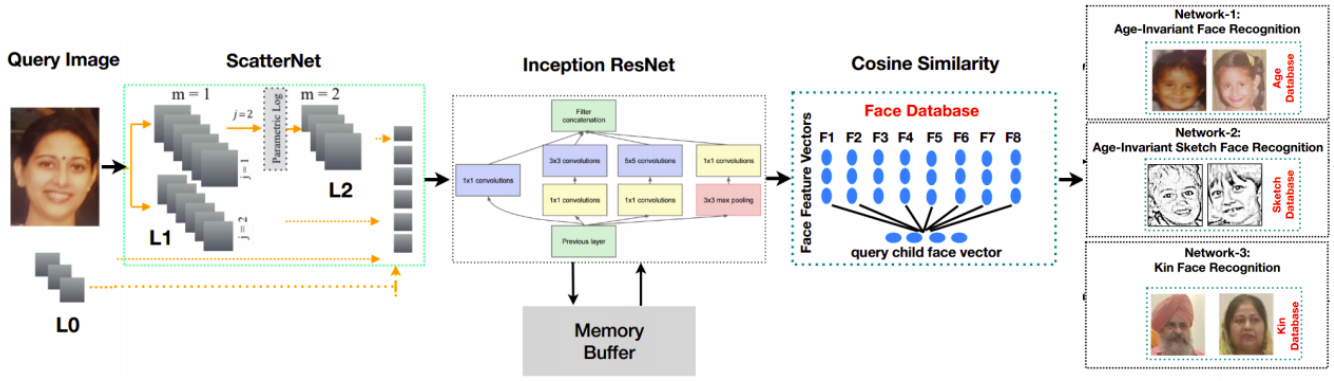

Figure 3: Pipeline of proposed CENSER system. The MTCNN is used to detect all the faces in input video frame to get the bounding boxes. The cropped faces with the target childhood image are sent to the MSRHN for embedding generation. The low level features extracted by ScatterNet at L0, L1, and L2 are 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.

the kinship relation using deep convolution neural networks. They extracted high-level features using a deep convolutional neural network with a softmax classifier at the end to identify the kinship relation between two persons. Dehghan (Dehghan et al. 2014) used a gated autoencoder with a discriminative neural network layer to learn the genetic features. The features and metrics discovered from the gated autoencoder are fused to identify the kinship relationship. Zhang (Zhang et al. 2015) in their work achieves an accuracy of 77.5% on the KinFace Wild 1 dataset(Lu et al. 2014, 2012), which has 533 pairs of child-parent images and an accuracy of 88.4% on the KinFace Wild 2 dataset(Lu et al. 2014, 2012) which has 1000 pairs of child-parent images.

Deep learning methods have been recently employed for the task of the sketch-photo recognition problem by learning the relationship between the two modalities. Recent approaches on sketch recognition problem have mainly focused on closing the gap between the two domains of sketch and photo and use of soft biometrics has not been investigated adequately. In (Klare et al. 2014) an approach was proposed to directly use facial attributes in suspect identification without using the sketch. Mittal et. al. (Mittal et al. 2017) fused multiple sketches of a suspect to increase the accuracy of their algorithm. They also employed some soft biometric traits such as gender, ethnicity, and skin color to reorder the ranked list of the suspects. To avoid this, many deep techniques utilize relatively shallow model or train the network only on the photo modality (Mittal, Vatsa, and Singh 2015). Ouyang et al. (Ouyang et al. 2014) achieves an accuracy of 65.19% on forensic dataset having 196 pairs of face and sketch pairs and an accuracy of 69.15% on Caricature dataset (Klare et al. 2012) having 207 pairs.

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 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.

Dataset	Images	Subjects	Blurred Faces	Water Damage	Poor Illumination
Age	1,076	358	103	56	68
Kin	705	305	–	–	–
Sketch	1,076	358	–	–	–

Table 1: The table presents the details of the (i) Age (ii) Kin, and (iii) Sketch sub-datasets.

CENSER Dataset

This work uses the real-world data obtained by scanning the manual records of 300 children between the ages of 5 and 11. The aim is to collect images at multiple ages for a child to construct the deep model which can perform age-invariant face recognition. There is only one image available for most cases and a handful of cases contains images at multiple ages. In such cases, Progressive Face Aging with Generative Adversarial Network (Huang et al. 2021) in used to artificially age the image at a different age to produce images at different ages using the available single image. The qualitative comparison of artificially aged images is shown in Fig. 2. In addition, the visual quality of several images is degraded by water and resolution, and are visually depicted in Fig. 1.

Several families never clicked the picture of their chil-

dren due to the lack of financial resources. We generated the sketch images of the child with the help of their parents. These images are used by CENSER to learn face sketch recognition.

We also obtained the image of the parents and the sibling which are used by the CENSER to learn kinship analysis and augment the sketch analysis module. The CENSER dataset has 3 sub datasets namely Age, Kin and Shetch datasets. The CENSER dataset has a total of 358 subjects(child image). Out of which only 305 subjects have either of father or mother image available. And the subject images at different ages and their sketches are available for all the subjects. The number of images available for each sub-dataset and further details are presented in Table 1. The Kin sub-dataset have these four relations: Mother-Daughter(M-D), Mother-Son(M-S), Father-Daughter(F-D), and Father-Son(F-S) with 127, 116, 134, and 156 pairs of kinship images of the above-mentioned pairs, respectively. The images are aligned, and also to cut out the background, the face region is cropped into 64x64 size. Some of the images of both the datasets are shown in Fig. 1

The datasets are collected in an unconstrained environment with no limitations on the pose, expression, partial occlusion, background, ethnicity, and lightening.

Proposed CENSER System

This section presents the age-invariant kinship recognition framework. The faces in the given video are first detected using MTCNN, and the Memory Augmented ScatterNet ResNet Hybrid (MSRHN) Network is then used on the extracted faces to obtain the features corresponding to the kinship relation. For realtime recognition, the proposed framework is deployed on AWS EC2 instance. The following subsections explain the age-invariant kinship recognition.

Face Detection and Alignment using MTCNN

MTCNN (Multi-task Cascaded Convolutional Neural Networks) is a deep CNN architecture consisting of three stages, namely Proposal Network (P-Net), Refine Network (R-Net) and Output Network (O-Net). MTCNN detects the bounding boxes of faces in a picture and five Face Landmarks along with a confidence score. Every stage improves the detection results by passing its input through a CNN that returns candidate bounding boxes with their scores, followed by non-max suppression (NMS).

Given a picture, it is at the start resized to different scales to create an image pyramid, that is that the input of the subsequent three-stage cascaded framework. Each stage in the pipeline is a deep CNN to extract the aforementioned features. After that, MTCNN employs NMS to merge extremely overlapped candidates. Each stage improves the results from the prior stage, finally resulting in accurate bounding box with accuracy of it being a face and facial landmarks. The complete architectural details of MTCNN are proposed in (Zhang et al. 2016).

Memory Augmented ScatterNet ResNet Hybrid (MSRHN) Network

The high-level kinship relation representational features of the faces, detected using MTCNN, are extracted using the proposed MSRHN network. The details of the proposed ScatterNet Hybrid Network, inspired from Singh et al.'s work in (Singh and Kingsbury 2017b,c; Singh, Hazarika, and Bhattacharya 2017; Singh and Kingsbury 2018), are explained in this section. This Hybrid network uses an Inception ResNet (IR) in the back-end and a handcrafted two-layered parametric log ScatterNet (Singh and Kingsbury 2017a) in the front-end, as shown in Fig. 3. The ScatterNet extracts the invariant edge features, accelerating the ScatterNet Inception Hybrid Network (SIHN) to learn the sophisticated features from the beginning. The subsection below explains the front-end ScatterNet of the proposed SIHN framework.

ScatterNet (front-end): One of the improvised versions of the handcrafted multilayered scatter network (Singh and Kingsbury 2016; Nadella, Singh, and Omkar 2016), proposed long ago, is the parametric log-based *dual-tree complex wavelet transform* (DTCWT) ScatterNet (Singh and Kingsbury 2017a) which is used to extract invariant features that are comparatively symmetric translation representations. The parametric log-based DTCWT ScatterNet uses a DTCWT (Selesnick, Baraniuk, and Kingsbury 2005), hence called DTCWT ScatterNet, followed by a log transformation layer. The features from DTCWT ScatterNet being extracted from images at higher resolutions, approximately 1.5 or 2 times the input image, are dense over the range. The framework of DTCWT ScatterNet that takes a single image as input is described below. The same framework described below is applied to each of the high-resolution images.

Already mentioned above, the parametric log-based DTCWT Scatter-Net is used to extract invariant features that are comparatively symmetric translation representations for an input signal or the input image. The dual-tree complex wavelet transform $\psi_{j,r}$, known to be better than cosine transform (Jeengar et al. 2012), extracts the invariant features from the input image or the input signal in the first stage at different ranges (j) and six pre-defined orientations (r) fixed to $15^\circ, 45^\circ, 75^\circ, 105^\circ, 135^\circ$ and 165°. A better representation invariant to translation is built by applying an element-wise L_2 non-linearity (complex modulus) to real and imaginary parts of signal or image filtered from DTCWT:

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

A parametric log transform layer follows the DTCWT layer. The log transformation layer introduces the relative symmetry of pdf (Singh and Kingsbury 2017a) (shown below), reducing the effect of outliers in all the oriented representations obtained from the DTCWT layer at $j = 1$ scale and parametrized by $k_{j=1}$.

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

The required representation $S_1[\lambda_{m=1}]$ which is invariant in translation is obtained by computing a local average on

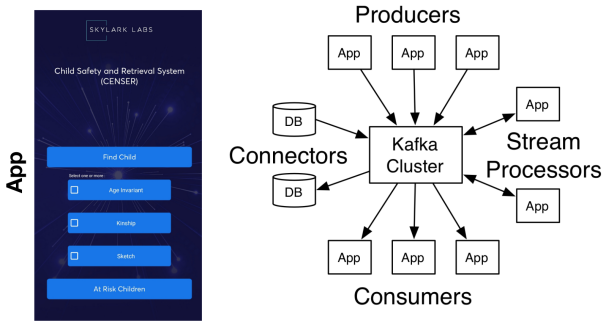

Figure 4: The illustration presents the CENSER App used by the Guria non-profit to retrieve the children. The app is used by multiple volunteers (producers) and the feed (stream) from them is processed by multiple servers using Kafka as shown in the figure.

$|U1[\lambda_{m=1}]|$ that amass the coefficients of the envelope. Furthermore, $S_1[\lambda_{m=1}]$ is given by the equation:

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

Due to smoothing done by the first layer, the high-frequency components succumb, which can be recovered by the cascaded wavelet filtering operation performed at the second layer. Furthermore, using L2 non-linearity followed by averaging like in the first layer (Singh and Kingsbury 2017a), a translational invariance is added to the features.

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

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

The invariance in scale and rotation can be obtained by joint filtering of the rotation (θ), scale(j) and the position (u) which is described in detail below: in (Sifre and Mallat 2013).

Memory Buffer The new or novel faces which are added after the model has been deployed need to be learned on the fly in a one-shot learning manner. In order to recognize these faces, the features for these newly added faces are extracted from the n^{th} layer of the ResNet network and added to the memory buffer along with the identity of these faces. When this face is seen as a query image, the features vector of the query image is matched with the features vectors in the memory buffer using cosine similarity to produce the correct face identity. The SRHN network won't produce the correct results for these novel faces as these are not learned by the network yet. Overtime if there is a certain novel face seen frequently, it is replayed to the SRHN network and consolidated after which it is removed from the memory buffer making space for newer faces to be added.

Face Comparison

MTCNN extracts the faces that are available in the test video arrangement; each of the extracted faces is compared with the faces present in the database of parent images. The CENSER system discovers a parent image from the database, by comparing using a quantitative measure. The face embeddings of the acquired images are extracted utilizing the SIHN Network. They are compared with the stored embedding of parent images in the database using the $Ł_2$ norm (Euclidean distance). The pair for which the Euclidean distance acquired is beneath a specific pre-set threshold confirms a potential match, which is besides supported utilizing three false-positive expulsion channels. The FaceNet Unified Embedding (Schroff, Kalenichenko, and Philbin 2015) framework embraces this methodology.

Uncertainty Estimates

The uncertainty is assessed by using dropout during test time, in the network, proposed by Yarin Gal et al. (Gal and Ghahramani 2016). To estimate the uncertainty, 50 embeddings are generated on a single test image with dropout used in the network. The embeddings obtained are used to find the kinship relation pair from the database, a crucial comparison for this application because kinship relation can not be identified for the faces in the video in most of the cases. In the before-mentioned instances, knowing the confidence score of the kinship relation estimate is essential; estimates with low confidence are disregarded. The confidence or certainty score is estimated by cosine similarity given by:

$$Cosine(\mathbf{e}_p, \mathbf{e}_c) = \frac{\mathbf{e}_p \cdot \mathbf{e}_v}{\|\mathbf{e}_p\| \|\mathbf{e}_c\|} \tag{5}$$

Where, $\mathbf{e}_p$ and $\mathbf{e}_c$ are parent and child image embeddings respectively.

Underage Children Detection

The CENSER system is focused to recover teenage girls from brothels. As a pre-filtering stage for the aforementioned use, we use an age and gender predictor to predict if a child is a minor. The minor girls are directly reported to the police. The CENSER system further analyses females with the age above 18 to a kin match from the database. In case a potential match is found, the kin-ship pair is reported to the police.

Elimination of False Positives

False positives allude to typical information being deceitfully inferred as cautions, which, in a specific way, diminishes the level of accurate predictions by the system. Because of different capricious elements, in actuality situations, it is apparent that the framework gets confounded while making predictions for a frame or two. Thus to improve the accuracy and ability of the framework, we propose a novel algorithm eliminating the false positives.

The input video sequences, (recording of the CCTV) are severed into frames for analysis. A parallel comparison among extracted frames, along with the prediction algorithm, eliminates the false positives in testing. The target is recognized, in the frame by frame interpretation, only if identified positively in three consecutive frames. This, more explicitly confirms that a kinship relation pair is surely identified for the person who appeared in the video. A threshold using the certainty measurements also eliminates further false positives.

App on Kafka Spark

The mobile application is used by multiple volunteers of the Non-profit at the same time. This results in several users sending multiple face recognition requests at the same time to the server. In order to effectively process these requests, they are distributed over a cluster of servers. Apache Kafka was used to manage the interaction of these servers and process the requests.

Volunteers of the Non-Profit access CENSER via mobile and web application consequently generating hundreds of requests at peek hours. A single node is not able to serve these many requests. Following major issues come into picture :-

- The request processing latency increases
- Server resources such as memory and CPU are bottlenecked due to increased number of pending requests in memory
- The user experience on our applications degrades causing impact on daily operations of the Non-Profit.

To solve this problem, we designed a Distributed Processing Stack using Kafka and Spark. Kafka is a distributed message broker which can scale horizontally to multiple nodes. It provides a Consumer-Producer pattern. When the web interface receives a request from our applications it sends a message to a Kafka topic which is nothing but a broker queue. Here, the web interface acts as a producer.

Spark is a distributed processing framework that supports Kafka integration. Spark master can ingest streaming messages from a Kafka topic, acting as a consumer. When Spark receives a batch of messages it distributes them onto the worker nodes which perform the actual processing. The results are combined by the master and sent back to another Kafka topic which essentially acts as a message sink.

Experimental Results

This section presents the results of the experiments performed on the CENSER datasets using the proposed MSRHN network for age-invariant, Kinship, and sketch face recognition. The CENSER system first uses the MTCNN detector to detect, crop, and align faces in a frame. These faces are given as input to the MSRHN network to extract a 512-dimensional vector representation for each face. These vector representations are compared using cosine similarity with the age-invariant, kin, and sketch datasets to find the appropriate match. The consequent sections detail the performance of each part of the system. We also compare the performance of three components of the system with the state-of-the-art results.

MTCNN for Face Detection

MTCNN's three sub-networks namely: P-Net, R-Net, and O-Net (detailed in Section 4.1) take input with size 12x12, 24x24, and 48x48 respectively to extract the faces. The MTCNN was trained on the WIDER-FACE (Yang et al. 2016) datasets with 32,203 images and 393,703 faces in different situations and locations. The detector was trained with the following parameters: 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 above 96%.

Training Details

The detailed training procedure of the Inception ResNet V1 model is presented on the CENSER dataset in this section. The Inception ResNet v1 model is pretrained on VGGFace2 (Cao et al. 2018) for better extraction of features.

The Inception ResNet (IR) extracts high-level features of age invariant kinship relation, taking the vector obtained by concatenating each feature of L0, L1, and L2 layers of the scatternet as input. The convolutional layers of SIHN thus learn high-level features from the beginning of learning, leading to faster convergence (Singh and Kingsbury 2016).

The proposed Memory Augmented ScatterNet ResNet Hybrid (MSRHN) Network is trained on the CENSER dataset by combining kinship pairs of both KinFaceW-I and KinFaceW-II datasets using triplet loss. The SIHN trained over the KinFace dataset is fine-tuned over the CENSER dataset. The triplet loss (Szegedy et al. 2014) uses three images to compute the loss, namely, an anchor, a positive and a negative example. An image is chosen as the anchor, and the image related to the anchor is chosen as a positive example, and an image not related to anchor is chosen to be a negative example. The triplet loss minimizes the distance between the anchor and positive example and maximizes the distance between the anchor and negative example. The SIHN generates 512-dimensional embeddings of the input image that apprehend the age-invariant kinship relation facial features.

The Inception ResNet V1 has in total of 490 trainable variables, which are trained for the first three epochs. By trail and error, it is found that the last 95 variables extract high-level features from the faces. For the rest of the epochs, the first 395 variables are frozen, and the last 95 variables are fine-tuned with the Image Size of 160x160 having a Face Margin of 5 pixels. The network is trained using RMSProp optimizer at a learning rate of 0.01 and 0.0001 weight decay. The batch has 45 negative pairs with two images per person.

Experiments on Datasets

Dataset	True Negative	False Negative	False Positive	True Positive
CENSER (Kinship)	92	44	37	378
KinFace Wild 1	69	27	21	416
KinFace Wild 2	78	53	41	828

Table 2: The table presents the confusion matrix of MSRHN method's kinship recognition performance on CENSER, KinFace Wild 1 and KinFace Wild 2 datasets

The performance of the proposed system is measured on the age-invariant, Kin, and Sketch CENSER dataset presented in this paper. The performance on the CENSER dataset is presented with True Positive (TP), False Positive (FP), True Negative(TN) and False Negative (FN) measures

Dataset	Accuracy .(%)	Mis Classification	Precision	Sensitivity	Specificity	F Measure
CENSER (Kinship)	85.3%	0.147	0.911	0.896	0.713	0.903
KinFace Wild 1	91.0%	0.090	0.952	0.939	0.767	0.945
KinFace Wild 2	90.6%	0.094	0.953	0.940	0.655	0.946

Table 3: Comparison of MSRHN method's kinship recognition performance on CENSER, KinFace Wild 1 and KinFace Wild 2 datasets

as well as the final classification accuracy. As for the standard datasets, only the final classification accuracy is presented.

CENSER 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 CENSER dataset, we chose to perform manual counting on a randomly chosen subset of 540 frames from CENSER dataset. The frames in FP and FN sets are not mutually exclusive (contain repeated frames). For better evaluation of the proposed system, it is also tested on KinFace Wild 1 and KinFace Wild 2 datasets. 20% of the subjects from each of the KinFace Wild datasets are used for testing. The True Positive (TP), False Positive (FP), True Negative(TN) and False Negative (FN) measures on all 3 datasets are provided in table 2. For a better evaluation of the system on these datasets the accuracy, mis-classification ratio, precision, specificity and F measure are provided in table 3. This MSRHN model is compared to other techniques on CENSER dataset with their accuracy for kinship analysis present in table 4.

Method	Accuracy
CARC (Chen, Chen, and Hsu 2014a)	83.47%
LF-CNNs (Wen, Li, and Qiao 2016)	88.14%
MSRHN	**93.175%**

Table 4: Comparison of different methods on CENSER dataset

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. We extensively sample frames to produce real-time predictions as it is not possible to process 30 frames per second due to the memory constraints of the mobile app. In comparison, under similar conditions, LF-CNNs operates at 6-7 fps.

App on Kafka Spark

Kafka topics can be configured to be partitioned and replicated across multiple nodes in the Kafka cluster thus giving high scalability and fault tolerance to our stack. We configure a 3 Node Kafka cluster with 2 Intel Xeon Cores and 4 GB memory. The nodes use high speed SSDs. SSDs are important as Kafka uses disk storage to retain topic information.

A service running at the web interface runs a Kafka consumer that listens to the sink topic to get the results and forwards them onto the client applications. We configure a 4 Node spark cluster – 1 master and 3 workers. All nodes have 4 Intel Xeon Cores and 8GB memory each.

With the given specifications we were able to process around hundred requests per minute. Each request matches the faces in input image with around 1000 faces in our datasets namely – Age Invariant, Sketch and Kinship.

Conclusion

The paper proposed the ChildrEN SafEty and Rescue (CENSER) System used by the Guria (Guria NGO) non-profit organization to collect evidence of child sex trafficking from brothels in Varanasi, India. The app played a part in rescuing 13 children till date. The system uses face recognition to predict the child's age, which is then used to establish if a child is a minor. This information is utilized to convince the local law enforcement agencies to raid the brothels leading to children's rescue. The system also recognizes the children uses age-invariant face recognition to establish the identify of the children against a database. The system is used as a stand-alone mobile application to identify children abandoned at railway stations or bus stops. This system is highly application-oriented and deployable in real-world scenarios. However, the images of the children are only stored as feature representations. We don't store the raw images, mitigating any privacy concerns. Given the growing concerns about sex-trafficking, our CENSER system is only a small step towards safeguarding high-risk children. The future versions of the software include some examples of which pictures were classified incorrectly, and which correctly can give a clearer picture plus help to improve the system.

Acknowledgements

The authors thank Saurabh Bodhe and Vishnu Nimmalapudi for their help with this research.

References

Best-Rowden, L.; Hoole, Y.; and Jain, A. K. 2016. Automatic recognition of newborns, infants, and toddlers: A longitudinal evaluation. In *BIOSIG*.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

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

Chen, B.-C.; Chen, C.-S.; and Hsu, W. H. 2014b. Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Dehghan, A.; Ortiz, E. G.; Villegas, R.; and Shah, M. 2014. Who Do I Look Like? Determining Parent-Offspring Resemblance via Gated Autoencoders. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1757–1764.

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

Flanagan, P. A. 2017. Face recognition prize challenge. In *NIST*.

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

General-Assembly, U. N. O. 1989. Convention on the Rights of the Child.

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

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

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

Grother, P.; Matey, J. R.; Tabassi, E.; Quinn, G. W.; and Chumakov, M. 2013. IREX VI: Temporal stability of iris recognition accuracy. In *NIST Interagency Report*.

Grother, P.; Ngan, M.; and Hanaoka, K. 2017. Face Recognition Vendor Test. In *Ongoing NIST Interagency report*.

Guardian. 2017. The scandal of the missing children abducted from India's railway stations. https://www.theguardian.com/global-development/2017/jul/30/global-development-india-child-trafficking. Accessed: 2022-03-24.

Guria. NGO. India. http://www.guriaindia.org/. Accessed: 2022-03-24.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Tech. Rep. 07-49, University of Massachusetts, Amherst*.

Huang, Z.; Chen, S.; Zhang, J.; and Shan, H. 2021. PFA-GAN: Progressive Face Aging With Generative Adversarial Network. *IEEE Transactions on Information Forensics and Security*, 16: 2031–2045.

Jeengar, V.; Omkar, S.; Singh, A.; Yadav, M. K.; and Keshri, S. 2012. A Review Comparison of Wavelet and Cosine Image Transforms. *International Journal of Image, Graphics and Signal Processing*, 4(11): 16.

Klare, B. F.; Bucak, S. S.; Jain, A. K.; and Akgul, T. 2012. Towards automated caricature recognition. In *2012 5th IAPR International Conference on Biometrics (ICB)*, 139–146.

Klare, B. F.; Klum, S.; Klontz, J. C.; Taborsky, E.; Akgul, T.; and Jain, A. K. 2014. Suspect identification based on descriptive facial attributes. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*.

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

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

Liao, S.; Lei, Z.; Yi, D.; and Li, S. Z. 2014. A benchmark study of large-scale unconstrained face recognition. In *IEEE IJCB*.

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

Lu, J.; Hu, J.; Zhou, X.; Shang, Y.; Tan, Y.; and Wang, G. 2012. Neighborhood repulsed metric learning for kinship verification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2594–2601.

Lu, J.; Zhou, X.; Tan, Y.; Shang, Y.; and Zhou, J. 2014. Neighborhood Repulsed Metric Learning for Kinship Verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2): 331–345.

Mittal, P.; Jain, A.; Goswam, G.; Vatsa, M.; and Singh, R. 2017. Composite sketch recognition using saliency and attribute feedback. In *Information Fusion*.

Mittal, P.; Vatsa, M.; and Singh, R. 2015. Biometrics (ICB), 2015 International Conference on. In *Biometrics (ICB), 2015 International Conference on*.

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

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

Ouyang, S.; Hospedales, T.; Song, Y.-Z.; and Li, X. 2014. Cross-modal face matching: beyond viewed sketches. In *Asian Conference on Computer Vision*.

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

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

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

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

Singh, A.; Hazarika, D.; and Bhattacharya, A. 2017. Texture and Structure Incorporated ScatterNet Hybrid Deep Learning Network (TS-SHDL) For Brain Matter Segmentation. *ICCV Workshop*.

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

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

Singh, A.; and Kingsbury, N. 2017b. Efficient Convolutional Network Learning using Parametric Log based Dual-Tree Wavelet ScatterNet. *IEEE ICCV Workshop*.

Singh, A.; and Kingsbury, N. 2017c. Scatternet Hybrid Deep learning (SHDL) Network For Object Classification. *International Workshop on Machine Learning for Signal Processing*.

Singh, A.; and Kingsbury, N. 2018. Generative ScatterNet Hybrid Deep Learning (G-SHDL) Network with Structural Priors for Semantic Image Segmentation. *IEEE International Conference on Acoustics, Speech and Signal Processing*.

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

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

Xiong, C.; Liu, L.; Zhao, X.; Yan, S.; and Kim, T. 2016. Convolutional Fusion Network for Face Verification in the Wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3): 517–528.

Yang, S.; Luo, P.; Loy, C. C.; and Tang, X. 2016. WIDER FACE: A Face Detection Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yoon, S.; and Jain, A. K. 2015. Longitudinal study of fingerprint recognition. In *National Academy of Sciences*, volume 112.

Zhang, K.; Huang, Y.; Song, C.; Wu, H.; and Wang, L. 2015. Kinship Verification with Deep Convolutional Neural Networks. In Xianghua Xie, M. W. J.; and Tam, G. K. L., eds., *Proceedings of the British Machine Vision Conference (BMVC)*, 148.1–148.12. BMVA Press. ISBN 1-901725-53-7.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *CoRR*, abs/1604.02878.