ASEAN Child Face Recognition System with FaceNet

[1] Shun Lei Myat Oo, [2]Aung Nway Oo
[1] Faculty of Information Science, [2]Faculty of Computer Science,
University of Information Technology, Yangon, Myanmar
shunleimyatoo@uit.edu.mm, aungnwayoo@uit.edu.mm

Abstract— **Current researches show that deep learning is the state-of-art techniques in machine learning which outperform the human level performance. It is also the popular technique which gives high accuracy in computer vision. Face recognition is one of the ongoing researches in biometric authentication and identification system. In this paper, we develop ASEAN child face recognition system using FaceNet. We use FaceNet as feature extractions, dlib as preprocessing and classifier as K-Nearest Neighbors (KNN). FaceNet produces 128 embeddings per face as feature vectors and calculate Euclidean distance between faces in order to measure face similarity. The proposed system achieves state-of-art face recognition performance on ASEAN child face dataset with high accuracy.**

I. INTRODUCTION

Face recognition system is a system which is ability to find or identify human faces in images or videos. Face recognition is one of biometric authentication and used in many fields, such as public security, social media, finance, military and business sector. When research area of face recognition becomes wide, numerous methods and algorithms are widely developed for face recognition. When hardware devices are well developed and more powerful, deep learning become more popular. Deep learning is the state-of-art machine learning technique which achieves state-of-art performance in computer vision. Generally, human brain can detect and recognize multiple human faces instantly. But it is a challenging for computer system to detect and recognize human faces correctly.

Over the past year the performance of face recognition systems is unstable and poor and countless failure was met in real-world applications. In early 1990s, holistic learning approach influenced face recognition area [3]. Local handcrafted dominated in early 2000s and local feature learning approach was announced in late 2000s. Since traditional methods used two or three layers of feature descriptors for face recognition, results were worse for unconstrained facial changes, such as lighting, expression and pose.

In 2012, AlexNet overcome previous state-of-art methods when it won the ImageNet competition by reducing the top-5 error from 26% to 15.3% on ImageNet, achieving a top-1 error rate of 37.5 % using a deep learning technique. Convolutional Neural Network (CNN), type of deep learning methods, use multiple layers for feature extraction and transformation. The early layers extract the basic features of face and the later layer extract the high level features. In 2014, DeepFace and DeepID achieve the high accuracy on LFW (Labeled Face in-the-Wild) dataset around 90% and network became deeper.

In this paper, we used FaceNet with KNN classifier to develop the system. FaceNet achieve accuracy 99.63% on Labeled Face in the Wild (LFW) dataset and 95.12% on YouTube Faces DB. FaceNet [2] use deep convolutional neural network approach to produce optimizes embedding of faces. We train this network to produce optimize 128 features vectors of faces as output of the model rather than the intermediate bottleneck layers as in previous network architecture approaches. We use Euclidean distance to calculate the distance between the distances of faces and decide face similarity: same person have smaller distance and different person have larger distance between them. We calculate the distance and classify the people by using K-Nearest Neighbors (KNN) algorithm. Compared to previous deep learning methods, FaceNet trained the output to 128-D embedding

so representation size per face is compact. Before face dataset feed into network, we make alignment and cropping the face area as a preprocessing step.

II. Related works

In recent years, convolutional neural networks (CNNs) have shown good results for various fields such as image recognition, pattern recognition, etc. Dr. Priya Gupta et al. [7] proposed a new way for face recognition using a deep neural network. In their approach, instead of providing raw pixel values as input, only the extracted facial features are provided and can reduce the complexity of system and provided the accuracy of 97.05% on Yale faces dataset. According to Yi Sun et al. [13], address the problem of face recognition in two steps: feature extraction (design or learn features from each individual face image separately to acquire a better representation) and recognition (calculate similarity score between two compared faces using feature representation of each face. The research work [14] proposed multiple face recognition framework which is implemented on the embedded GPU system and framework contains face detection based on convolutional neural network (CNN) with face tracking and state of the art deep Convolutional Neural Network (CNN) face recognition algorithm. Zhang et al. [10] proposed a framework to detect face and alignment using unified cascaded Convolutional Neural Networks (CNN) by multi-task learning. In paper [8] used the combination of Convolutional Neural Network (CNN) and Support Vector Machine (SVM) to recognize face images. CNN is used to extract feature and train the CNN by the target dataset to extract more hidden facial features. SVM was used as a classifier instead of CNN to recognize all the classes. Deep learning for identifying missing child from the photos with the help of face recognition was presented in [9]. Face descriptors are extracted from the images using a pre-trained CNN model VGG-Face deep architecture and the child recognition is done by the trained SVM classifier. Schroff et al. [2] proposed FaceNet which is based on Deep convolutional network. The network is trained to directly optimize the embedding of the input image. This approach learns directly from the pixels of the face image and provides the higher accuracy for face recognition. The Convolutional Neural Network (CNN) with three well-known image recognition methods such as Principal Component Analysis (PCA), Local Binary Patterns Histograms (LBPH) and K–Nearest Neighbors (KNN) is tested in paper [11]. Sahar Siddiqui and group [12] proposed the face recognition for newborns, toddler and pre-school. In this research, representation learning algorithm to extract unique and invariant features from face images of newborns and toddlers, to design an efficient face recognition algorithm.

III. Background theory

A. Deep Learning

Deep Learning is a part of machine learning which algorithm work like a human brain, i.e., learns by itself. Its architecture is based on the layers of artificial neural network and learning type can be supervised, semi-supervised and unsupervised. It is also known as deep neural network. The term “deep” refers to number of hidden layers in the neural network. Traditional neural network has 2 or 3 hidden layers, while deep neural networks can have more than this. Deep learning uses a flow of multiple layers of nonlinear processing units (neurons) for feature extraction and transformation. Output of previous layers is the input of successive layers. With the flow of layers, it learns multiple levels of representations that correspond to different levels of abstraction. Mainly, deep learning can learn optimal features descriptors by itself and no more hand-crafted or hand-tuning are needed.

Recurrent neural networks, convolutional neural network, deep neural networks and deep belief networks, which are deep learning architectures, have been used in many fields such as machine translation, audio recognition, medical image analysis, drug design, computer vision, natural language processing, social network filtering, speech recognition and board games. Nowadays, one of the big examples of deep learning technology is driverless cars, robot car or autonomous car, which enable to recognize a stop sign, human, traffic lights, pedestrian and others.

A computer model, which are trained by a large dataset and powerful hardware resources using deep learning techniques, can do classification tasks from image, voice and text with state-of-art accuracy, and may be outperforms human-level performance.

Deep learning was started as theory in 1980s. It is recently popular technology due to significantly increasing amount of labeled data and computing power. With the combination of high-performance GPUs and cloud computing technology is well developed, training time consuming can reduce from weeks to hours or hours to minutes [5].

B. Face Recognition

Face Recognition system is the system which identifying or verifying human faces from digital video frame or image capture from camera. Generally, it extracts facial features from a given image or video frame and then compare unknown extracted features with the existing face features within the database. It is one of the artificial intelligence-based techniques that can tell who he/she is. It is widely used because it is non-intrusive nature or contactless process. For example, fingerprint recognition system needs user finger to place in a sensor, speech recognition system need user need to speak loudly and clearly [1].

It is widely used in today mobile phone face ID unlocks. It is also one of the popular methods in biometric authentication system. It is also applied in many areas such as public security, business like marketing, transportations, social media like Facebook and military. For example, Facebook use facial recognition software to tag individuals in photographs. Each time a person is tagged in a photo, a system store mapping information about that person's facial characteristics. Once system collected enough data of that person, system can detect that person when that person appears in a new photo. We can see it in another form of technology like robotics vision system.

Face recognition is performed in two steps, namely extraction and selection, and classification. Face recognition algorithms extract landmarks or features from an image of human face. An algorithm may analyze the relative position of size and shape of eyes, nose, cheekbones, mouth, eyebrows and jaw. These features are used to match with other features.

Input image data are transformed into a slightly more abstract and composite representations at each level in learning process of deep learning. In face recognition, the raw inputs are a matrix of pixels. The first representation layer may extract the pixels and encode edges, the second layer may compose and extract arrangements of edge, the third layer may extract a nose and eyes, the fourth layer may extract eyebrows and fifth layer may recognize that the image contains a face.

C. Convolutional neural network

Convolutional neural network (CNN) is mainly used in computer vision. It is an efficient machine learning method in face recognition for extract optimizes features from faces. It is a type of Feed-Forward Neural Network and Back-Propagation algorithm is used in training process. Convolutional neural network (CNN) can also be used for handwritten character recognition for feature extraction [4].

CNN use 2D convolutional layers to learn features from input image and this architecture is suitable for processing 2D data, such as images. CNN can do features extraction by itself without manually identify the feature descriptors to classify images. CNN can extract optimal features from image by using automated feature descriptors which are obtained by training the network on a largest number of images. So, deep learning model can extract features for computer vision tasks with highly accurate using this automated feature extraction.

CNN is the type of deep neural network with an input layer, one or more hidden layers and an output layer. CNN generally consists of convolution, pooling (sub-sampling) and fully-connected layers [6]. Each layer consists of one or more neurons that have learnable weights and biases. Each neuron receives inputs and produce non-linearity output by activation function. Neurons of input layer receive the original image vector as input. The intermediate layers are called hidden layers

and neurons in hidden layers are called hidden units. Loss function (e.g. Softmax or KNN) is applied on the fully connected layer to calculate the probability distribution of classes. CNN transform the final class scores from the original image by passing layer by layer. [4] For example, AlexNet use Rectified Linear Units (ReLU) as activation function and include layers for local contrast normalization, local response Normalization (LRN) and Dropout. Fully connected layers perform classification process. Final layer, output layer, is often Softmax layer when many class label neurons which represent the probabilities of each class are existing.

CNNs use ten or more hidden layers to detect different features of image. Every hidden layer increases the complexity of the learned image features. A CNN model applies different filters to identify edges or parts in an image to detect the specific object in an input image. For example, in face recognition, first hidden layer learns how to detect edges, the second learn how to detect eyes and successive layers learn the more complex parts of the given image, and searching for a face.

D. FaceNet

FaceNet [2] is developed by Google Inc. in 2015. It used Euclidean distance to calculate the face similarity. FaceNet embedding can make face recognition system to implement easily. FaceNet are trained to produce optimize embedding rather than used as intermediate bottleneck layer in previous deep learning approaches. It used novel online triplet mining method to train the network with triplets of roughly aligned matching / non-matching face patches generated. This method is using a deep convolutional network in order to learn a Euclidean embedding per image. The network is trained that embedding space of faces of same person have small distances and different person have large distances. Distance threshold involved between two embedding and can be considered as KNN classification problem. FaceNet used triplet loss function based on LMNN to train its output to be a compact 128-D embedding. Triplets consist of two matching face thumbnails and a non-matching face thumbnail and loss aims to separate the positive pair from the negative by a distance margin while two positives are close each other. Choosing which triplets to use to train the network is very important for achieving good performance.

FaceNet Method Overview: FaceNet use deep convolutional network and it is inception type network. Model black box structure is shown in Figure 1.

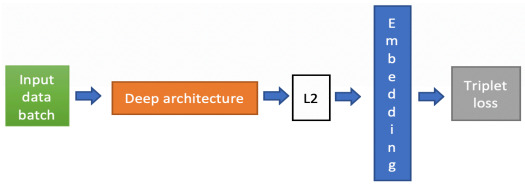

Figure 1. Model structure.

Data batch as input to deep CNN followed by L2 normalization to result face embedding. This is followed by triplet loss during training.

End-to-end learning is the important part of the whole system. In the end, triplet loss is used to achieve optimal face recognition. We retrieve an embedding from an image and calculate the square distance between all faces by using Eq. (1); a pair of face image with same identities is small, whereas a pair of face images with different identities is large. We choose the using pairs of positives and negatives to compare rather than directly used all pairs.

$$\sum_{i}^{N}\left[\|f(x_i^a)-f(x_i^p)\|_2^2-\|f(x_i^a)-f(x_i^n)\|_2^2+\alpha\right] \quad (2)$$

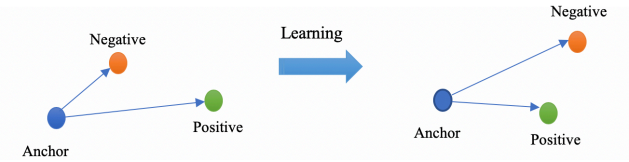

Figure 2. The triplet loss minimizes the distance between anchor and positive (both of them have same identity), and maximize the distance between anchor and negative (both of them have different identity).

Triplet Loss: The embedding is represented by $f(x) \in R^d$. It embeds an image x into a d-dimensional Euclidean space. We need to make sure that an image x_i^a (anchor) of a specific person is closer to all images x_i^p of the same person (positive) than any image x_i^n of any

other person (negative) in Eq. (2). This is shown in Figure 2.

$$|| x_i^a - x_i^p ||^2 + \alpha < ||x_i^a - x_i^n||^2, \forall (x_i^a, x_i^p, x_i^n) \in T \quad (2)$$

Where, α is a margin that is enforced between positive and negative pairs. T is the set of all possible triplets in the training set and has cardinality N.

We need to choose triplets type to train. There are 3 types of triplets; easy triplets, semi-hard triplets and hard triplets. Easy triplets may not contribute the model because it is easy to become loss of 0. Hard triplets may improve the performance of models. We describe detail in triplet selection.

Triplet Selection: We need to choose hard triplet to fast convergence and to improve model performance. This mean that, given x_i^a we need to select an x_i^p :: $argmax_{x_i^p} || f(x_i^a) - f(x_i^p) ||^2$ (hard positive) and an x_i^n such that $argmin_{x_i^n} || f(x_i^a) - f(x_i^n) ||^2$ (hard negative). Two ways of triplet mining are offline and online triplet mining. In offline triplet, embedding is calculated from all dataset and then select semi-hard or hard triplets are selected to train for one epoch. In online triplet, we only need to compute embedding for each batch of inputs. Online triplet mining is more efficient than offline because offline need to pass all training data to generate triplets.

IV. Overview of the Process

A. Preprocessing

We use the dlib for preprocessing to detect, align and crop of raw data image. We use 68 face landmarks to identify the geometric structure of faces in digital images and attempt to obtain canonical alignment of face based on scale, translation and rotation. It detects face in image and extracts the facial features around eyebrows, eyes, mouth, chin and nose as shown in Figure 3. Face images, which are preprocessed, can be input dataset for FaceNet model. Augmentation process is done to increase the number of datasets to train.

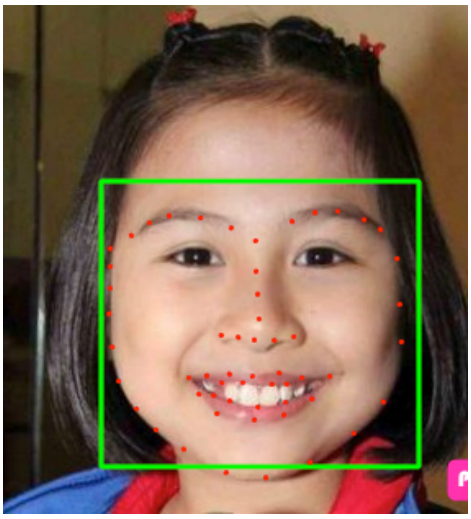

Figure 3. Facial features detection in image.

B. Learning

CNN is trained using Stochastic Gradient Descent (SGD) with standard back propagation and use AdaGrad optimizer. Learning rate is start with 0.05 which are gradually decreased at the end of epoch. Model architecture consists of 22 layers deep. It is total of 140 million parameters. We use pre-trained model and use fine tuning technique. Last layer of model generates 128 embedding for each image. We retrain it with softmax cross entropy. This network is train on about 30 epochs and iteration over 350. The detail architecture of FaceNet is shown in Figure 6.

Firstly, training data pass preprocessing step. Network (FaceNet) is trained on preprocess data to generate embedding. Classifier is trained using embedding to output result.

C. Classification

K-Nearest Neighbors (KNN) is used for classification. Extracted facial features are stored as training input data for KNN. KNN used this embedding for training. The training process is shown in Figure 4.

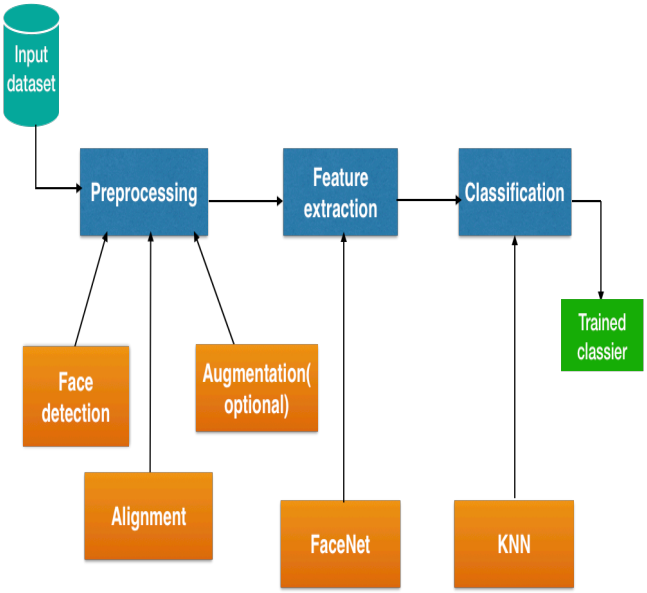

Figure 4. Training process.

After training, KNN match between existing faces and unknown faces, and output the prediction result. Figure 5 illustrate the overview of the system.

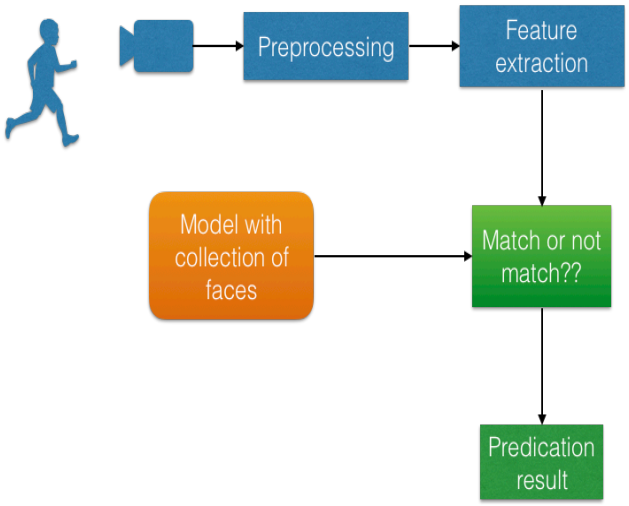

Figure 5. Overview of face recognition System.

Test data is preprocessed to generate embedding. Generated embedding and existing embedding are matched to determine who he or she is.

layer	size-in	size-out	kernel	param	FLPS
conv1	220×220×3	110×110×64	7×7×3, 2	9K	115M
pool1	110×110×64	55×55×64	3×3×64, 2	0	
rnorm1	55×55×64	55×55×64		0	
conv2a	55×55×64	55×55×64	1×1×64, 1	4K	13M
conv2	55×55×64	55×55×192	3×3×64, 1	111K	335M
rnorm2	55×55×192	55×55×192		0	
pool2	55×55×192	28×28×192	3×3×192, 2	0	
conv3a	28×28×192	28×28×192	1×1×192, 1	37K	29M
conv3	28×28×192	28×28×384	3×3×192, 1	664K	521M
pool3	28×28×384	14×14×384	3×3×384, 2	0	
conv4a	14×14×384	14×14×384	1×1×384, 1	148K	29M
conv4	14×14×384	14×14×256	3×3×384, 1	885K	173M
conv5a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv5	14×14×256	14×14×256	3×3×256, 1	590K	116M
conv6a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv6	14×14×256	14×14×256	3×3×256, 1	590K	116M
pool4	14×14×256	7×7×256	3×3×256, 2	0	
concat	7×7×256	7×7×256		0	
fc1	7×7×256	1×32×128	maxout p=2	103M	103M
fc2	1×32×128	1×32×128	maxout p=2	34M	34M
fc7128	1×32×128	1×1×128		524K	0.5M
L2	1×1×128	1×1×128		0	

Figure 6. FaceNet network architecture.

V. Experimental Results

The proposed method is tested on my own created dataset which consists of color images of celebrity children with different expression. 8 photos per person and total are 35 people. Each person represents one class. Their facial expression is happy, disgust, normal and their facial pose is frontal and non-frontal. Generally, total data is split as 20% for testing and 80% for training. Train accuracy and loss of testing is show in Figure (7) and (8). Accuracy of proposed system is over 99%. We can see distance of output embeddings using 3D graph as shown in Figure (9).

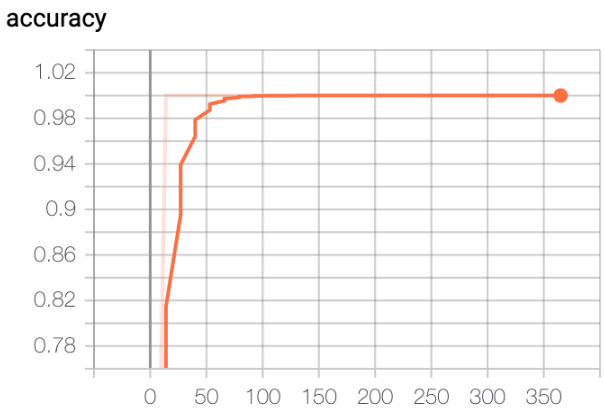

Figure 7. Training accuracy is over 99%. Row is number of iteration and column is percent of accuracy.

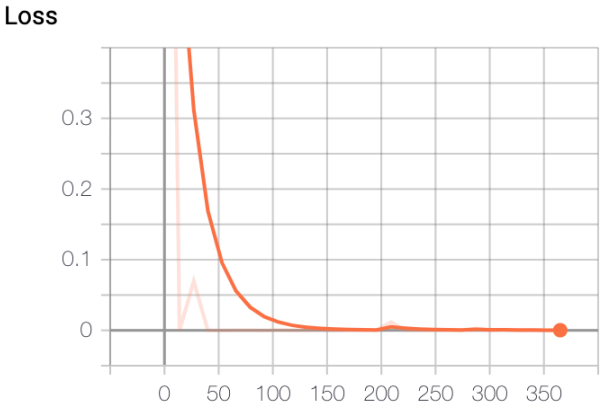

Figure 8. Training loss is nearly 0. Row is number of iteration and column is percent of loss.

Figure 9. Embeddings which generate from FaceNet show as 3D graph. One color represents one class.

VI. Conclusion

In this proposed system, raw input data are preprocessed to detect, align and crop face using dlib and then feed them into FaceNet to get good accuracy rather than directly feed raw image which contain other parts such as background or shoulder. FaceNet extract features from these preprocess data and generate 128 embeddings. These 128 embeddings are fed into KNN classifier to classify child face. The proposed method gets high accuracy.

Acknowledgment

We would like to express our gratitude to all my Teachers from University of Information Technology for their suggestion and supports throughout this research work. I would like to express my gratitude to Codigo company which helps and supports me throughout my research. Finally, I also thank all of my colleague for their participation and contribution for this study.

References

[1] Daniel Sa´ez Trigueros, Li Meng, Margaret Hartnett, “Face Recognition: From Traditional to Deep Learning Methods”, arXiv:1811.00116v1 [cs.CV], 2018, pp. 1-2

[2] Florian Schroff, Dmitry Kalenichenko , James Philbin , “FaceNet: A Unified Embedding for Face Recognition and Clustering” , CVPR 2015, pp. 1-4

[3] Mei Wang, Weihong Deng, “Deep Face Recognition: A Survey”, arXiv:1804.06655v 8 [cs.CV], 2019, pp. 1-2

[4] Jakob Grundstro¨m, “Face Verification and Open-set Identification for Real-Time Video Applications”, Master’s Thesis, 2015, pp.21

[5] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, “Deep Learning for computer vision: A Brief Review”, Hindawi, 2018, pp. 1

[6] Jianxin Wu, “Introduction to Convolutional Neural Networks”, Sematic Scholar ,2017, pp.5-6

[7] Dr. Priya Gupta, Nidhi Saxena, Meetika Sharma, Jagriti Tripathi. ” Deep Neural Network for Human Face Recognition”, I.J. Engineering and Manufacturing, 2018, 1, pp. 63-71

[8] Sun, Yi, Xiaogang Wang, and Xiaoou Tang. "Hybrid Deep Learning for Face Verification." IEEE Transactions on Pattern Analysis and Machine Intelligence 38.10 (2016), pp. 1997-2009.

[9] Savath Saypadith and Supavadee Aramvith.“Real-Time Multiple Face Recognition using Deep Learning on Embedded GPU System”, Proceedings, APSIPA Annual Summit and Conference 2018, pp. 1318-1324

[10] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using

multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, 2016, pp. 1499-1503

[11] Patrik KAMENCAY, Miroslav BENCO, Tomas MIZDOS, Roman RADIL, "A New Method for Face Recognition Using Convolutional Neural Network", Advances in Electrical and Electronic Engineering, Volume: 15, Number: 4, 2017, pp. 663-672

[12]Sahar Siddiqui, Mayank Vatsa, and Richa Singh , "Face Recognition for Newborns, Toddlers, and Pre-School Children: A Deep Learning Approach", 24th International Conference on Pattern Recognition (ICPR), 20-24 Aug. 2018.

[13] Shanshan Guo ; Shiyu Chen ; Yanjie Li, "Face recognition based on convolutional neural network and support vector machine", IEEE International Conference on Information and Automation (ICIA), 2016.

[14]Pournami S. Chandran ; N B Byju ; R U Deepak ; K N Nishakumari ; P Devanand ; P M Sasi, "Missing Child Identification System Using Deep Learning and Multiclass SVM", IEEE Recent Advances in Intelligent Computational Systems (RAICS), 6-8 Dec. 2018.