Learning Deep Features for Hierarchical Classification of Mobile Phone Face Datasets in Heterogeneous Environments

Neeru Narang[1], Michael Martin[1], Dimitris Metaxas[2] and Thirimachos Bourlai[1]
[1] MILab, LCSEE, West Virginia University, Morgantown, WV, USA
[2] Department of Computer Science, Rutgers University, NJ, USA

***Abstract*— In this paper, we propose a convolutional neural network (CNN) based, scenario-dependent and sensor (mobile device) adaptable hierarchical classification framework. Our proposed framework is designed to automatically categorize face data captured under various challenging conditions, before the FR algorithms (pre-processing, feature extraction and matching) are used. First, a unique multi-sensor database (using Samsung S4 Zoom, Nokia 1020, iPhone 5S and Samsung S5 phones) is collected containing face images indoors, outdoors, with yaw angle from -90° to +90° and at two different distances, i.e. 1 and 10 meters. To cope with pose variations, face detection and pose estimation algorithms are used for classifying the facial images into a frontal or a non-frontal class. Next, our proposed framework is used where tri-level hierarchical classification is performed as follows: Level 1, face images are classified based on phone type; Level 2, face images are further classified into indoor and outdoor images; and finally, Level 3 face images are classified into a close (1m) and a far, low quality, (10m) distance categories respectively. Experimental results show that classification accuracy is scenario dependent, reaching from 95 to more than 98% accuracy for level 2 and from 90 to more than 99% for level 3 classification. A set of experiments is performed indicating that, the usage of data grouping before the face matching is performed, resulted in a significantly improved rank-1 identification rate when compared to the original (all vs. all) biometric system.**

I. Introduction

Standard face recognition (FR) systems typically result in very high identification rates, when the face images are taken under highly controlled conditions, i.e. indoors, during day time, at short range etc. However, in law enforcement and security applications, investigators deal with mixed FR scenarios that involve matching probe face images captured by different portable devices (cell phones, tablets etc.), and at a variable distances against good quality face images (e.g. mug shots) acquired using high definition camera sensors (e.g. DSLR cameras) [16].

The worldwide popularity of mobile devices, due to rapid increase in processing power, sensor size and storage capacity offers a unique collection of mobile databases for studying more challenging FR scenarios [16]. Most modern smartphones contain both rear and front facing cameras capable to capture both images and videos. Online statistics from 2016 depicted that the total number of smartphone users have reached 2.08 billion within 2016. Mobile based face identification systems are used by law enforcement agencies for security purposes [1], [3]. The major challenges for mobile based FR are variation in illumination conditions, poor face image quality (due to various factors including noise and blurriness due to movement of hand-held device during collection), variations in face pose and camera sensor quality. These factors can degrade the overall performance of FR systems (including pre-processing, face detection, eye detection and face matching). To facilitate recognition performance, knowing the specific image category (phone type, indoors, outdoors, distance of the subject from the camera based on origin) is important in order to set the proper parameters for image quality prediction, as well as face and eye detection.

Predicting this image category is a task that humans can perform easily. However, such a process is time consuming, especially, when dealing with large scale face datasets. Therefore, an automatic process of classifying images into a specific scenarios is needed. A lot of early work in the area of image based data grouping is based on the usage of low level features to classify scenes into indoors and outdoors categories. Vailaya et al. [4] proposed an (visible band) image classification that depends on using low level features extraction and processing. The authors reported hierarchical classification results, where, first, images are classified into indoor or outdoor. Then, outdoor images are further classified as city or landscape categories and, finally, landscape images are classified into a forest, sunset or mountain categories. Recently, deep convolutional neural networks have achieved great success in the area of computer vision, machine vision, image processing and biometrics for the classification of scenes, object recognition, detection, face authentication and quality assessment. Some examples are the work of Gupta et al. [6] that proposed a probabilistic neural network (PNN) based approach for the classification of indoor vs. outdoor visible band images. Sarkar et al. [9], proposed a deep feature based face detector for mobile devices using front-facing cameras. The authors applied their proposed method on the database collected under varying illumination conditions, poses and partial faces. They reported that the deep feature method outperformed the traditional methods and the developed system can be implemented for offline systems. The use of deep CNNs has also been extended for face recognition applications. In [15], Parkhi et al. conducted a large-scale collection of face images from online search engines and trained the VGG-16 network to extract deep face features that can be used for face recognition. This approach

This work was supported by Department of Homeland Security (DHS).

978-1-5090-4023-0/17/$31.00

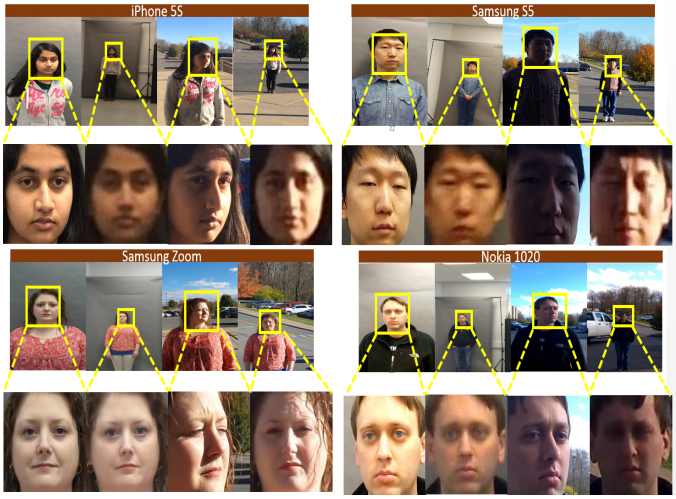


Fig. 1. Multi-sensor face image database collected using various cell phones under the challenging conditions.

achieved high face recognition and verification accuracy in a variety of scenarios.

Most of the existing classification [6], [4] systems are based on image scene classification into indoors or outdoors and result in operationally acceptable classification rates. In this paper we propose a solution to a more complicated problem as we are dealing with a multi-sensor cell phone face database captured at variable standoff distances, illumination conditions and pose angles. We propose a deep learning based, scenario-dependent, and sensor-adaptable algorithmic approach for the classification of data in terms of phone type, conditions, and standoff distances. To show the impact of classification or database pre-screening, face matching experiments are performed using local binary patterns (LBP) and the VGG Face matcher with data grouping using our proposed classification approach, or without data grouping, i,e, using the original face database to apply a set of FR matchers.

II. Motivation

There are various studies reported in the literature, where face images captured from different devices are matched against visible good quality face images [11]. To our knowledge, there is no study reported where all face datasets available are simultaneously collected (i) using variable portable devices that have the capability (sensors) to acquire mid range (>10 meters) face images, (ii) at different standoff distances and, (iii) at indoors vs. outdoors conditions.

In this paper, our main contributions are the following: A multi-sensor (MS) face image database is collected using a set of cell phone devices including Samsung S4 Zoom, Nokia 1020, Samsung S5 and iPhone 5S. The visible band face database is collected indoors, outdoors, at standoff distances of 1m and 10m respectively, and with different pose angles as shown in Fig. 1. Automated face detection and pose estimation method is designed and developed that selects full-frontal face images that will be used to perform the hierarchical classification experiments. Our proposed hierarchical classification framework is composed of three levels of operation: at the top level (Level 1), images are classified into phone types, which are then further classified as indoor or outdoor face images (Level 2); finally, indoor and outdoor face images are classified as either close or far standoff distance images (Level 3). The complete proposed framework is represented in Fig. 2.

III. Methodology

In this section, we outline the challenging database collected in our lab and discuss the proposed CNN architecture that we used to perform grouping on data based on a three level classification scheme.

A. Database

A multi-sensor visible database (DB1) was collected to perform the classification experiments using four phones. Face videos were collected indoors, outdoors, at a standoff distance of 1m and 10m as shown in Fig. 1. In total, the database consists of 50 subjects. For each subject, 16 videos are collected, including 4 videos (2 videos: indoors 1m and 10m and 2 videos: outdoors 1m and 10m) from each phone. Each video consists of around 700 frames and in total almost 11,200 frames for a single subject. Each video is captured with head poses varying from -90° to +90°. Two scenarios are selected to collect the database.

- Close Distance ($\sim 1m$): Involves both the face and shoulder part of the body.
- Far Distance ($\sim 10m$): Involves full body images. Please check in Fig. 1 sample images of the database collected for the aforementioned scenarios. In the left side of the figure, the top 2 rows represent, video frames collected from a iPhone 5S and the bottom 2 rows represent, the video frames collected from a Samsung S4 Zoom (equipped with $10\times$ optical zoom to capture close-ups from far distances). In the right side of Fig. 1, the top two rows represent the video frames captured from a Samsung S5 and the bottom two rows from a Nokia 1020. A multi-sensor database (DB2) collected from 92 subjects in our lab under un-constrained conditions including, outdoors, at standoff distances of 2.5 to 10 meters for Nokia, Samung S5 and iPhone 5S and 25 to 80 meters for Samsung S4 Zoom is used to train the CNN network.

B. Classification of Frontal vs. Non-Frontal Face Images

Face recognition and image classification systems perform well when good quality full-frontal face images are used. In this work, the data we are using is very challenging as discussed. In order to keep only full-frontal face images to perform the other pre-processing and face matching experiments, we selected an automated face detection and pose estimation method [13]. Finally, the frontal vs. non-frontal face classification (for all the phones, indoors, outdoors, at close and far distance) is performed based on the automatically estimated pose angle. We achieved good results for the

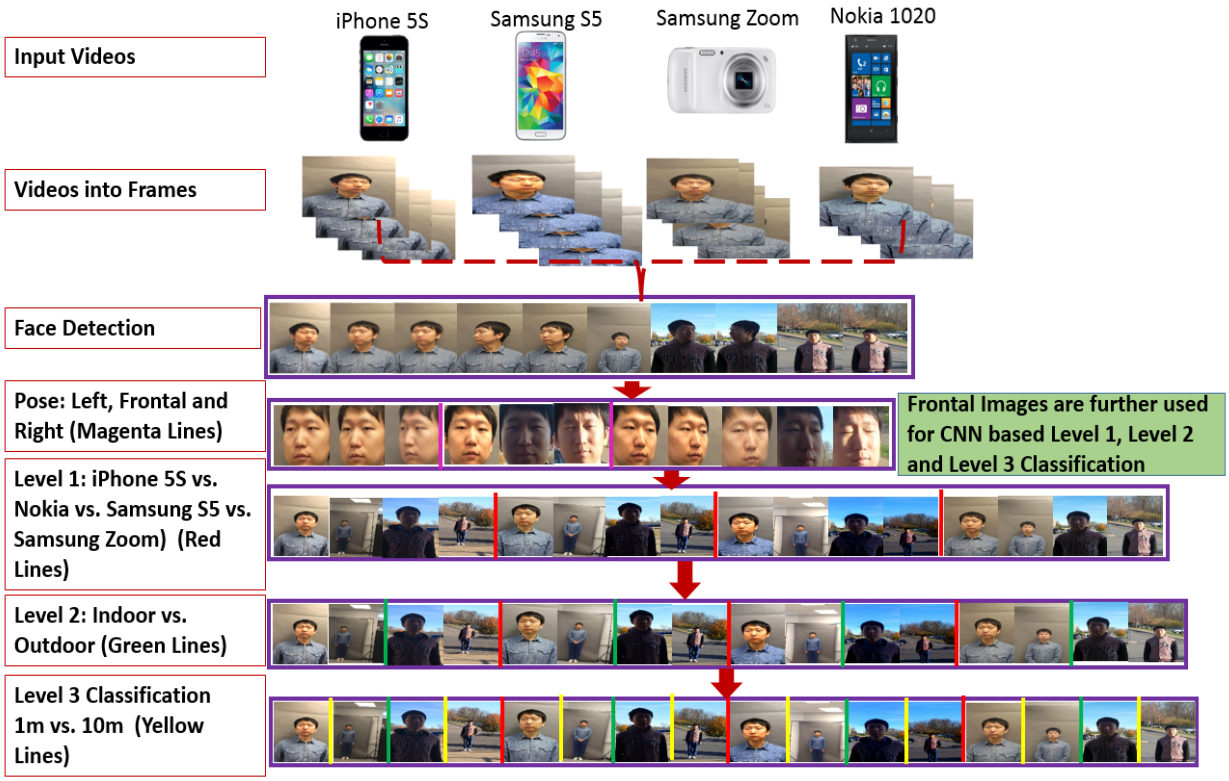


Fig. 2. An overview of our proposed hierarchical classification approach when using the face images captured from the mobile phones under various challenging conditions. Please note that after performing the first three pre-processing steps converting (video into frames, face detection and pose estimation), the frontal face images are selected to perform the 3-Level classification approach discussed in Section III.

face images collected at close indoors and outdoors distances as shown in Fig. 3. The main challenge was to detect the face for the images collected outdoors, at far distances, with challenging pose angles and for the lower quality camera phones with no optical zoom in capabilities. For some cases, we got either no output or multiple outputs with various false positives and false negatives rates as shown in Fig. 4.

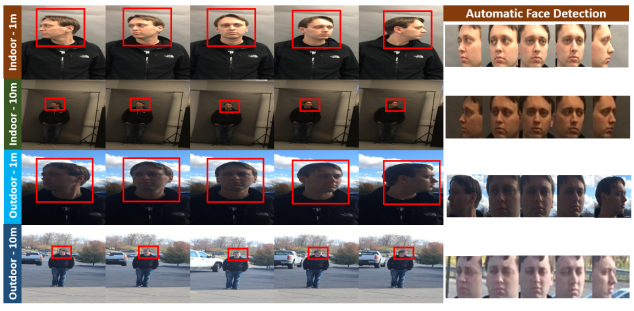


Fig. 3. Face detection on face images captured from the mobile phones under un-constrained conditions.

- **Face Detection:** In our work, we used a cascaded adaboost classifier for face detection [2] and adapted the algorithm for the challenging multi-sensor cell phone face database collected in our lab. To address the issue, for the output with multiple images, once we find the bounding boxes, we applied the condition to search for a face bounding box based on the size of box and the pixel location.

Fig. 4. The Viola and Jones face detector with *True Positives, False Positives and False Negatives* results.

- **Pose Estimation:** There are three types of face pose rotation angles such as yaw, pitch and rolling angle [12]. We used the algorithm developed by Aghajanian et al. [13] to estimate the face pose, for the database collected under un-controlled conditions. This algorithm classifies the detected face images into three categories left profile, frontal and right profile, with yaw angle from -90° to 90°.
  Next, a probabilistic framework is generated, where the subjects' faces are represented by non-overlapping grid

of patches and a generative model, based on this patch representation, is further used for pose estimation on test images. To perform the pose estimation experiments, we selected the radial bias functions (RBF) with size of RBF9D, patch grid resolution of 10×10, number of patches of 100, and a standard deviation (for RBF) of 45. We were able to achieve good results for the database collected from sensors with large size face images (when using Samsung S4 Zoom and Nokia 1020) as shown in Fig. 5. It was more challenging to process face images of smaller spatial resolution (when using the iPhone, Samsung S5). For the outdoor data at far distances, some of the faces are mis-classified with a wrong pose angles. The face images with frontal view and yaw angle of 0° are classified as frontal and face images with left and right profile are classified as non-frontal (yaw angle less than and greater than 0°).

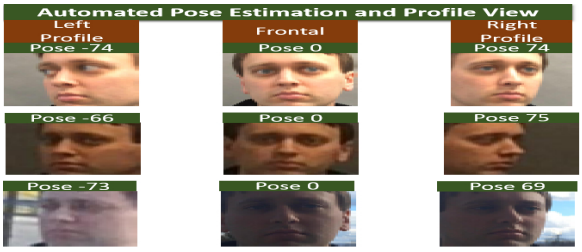


Fig. 5. Pose estimation from -90° to +90° angle for the detected face images.

C. Convolutional Neural Network

We proposed a scenario dependent and sensor adaptive CNN network, which is capable of classifying test images with class label of phone type, illumination condition and standoff distance (see in Fig. 6). The lower layer features are favorable for landmark location and pose estimation [9]. Whereas, the higher layer features are best fit to perform the classification task. Our work is focused on higher layer features to perform the hierarchical classification. To generate the model, we selected the visual geometry group (VGG) CNN architecture [14]. The network consists of convolutional layers (a bank of linear filters), followed by a rectification layer such as rectified linear unit (ReLU) and max pooling layer, along with fully connected layers [14], [10], [15].

Model Architecture: The architecture consists of 8 layers including: 3 convolutional layers followed by 2 pooling layers, 1 rectified layer and 2 fully connected layers (see in Fig. 6). The convolutional layers output the feature maps, where each element is computed by a dot product between the local region and the selected set of filters [14]. The pooling layer is applied to perform the down-sampling operation via computing the maximum of local region. The last fully connected layer is softmax that computes the scores for each class.

The first convolutional layer of the network has 20 operational filters of size 5×5 followed by max pooling layer which takes the maximum value of 2×2 regions with 2 strides. The output of the previous layer is processed by the second convolutional layer, which consists of 20 filters with size of 5×5 filters followed by a max pooling layer. The output of the previous layer is processed by a third convolutional layer, which consists of 50 filters with a 4×4 filter size, followed by a rectified layer. The output of the third convolutional layer is processed through the fully connected layer and finally, the output is fed to a softmax layer that assigns a label to each class. For level 1 classification, the last softmax layer assigns a label in terms of the of phone type, iPhone 5S, Samsung S5, Samsung S4 Zoom and Nokia 1020. For the level 2 classification, the last softmax layer assigns either an indoor or outdoor label. Finally, for level 3 classification, softmax assigns a label to each input face as a close or far distance face image (see in Fig. 6).

Training and Testing: In our work, due to a limitation of resources, we did not collect a large scale training database from available image repositories and then, label the database manually since this would be a very time consuming process. Also, there was no pre-trained multi-sensor network models available to use for our CNN network. Thus, we trained the models on our original database for each level, i.e. from level 1 to level 3. To train our system for each level, we selected a fixed value of 0.92 for the momentum parameter, a batch size of 100 and a learning rate of 0.002. The classification framework is performed for 13 different set of epoch values, namely $4, 8, 12, 16, ..., 52$, i.e. for each level of the classification ranging from level 1 to level 3. The classification results are represented in the experimental results section.

- **Level 1**: The input face database used consists of images collected under variation in illumination conditions, standoff distances, sensor type, ethnicity and gender. To train the system, 4 labeled classes in terms of the phone type are used to train the system. The network is trained to classify each of the test images into the right phone type face image (e.g. face images collected using an iPhone, are categorized into the iPhone face folder).
- **Level 2**: We trained our level 2 classifier using both indoor and outdoor face images for each phone. Thus, the original face data categorized into a phone type face folder (Level 1), are now further categorized into either an indoor or outdoor category.
- **Level 3**: We trained our level 3 classifier using 1m and 10m face images for indoor and outdoor class face images captured from any phone used in our data collection process. Thus, the data used for the training the level 2 classifier are now further classified into either an 1m or 10m distance category.

D. Face Matching

Local Binary Patterns: LBP matcher is used to extract the appearance and texture information from human faces and is invariant to changes in illumination conditions [5].

VGG Face Network: Although the VGG-Face network is

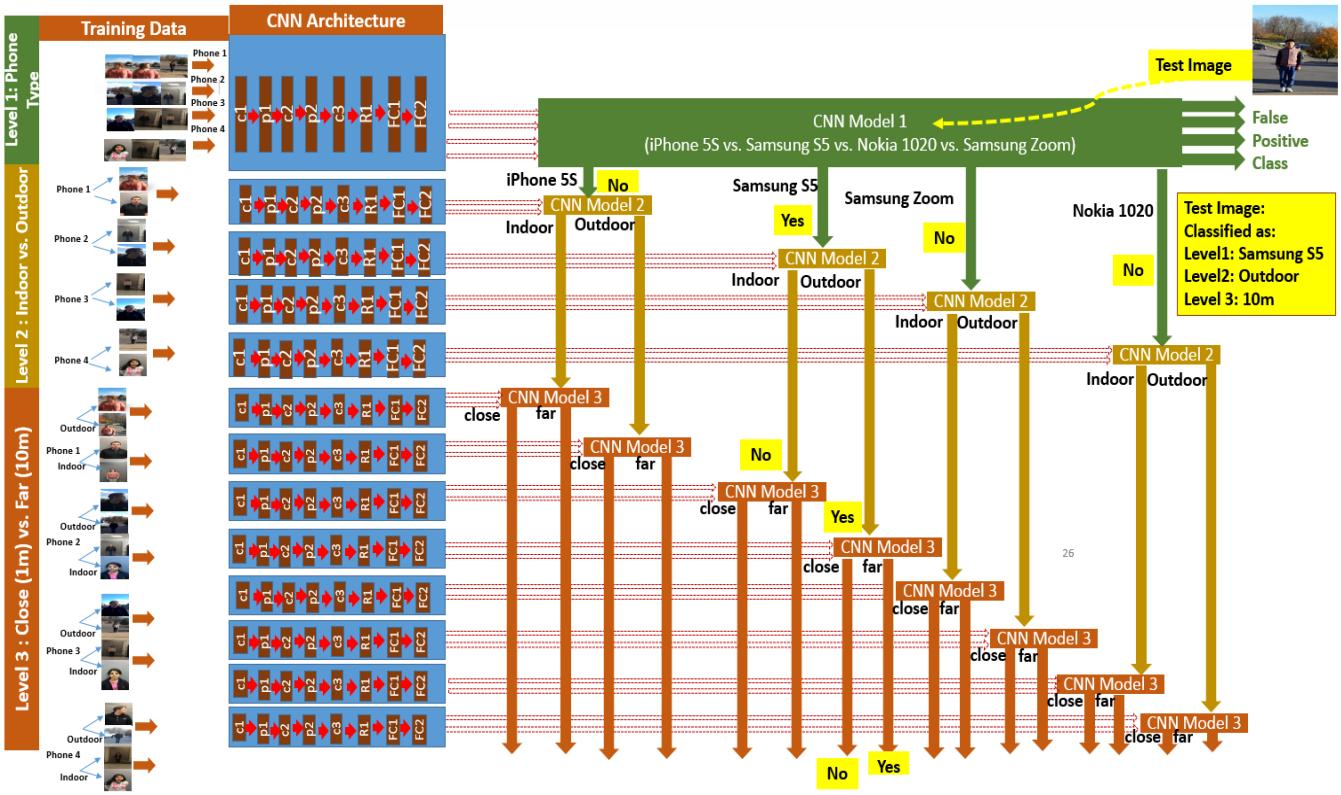


Fig. 6. Proposed CNN scheme to perform hierarchical classification: C represents the convolution layer, P the pooling layer, R the rectification layer and FC the fully connected layer. The layer number (for example C1) represents the first convolution layer. Thus, in summary, the proposed CNN architecture consists of three convolution layers (C1, C2 and C3), two pooling layers (P1 and P2), one rectified layer (R1) and two fully connected layers (FC1 and FC2).

trained on a specific set of identities, the features extracted in intermediate layers can be applied to other identities for recognition. To evaluate our dataset we used the features extracted in the last fully connected layer before the softmax as deep facial features (fc8 in [15]), which can be compared using the Euclidean distance to generate a distance score between two face images.

IV. Experimental Results

In the first set of experiments, we aim to illustrate how the mobile phone adaptable deep learning system performs for Level 1 classification, where the multi-sensor data collected under un-controlled conditions is used to classify phone types. For Level 2 classification, a set of experiments is performed, where for each phone type the face images are further classified into an indoor or outdoor class. For Level 3 classification, both indoor and outdoor face images are further classified into close (1m) or far (10m) distance face images (see Fig. 2).

A. Level 1: CNN based Phone Type Classification

In this CNN network is proposed for the grouping of the database into four classes with labels iPhone 5S, Samsung S4 Zoom, Nokia 1020 and Samsung S5. To train the CNN network, three scenarios are selected including, *Scenario 1*, the subjects in the training and test sets are different, and the images are taken at different locations and days. The database DB2 (collected outdoors at standoff distance from 2.5 to 80 meters away) is used for the training and DB1 for testing. For *Scenario 2*, we selected DB1 for training (50%) and the rest of the database for testing without any overlap of subjects. For Scenario 3, the images collected from both DB1 (50%) and DB2 (All Data) are used for training, while the DB1 for testing. There is no overlap of subjects in the training and test sets. Note that we could use 90% of the data for training (10-fold cross validation) and the expected accuracy would be much higher. Due to time constraint this was not possible. Table 1, depicts the accuracy results for the grouping of data in terms of phone types from CNN used. Based on the results, we concluded that the classification results of highest accuracy was achieved from scenario 3 and, the classification accuracy reaches almost 75%.

TABLE I

Classification results from CNN: Phone Type.

Class Type: iPhone 5S vs. Samsung S4 Zoom vs. Nokia 1020 vs. Samsung S5		
Scenario 1: Training DB2 and Testing DB1		
Accuracy	0.40	
Scenario 2: Training DB1 and Testing DB1		
Datasets	**Set 1**	**Set 2**
Accuracy	0.71	0.69
Scenario 3: Training DB1+DB2 and Testing DB1		
Datasets	**Set 1**	**Set 2**
Accuracy	0.75	0.70

B. Level 2: CNN based Conditional Classification

The output face images of the Level 1 classification task, were further classified into indoor and outdoor face images. To train the CNN network, the data was divided into different set sizes (10%, 20%, 30%, 40% and 50%) for training , while the rest of the data was used for testing. In order to examine the effectiveness of the classification system, we repeated this process five times, where each time a different training set was randomly selected and rest of the data was used for testing (without overlap of subjects).

Classification was performed for 13 different set of epoch values, namely $4, 8, 12, ..., 52$ for each phone and, finally, selected the value where we achieved the best classification results from all five sets. The results for Samsung S4 Zoom are shown in Fig. 7, where we achieved the highest classification accuracy at the following setting, i.e. when the epoch value was set to 16 and 50% of the data was used for training. The same set of experiments were performed for rest of the remaining three phones.

In Table II, classification results are presented with the epoch

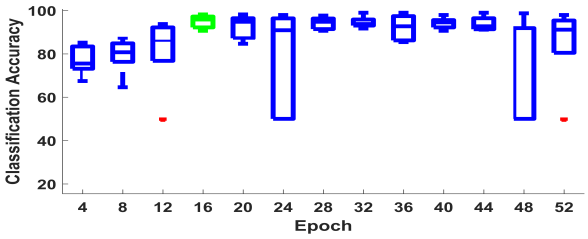


Fig. 7. Classification accuracy vs. Epoch after running a set of five experiments.

value resulted in the highest accuracy from Samsung S4 Zoom and iPhone 5S. The same set of experiments were repeated for Samsung S5 and Nokia 1020. Based on the results, we concluded that the classification results of highest accuracy were achieved when 50% of the data was used for training. For Samsung S4 Zoom, Nokia 1020 and Samsung S5, the classification accuracy on average from 5 sets reaches greater than 93%. For iPhone 5S, the classification accuracy on average reaches more than 91%.

In Table III, the classification results of highest accuracy from each phone are represented. Based on mean and variance plots we concluded that, the classification accuracy reaches approximately 95% when the Nokia 1020 face dataset was used (this was the highest accuracy when compared to any other phone specific face dataset used. See Fig. 8).

C. Level 3: CNN based Standoff Distance Classification

The labeled indoor and outdoor data from Level 2 classification is used to classify into either a close or a far distance class. The classification experiments were performed for 13 different set of of epoch values, namely $4, 8, 12, ..., 52$. In Table IV, the highest classification accuracy results are presented for each of the Samsung S4 Zoom, iPhone 5S, Nokia 1020 and Samsung S5 face datasets.

TABLE II

Classification results from CNN: Indoor vs. Outdoor

Datasets	set 1	set 2	set 3	set 4	set 5
Samsung S4 Zoom					
train 10%	78.05	70.35	80.52	50.0	82.70
train 20%	86.84	90.39	83.55	91.56	87.87
train 30%	88.97	50.00	89.71	89.34	94.67
train 40%	96.34	90.09	85.34	94.83	92.67
train 50%	95.83	92.71	96.88	98.18	90.63
iPhone 5S					
train 10%	87.99	50.00	50.00	50.00	84.21
train 20%	50.00	50.00	90.13	89.80	50.00
train 30%	92.85	93.57	91.91	90.71	50.00
train 40%	94.17	50.00	91.13	92.89	93.97
train 50%	94.27	88.54	88.54	95.57	94.01

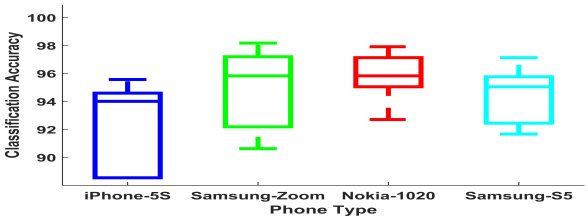


Fig. 8. Classification accuracy results with a selected set of epoch and training sets for CNN. Each boxplot is based on results from 5 randomly selected training and testing sets.

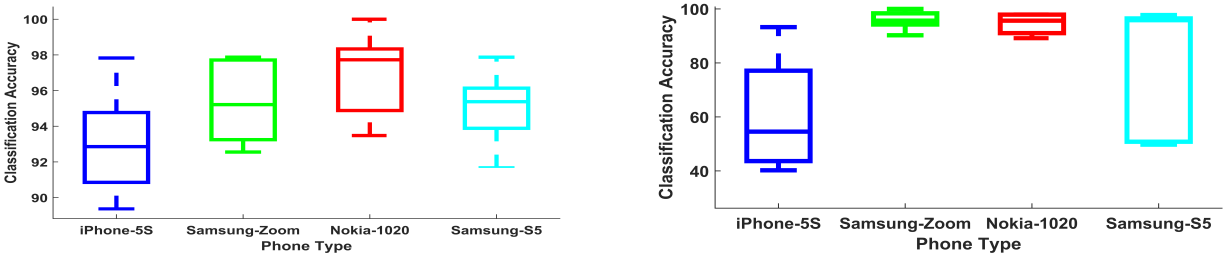
TABLE III

Highest classification accuracy results in terms of our Level 2 classification task, i.e. indoors vs. outdoors data for each phone used.

Phone Type	iPhone 5S	Samsung S5	Nokia 1020	Samsung S4 Zoom
	95.57	97.13	97.91	98.17

Based on the results we concluded that, for Samsung S4 Zoom, for both the indoor and outdoor class, the classification accuracy on average from 5 sets reaches greater than 94%. For iPhone 5S, with the indoor class, the classification accuracy reaches greater than 90% and for the outdoor class, the classification accuracy reaches almost 60%. For Samsung S5, with the indoor class, the classification accuracy reaches greater than 94% and for the outdoor class, the classification accuracy reaches greater than 75%. Based on mean and variance plots, the highest classification accuracy results for the indoor dataset that was classified into close or far distance was achieved for Nokia 1020. For the same classification problem, when using outdoor dataset, the highest accuracy classification results were achieved when using the Samsung S4 Zoom (see Fig. 9).

In Table V, the results are represented based on two scenarios. In scenario 1, the original raw images are selected, and in scenario 2 the detected face images are selected to perform classification using CNN. Based on the results, we concluded that overall the highest classification accuracy results were achieved when the original database was selected to perform the classification. For the indoor class, the results are similar. However, for the outdoor class, scenario



(a) Indoor Class: Close vs. Far Distance

(b) Outdoor Class: Close vs. Far Distance

Fig. 9. Classification results for Level 3 classification for all the phones.

TABLE IV

Classification results from CNN: Close (1m) vs. Far (10m) distance.

Datasets	set 1	set 2	set 3	set 4	set 5
Samsung S4 Zoom					
Indoor	97.87	97.66	95.21	92.55	93.48
Outdoor	95.56	97.82	95.65	100	90.24
iPhone 5S					
Indoor	93.75	97.83	92.86	89.36	91.35
Outdoor	71.73	93.25	44.78	40.22	54.54
Nokia 1020					
Indoor	100	93.48	97.78	97.73	95.35
Outdoor	97.87	95.65	91.67	97.92	89.13
Samsung S5					
Indoor	97.87	94.62	95.38	91.67	95.56
Outdoor	96.06	49.73	95.83	51.16	97.72

1 outperformed scenario 2. We have the valid reason for the outcome results: for the outdoor dataset, the images with background have more features to offer to perform the classification, while, for the indoor class the background is uniform.

TABLE V

Best classification results from CNN: Close (1m) vs. Far (10m) distance.

Phone Type	iPhone 5S	Samsung S5	Nokia 1020	Samsung S4 Zoom
Scenario 1: With Raw Database				
Indoor	97.83	97.87	100	97.87
Outdoor	93.25	97.72	97.92	100
Scenario 2: With Detected Face Database				
Indoor	95.23	91.12	98.40	92.02
Outdoor	89.20	92.83	93.08	91.01

D. Face Matching Results: With and Without Grouping

First, experiments are performed with the original FR system, namely when no grouping is used. Second, we used grouped datasets in terms of cell-phone type, indoors or outdoors, close or far distance. The identification results using LBP-CHI method are summarized in Table VI. Based on the results, we determined that the rank-1 score is improved from 53% (All Data) to 66% - Level 1 (Samsung S4 Zoom), 82% - Level 2 (Samsung S4 Zoom Indoor) and to 95% - Level 3 (Indoor Close) and 71% (Indoor Far). The face matching experiments are represented in Table VI. To show

TABLE VI

Face matching for with and without grouping of data.

Rank-1 Score for - All Data Without Grouping				
LBP	0.53			
VGG-Face	0.52			
Rank 1 Scores using LBP - With Grouping using Proposed CNN				
Level 1 Labeled Class: Cell Type				
Phone Type	**iPhone 5S**	**Samsung S5**	**Nokia 1020**	**Samsung S4 Zoom**
	0.47	0.49	0.50	0.66
Leve 2 Labeled Class: Indoors or Outdoors				
Indoor	0.65	0.65	0.65	0.82
Outdoor	0.29	0.32	0.33	0.49
Leve 3 Labeled Class: Close or Far Distance				
Indoor Close	0.95	0.93	0.98	0.95
Indoor Far	0.34	0.38	0.31	0.71
Outdoor Close	0.44	0.51	0.42	0.52
Outdoor Far	0.15	0.23	0.17	0.46

the impact of grouping or pre-screening of the database on face matching, the first 5 rank identification rates are shown in Table VII.

TABLE VII

Impact on Face Matching Accuracy when either of Data Grouping or databased Pre-Screening is used

Rank	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Face Matching - Without Grouping					
LBP	0.53	0.61	0.66	0.70	0.73
VGG-Face	0.52	0.59	0.63	0.66	0.68
Face Matching - With Grouping (Level 1)					
LBP	0.66	0.74	0.78	0.80	0.82
VGG-Face	0.68	0.74	0.78	0.80	0.83

TABLE VIII

Classification results for extended database (100 subjects) from CNN for all the phones: Indoor vs. Outdoor.

Phone Type	iPhone 5S	Samsung S5	Nokia 1020	Samsung S4 Zoom
CNN-1	90.66	90.23	90.00	93
CNN-2	95.23	95.33	96.33	97

E. Classification and Matching: With Extended Database

We have doubled the database size in terms of all scenarios, including cell phone types, indoors, outdoors, with different poses, at short and far distances and repeated the

classification (Level 2) and face matching experiments. For the classification where 50% of the data is used for training and the rest for testing (without any overlap of subjects), we achieved 93% accuracy (Samasung Zoom). To further improve the classification results, we selected the model with more layers (CNN-2: 12 layers based on MatConvNet) and based on results the classification reaches to 97% as represented in Table VIII.

Based on the face matching results, we determined that the rank-1 identification rate is improved from 82% - VGG Face and 80% - LBP (All Data) to more than 98% - Level 2 (All Phones Indoors 1 meters) as represented in Table IX.

TABLE IX

Face Matching Results for extended database (100 subjects) using LBP matcher for with and without grouping of data.

Rank-1 Score - All Data Without Grouping for 1 meters					
Rank	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**
LBP	0.80	0.83	0.85	0.85	0.86
VGG-Face	0.82	0.83	0.84	0.85	0.86
LBP Matcher for Database With Grouping					
Level 2 Labeled Class: Indoors or Outdoors					
Phone Rank 1		**iPhone 5S**	**Samsung S5**	**Nokia 1020**	**Samsung S4 Zoom**
Indoor		0.99	1.00	0.99	1.00
Outdoor		0.57	0.46	0.56	0.58

V. Conclusion

We investigated the advantages and limitations of our proposed multi-sensor mobile phone adapted convolutional neural network based, hierarchical classification framework, designed to automatically categorize the face images captured under various challenging conditions, before the FR algorithms are used. First, a multi-sensor database is collected (videos) indoors, outdoors, at close and far standoff distances and with different poses using iPhone, Samsung and Nokia phones. The performance of the image classification system is sensitive to face pose variations. To deal with this issue, an algorithmic approach for the selection of frontal face images was generated: first, faces were detected, next a pose estimation algorithm was applied, and, finally, based on the left, right and frontal view, the face images were classified into a frontal or a non-frontal (left and right profile) class. The frontal face image dataset generated was then used to apply our proposed CNN framework for hierarchical classification (Level 1, Level 2 and Level 3).

We trained the CNN model using our challenging multi-sensor mobile phone face database and, for each classification level, a series of tests was performed to select the network parameters that result in high classification accuracy. Our experiments showed that for Level 1 classification, our proposed CNN provides us with significant classification accuracy, i.e. more than 80%. We achieved more than 95% classification accuracy for Level 2 classification in most scenarios tested. We also achieved more than 96% accuracy for Level 3 classification for the indoor class into close or far distance. For the outdoor class into close or far distance, we achieved more than 90% accuracy.

Our face matching results provide important evidence that data grouping in terms of cell-phone type, indoors or outdoors and close or short distance provide significant improvement in the rank-1 identification rate, e.g. the performance is improved from 53% to 82% - Level 2 and to 95% - Level 3 (Indoor close) and 71% (Indoor far) for database labeled as Samsung S4 Zoom based on our proposed architecture.

Experimental results show that CNNs, when properly designed, can be a useful tool in multi-sensor mobile face recognition settings, even though a limited mobile dataset is used (as the one used in this work). In the future, we expect to achieve more accurate results by using a much larger face dataset for training. However, the challenge is the complicated collection scenarios that require many resources when a larger number of subjects is involved.

References

[1] A. Hadid, J. Y. Heikkilä, Olli Silvén and M. Pietikäinen, "Face and eye detection for person authentication in mobile phones", *First ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC'07*, 2007, pp 101–108.

[2] N. Narang, and T. Bourlai, "Face recognition in the SWIR band when using single sensor multi-wavelength imaging systems", *Image and Vision Computing*, vol. 33, pp. 26-43, Jan 2015.

[3] G. Dave, X. Chao and K. Sriadibhatla, "Face Recognition in Mobile Phones", *Department of Electrical Engineering Stanford University, USA*, 2010.

[4] A. Vailaya, M. A. T. Figueiredo, A. K. Jain and H. J. Zhang,"Image classification for content-based indexing", *IEEE Transactions on Image Processing*, vol. 10, 2001, pp 117–130.

[5] T. Bourlai, , N. Mavridis, and N. Narang,"On designing practical long range near infrared-based face recognition systems", *Image and Vision Computing*, vol. 52, 2016, pp 25-41.

[6] L. Gupta, V. Pathangay, et al., "Indoor versus Outdoor scene classification using probabilistic neural network", *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2006, pp 1–10.

[7] A. Payne and S. Singh, "A benchmark for Indoor/Outdoor scene classification", *Pattern Recognition and Image Analysis*, 2005, pp 711–718.

[8] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems*, 2012, pp 1097–1105.

[9] S. Sarkar, V. M. Patel and R. Chellappa, "Deep Feature-based Face Detection on Mobile Devices", *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, Sendai, 2016, pp. 1-8.

[10] G. Levi and T. Hassner,"Age and Gender Classification using Convolutional Neural Networks", *Comput. Vision Pattern Recognition CVPR Workshops*, June, 2010.

[11] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, et al., "Overview of the face recognition grand challenge", *IEEE computer society conference on Computer vision and pattern recognition, CVPR*, vol. 1, 2005, pp 947–954.

[12] J. Whitehill and J. R. Movellan,"A discriminative approach to frame-by-frame head pose estimation", *Int. Conf. Automatic Face and Gesture Recognition*, 2008.

[13] J. Aghajanian and S. Prince,"Face Pose Estimation in Uncontrolled Environments", *Proceedings of the British Machine Vision Conference, BMVC*, vol. 1, September, 2008, pp 1–11.

[14] A. Vedaldi, and K. Lenc, "MatConvNet-convolutional neural networks for MATLAB", *Proceedings of the 23rd ACM international conference on Multimedia*, 2015.

[15] O. M. Parkhi, A. Vedaldi and A. Zisserman, "Deep face recognition", *Proceedings of the British Machine Vision Conference, BMVC*, vol. 1, 2015, pp 6.

[16] J. R. Beveridge, P. Jonathon Phillips, et al.,"The challenge of face recognition from digital point and shoot cameras, *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Arlington, VA, 2013, pp. 18.