ISBN-978-989-9121-03-4
41st ISTANBUL International Conference on "Advances in Science, Engineering & Technology" (IASET-22)
Istanbul (Turkey) May 11-13, 2022

Man Vs Machine: The Ethnic Verification of Pakistani and Non-Pakistani Mouth Features

Shelina Khalid Jilani[1*], Hassan Ugail[1] and Andrew Logan[2]

[1] Centre of Visual Computing, University of Bradford, Bradford, United Kingdom.
[2] Department of Vision Science, Glasgow Caledonian University Glasgow, Scotland.

Abstract: *Complex computer vision experimentations have achieved state-of-the-art classifications on a number of image-based problems. Accordingly, the classification of ethnicity from facial images has also gained significant popularity over the past decades and remains an open challenge. In an attempt to address the problem of demographic, i.e., ethnicity verification, a novel framework is proposed for the Pakistani ethnicity, using a two-tier, human and machine perspective. Fundamental to the processing of demographic information, is the requirement of an ample and balanced training data. Hence, a challenging, criterion-specific dataset has been created. To the best of our knowledge, this two-tier framework, is the first of its kind to attempt ethnicity verification based on information from images of the human mouth alone.*

This work evaluates both Computer and Human-based procedures, for the classification of Pakistani vs non-Pakistani mouth features. The computer-based route utilised a combination of Deep learning (ResNet) and ConvNet pre-trained models. The experimental design is based on (i) feature extraction using deep ResNet and ConvNet-based features and (ii) ethnicity classification using a Linear Support Vector Machine (SVM) algorithm. In the human-based experiments, we quantified the accuracy with which the human visual system descriminates between isolated facial components (mouth) of different ethnicities. Essentially, this part of the study aimed to understand how humans use the visual information contained within face components to make decisions about ethnicity, focussing specifically upon the information used to classify Pakistani features from those of other ethnicities.

Keywords: *Deep Residual Neural Networks; Ethnicity Verification; Machine Learning; Human Visual System; Human Mouth; Pakistani Face Database.*

1. Introduction

The face is a complex visual object and a rich source of information. The face enables humans and machines to extract socially-relevant information relating to an individual such as details of emotional state (Krumhuber et al. 2019), gender (Lin et al. 2016), trustworthiness (Todorov et al. 2009), as well as demographic information relating to ethnicity (Greenwell et al.; Islam et al.; Fu et al. 2014, David Belcar, Petra Grd and Igor Tomičić, 2022). The classification of ethnicity from image-based tasks within a Machine Learning and a human-based framework, helps to provide an insight into understanding how the face and its components are perceived and categorized for a given ethnic label.

The motivation to categorise ethnicity from images of the mouth stems from the role ethnicity plays in technology with applications in isolated face feature-related biometric systems such as iris recognition (Gangwar and Joshi; Dua et al. 2019). Moreover, within the forensic arena, when the face of a perpetrator is disguised and the identification is limited to the region of the mouth and periocular, a framework of this sort can pave the way in leading a criminal investigation and fine-tuning the perpetrator search. Furthermore, demographic-based recognition systems toughen security applications and have been used in crowd race determination (Huh 2018),

and smartphone security (Wu et al. 2014). Hence, the attraction for using demographic information is its multi-disciplinary application.

The charm of studying the problem of ethnicity verification, with deep learning models is because they have worked on-par with humans while in other conditions, surpassed human level performance (Schroff et al.). Ethnicity is a stand-alone and a non-variable trait (Fu et al. 2014), and within the machine learning arena demographic (ethnicity), classification has been reported for a spectrum of face datasets. For example, whole-face images (Hosoi et al. 2004; Jilani et al. 2019; Lu and Jain; Riccio et al. 2012), face profiles (Jilani et al. 2017) including silhouetted face profiles (Tariq et al.), the use of 3D Facial Landmark data in Kendall Shape Space (Lv et al. 2020), and more recently with the use of facial colour and texture (Sallam et al. 2021). In contrast, researchers have also used isolated face components namely the eyes (Mohammad and Al-Ani; Qiu et al.) and the nose (Chang et al. 2006; Song et al. 2009), to address the problem of ethnicity classification. However, the mouth as a single modality for ethnicity classification which remains unchartered territory.

In this paper, we focus on ethnicity as a visible demographic feature in the form of the mouth. The proposed framework is built on a two-tier approach **(i)** automated mechanisms and **(ii)** human discrimination ability, with an aim to conduct comparative analyses between the performance of each. The automated approach makes use of Residual Neural Network (ResNet) and ConvNet-based machine learning algorithms. Deep features are extracted using various deep networks; ResNet-50, ResNet-101, ResNet-152, VGG-Face, VGG-16 and VGG-19. After which, the weights from each model are passed to a Linear Support Vector Machine (SVM) algorithm for a two-class (Pakistani Vs. non-Pakistani) output. For the human experiment novice participants were recruited to partake in a computerized ethnicity classification task, to understand how humans use the visual information contained within mouth images to make decisions about the ethnicity of individuals.

2. Race and Ethnicity

The taxonomy of human populations into sub-groups is not an emerging science. There have been many contributions to the grading of humans 'Homo-Sapiens' into specific assemblies. While some scientists have proposed a categorization system based purely on skin colour (Haller 1995), others have suggested the distinction of people based on the division of the earths continents (Pickering and Hall 1854). Both views are now outdated, since skin colour is fluid and too variable for classifying human populations (Brooks and Gwinn 2010). And secondly, the mass migration of people has meant that a geographical location which once may have represented a person, is now irrelevant. Moreover, there is the factor of gene-pool mixing through interracial relations, which naturally allows for the development of facial characteristics which do not adhere to a specific group.

Amidst Epigenome-Wide Association Studies (EWAS) research, there is an abundance of anthropological research which reports that discrepancies exist between multiple groups of humans, based solely on metrics from the face and its features (Farkas et al. 2005; Jilani et al. 2019). This provides a grounding for computational methods when embarking on ethnicity classification challenges, as it highlights that on a primary level, there are variances amongst both human populations and gender. Within the field of machine learning, the terms race and ethnicity are used interchangeably.

In the context of this research, ethnicity refers to a person's cultural and ancestral background. For example, a person belonging to the South Asian race can belong to either the Bangladesh, Gujarati or Pakistani ethnicity.

3. Literature Review

The recognition of a face by a human is almost effortless compared to a computational model, which must learn fine feature discrepancies of a face. There is literature to suggest that humans have far better face-recall for a familiar, relative to an unfamiliar face and that performance of a human matching unfamiliar faces is error-prone (Ellis et al. 1979; Johnston and Edmonds 2009; Young and Burton 2018). It is suggested that, in humans, the internal representation of a familiar face is derived by computing an average across multiple viewing conditions and considers variances in appearance. Importantly, it does not consider any artificial or extraneous information, which may govern the representation of an unfamiliar face (Jenkins and Burton 2011).

Over the years, a range of publications in the field of cognitive science and psychology, have reported on how the human visual system identifies, interprets and recalls a human face (Bruce and Young 1986; Kanwisher et al. 1997; Kanwisher et al. 1999). A range of aspects relating to human demographic (ethnicity) discrimination

ability has also been reported (Hayward et al. 2008; Burns et al. 2019). Such studies provide context to computational methods, when classifying ethnicity and function as a benchmark for performance analyses.

While anthropometric studies demonstrate differences between human face features, such as the nose (Ugochukwu et al. 2014; Ozdemir and Uzun 2015; Elsamny et al. 2018; Mohammed et al. 2018) and the eyes (Öztürk et al. 2006; Bukhari 2011), there does not appear to be any literature on the mouth in isolation as a determinant of ethnic origin. A potential reason for this may be its variability caused by factors such as cosmetics, facial hair and changes in facial expression. There are, however, computational studies (Heng et al. 2018; Wang et al. 2016; Masood et al. 2018) which have used components of the face including the mouth, to extract distance ratios between the features, to classify ethnicity.

The classification of ethnicity from distances between the internal components of the face (i.e., eyes, nose and mouth) was carried out by Masood et al, 2018. Using a total of 357 images for training and 90 for testing, three ethnic groups were investigated; Mongolian, Caucasian and Negro (described as per the researcher), and a comparative analysis was conducted between 2 methods of analysis. Artificial Neural Network (ANN) and Convolutional Neural Network (CNN). It was concluded that CNN reported higher results with 98.6% accuracy compared to 82.4% for ANN (Masood et al. 2018).

A CNN-based framework was used for predicting fine-grain ethnicity (defined by the authors as a refined category of a human group) for 5 classes: Chinese, Japanese, Korean, Vietnamese and Filipino (Srinivas et al. 2017). Face and face regions from the Wild East Asian Face Dataset (WEAFD), were located and annotated using Dlib. Regions included the eyes, mouth, lower face (inclusive of all three face features as well as the chin), center of the face (nose only) and the left and right side of the face. The methodology employed by the Srinivas et al. 2017 study, made use of two different CNN models, one for the whole face condition (Network A), and another for the sub-region condition (Network B). A key difference between the two models were the feature parameters; Network A had 3,573,028, while Network B had 1,688,550. While the results for fine grain ethnicity were low, 24.0% for experiments 1 and 33.33% for experiment 2. The authors acknowledged the limitations of their training and testing data, was the predominant reason for the low accuracy. Although, the results did highlight that the features learned in the first convolutional layers and ultimately used to discern ethnicity include the shape of the nose, eyebrows and the lips. This provides a rational foundation to suggest that the region of the mouth can be an enriched feature for demographic classification. Although, developing a large dataset is a critical requirement.

From the survey of literature presented above, it does not appear as though any mouth-centered research has been conducted within the field of machine learning and this may be an outcome of unavailable, criterion-specific datasets, until now.

4. Pakistani Face Database (PFD)

A preliminary aim of the present study was to create a face-image database rich in ethnic diversity and high in ecological validity. There is a need for an open-access, multi-racial and multi-ethnicity face database, which includes subjects of both genders. To maximise ecological validity and emulate naturalistic settings (e.g., an international airport), items of religious dress (e.g., headscarf) were not removed. Moreover, many of the female subjects were wearing make-up. Objects without ethnicity-specific associations, on the other hand, were removed such as spectacles, lanyards, high collar jackets and scarfs.

A total of 200 images, split equally in gender and ethnicity (100 per ethnic class and 50 per gender) were used. To ensure the database was not homogenous images of non-Pakistani participants were also included, for example, Caucasian (including Irish), East/Central Asian (Chinese) Middle Eastern (Egyptian, Jordanian, Kurdish, Omani and Syrian), Black (Nigerian and Ugandan) Japanese and Polish participants. While, the ethnicity of each participant was self-assigned, there was a strict selection process for those considered Pakistani. To ensure the database represented pure Pakistani features, for participants of Pakistani heritage, admissibility was reliant on both the maternal and paternal parents being of Pakistani origin.

Images were captured in a darkened room, with a plain white background which was illuminated with a custom-designed lighting set-up, using a capture system called Halo (Jilani et al. 2018). Photographed subjects were asked to look straight ahead and maintain a neutral facial expression. To quantify the contributions of different face features to judgements of ethnicity, the full-face images were then manipulated to create images of individual components (eyes, nose and mouth). To do this, the full-face images were cropped using Adobe Photoshop (CS6 Extended, version 13.0 x 64). To ensure consistency across images, a cropping template was

created, within which each image was aligned and manually cropped to factor-in individual head shape and headscarf type.

For the human-based experiments, the original colour images were converted to greyscale to prevent any colour cues from assisting the participant during the computer-based ethnicity task. While, for the machine learning based experiments, data augmentation was carried out on the original data of 200 mouth images to create a challenging sample of 2,200 images (split equally per class), with a total of 11 images per subject. This led to a minimum of a twelvefold increase in the original data size.

5. Methodology

A two-tier approach is used for the ethnicity classification framework. For the machine learning experiments, our approach was to extract deep mouth features from a dataset of 2,200 images, using ResNet-50, ResNet-101, ResNet-152, VGG-Face, VGG-16 and VGG-19. After feature extraction supervised learning is used to perform a binary classification using a Linear SVM. For the human ethnicity discrimination task, the aim was to use our novel stimuli to quantify the ability of human participants to make ethnicity discrimination judgements based solely on information available from the mouth. Participants took part in a **(i)** Two-Alternative Forced Choice (2-AFC) procedure and, **(ii)** single image procedure.

5.1 Human-based Ethnicity Verification

Participants

A total of 74 students from the University of Bradford (UK) took part in this experiment, of which 38 were Male and 36 were Females. From the male participants, 15 were of British Pakistani (British in the sense that they were born in the UK, hence nationals) ethnicity, while 23 were non-Pakistani, in comparison the ethnic split for the female participants were as follows: 18 British Pakistani females and 18 non-Pakistani females. Collating all this information into the 2-class ethnic groups, a total of 33 Pakistani participants and 41 non-Pakistani participants were recruited. All the participants reported normal or corrected-to-normal vision. Observers gave informed consent, and the study was approved by the Chair of the Biomedical, Natural, Physical and Health Sciences Research Ethics Panel at the University of Bradford.

Design

Participants were categorised by ethnicity (Pakistani/ Non-Pakistani/other ethnicity) and gender (Male/Female). The stimuli on which participants were tested were also organised into categories based on Gender (Male and Female) and Ethnicity (Pakistani and Non-Pakistani).

Apparatus

Images were presented, using routines from the Psychtoolbox (Brainard 1997) on a Sony Trinitron CRT monitor (1024 X 768 at 85Hz) of 65 cd/m^2 mean luminance which was controlled by a Mac mini-computer. 150 equally spaced grey levels were used to maximize contrast linearity. Participants were seated 1.2m from the monitor. Accurate viewing distance was maintained with a chin and forehead rest. At the test distance, the computer monitor subtended 13.4° by 10.1° visual angle; one pixel was 0.018°.

2-AFC Procedure

Ethnicity discrimination accuracy was measured with a custom-designed computerised test. On each trial, participants were presented with two simultaneously visible mouth images. One of the mouth images (target) depicted an individual of Pakistani origin, the other mouth image (distracter) belonged to a different ethnicity. The pairings of target and distracter were kept consistent for all participants.

The participant was asked to indicate the location of the Pakistani mouth (i.e., on the left or right of the screen) of the Pakistani mouth via computer key press. The position (left or right) of the target mouth (i.e., Pakistani origin) was randomly determined prior to each trial. To minimise eye movements and preclude scrutiny of individual featural cues (e.g., lip thickness), mouth images were presented for a maximum of 1 second. After this time, the images were replaced by a low-level, greyscale luminance noise mask until a decision was made. Early responses (<1s) were accepted and resulted in immediate progression to the next trial.

Ethnicity discrimination accuracy was tested for stimuli of each gender in separate blocks. Each block comprised 25 trials (i.e., 50 mouth images in total). At the end of each block, ethnicity discrimination accuracy was calculated as the total number of correct responses as a proportion of the maximum possible score (i.e., 25).

Single Image Procedure

The single image procedure was identical to that outlined above, other than the participants were now shown a single image only. Participants were required to indicate the ethnicity of the presented mouth (Pakistani or non-Pakistani) via keyboard press. Within each single gender block, discrimination accuracy was measured for 25 sets of mouths, see fig. 1.

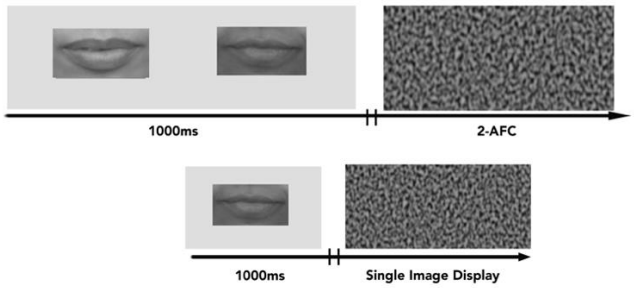

Fig. 1: Procedure schematic. Top: A Two-Alternative Forced Choice (2-AFC) trial for the mouth condition where gender is tested in separate blocks. A pair of target mouth images are shown for 1000ms, followed by a luminance noise mask, presented until the participant responds. Participants must select the mouth which they believe is of Pakistani origin. Bottom: the single image display condition. This trial progressed in the same way as the 2-AFC, although participants were now only presented with a single image on which to base their decision of ethnicity.

5.2 Machine-based Ethnicity Verification

Feature Extraction

All images were resized to 224 × 224 pixels to ensure they conformed to the input criteria of the machine learning models. Feature extraction is a procedure of parameterising an input i.e., the mouth images, with a view to defining the most discriminating features. Given some input, parameters such as image height, image width, colour channels, skin colour image edge information and mouth-shape data are learnt. Essentially, each input image is filtered through different layers of each pre-trained model independently. Each pre-trained model consists of learnable parameters **C** which in turn generate outputs $\mathbf{C}\chi$ for each layer within the network.

The activation of the last pooling layer (2048 dimension) of each of the ResNet models was used, for feature representation. Whereas, for ConvNet-feature representation the Fully Connected layer 7 (FC-7) with 4096 dimension was used. In total, three sets of features were retrieved using ResNet-50, RestNet-101, and ResNet-152 Neural Networks. Similarly, 3 sets of features were extracted using each of the three ConvNet models. It was decided to not use the last output layer of the Fully Connected layer (FC), since it was trained on a set of different data (i.e., objects) compared to the facial feature which are presently used. Moreover, research suggests that the lower layers of the Deep Neural Network are adequate in learning generic features (Karpathy et al.; Chang et al. 2006).

Classification

Having used machine learning-based feature extraction, a linear classifier was employed for binary classification (Pakistani vs non-Pakistani ethnicity). SVMs are a powerful binary classifier and a supervised machine learning model. SVM operates by defining an Optimum Separating Hyperplane (OSH), which typically involves solving an optimization problem between two classes of data (Chierchia et al.; Nalavade and Meshram);

$$\begin{cases} \min\limits_{w,b} & \frac{1}{2} \| w \|^2 \\ \text{with} & y_i . (w . x + b) \geq 1 \end{cases} \tag{1}$$

Model Evaluation

A ***k***-fold cross validation technique was employed to evaluate the performance of all 6 of the pre-trained models. The data ***M*** is split into 10 mutually exclusive parts of the same size, $\boldsymbol{M_1}$ $\boldsymbol{M_2}$ $\boldsymbol{M_3}$…, $\boldsymbol{M_k}$. The SVM algorithm is trained 10 times and during each rotation, 1 set of data is selected for testing while the remaining 9

sets are combined to create a training set. This process is repeated 10 times to ensure each of the ten partitions of data have been used for training and as a testing. During each cycle, relevant features are extracted from the training set and the classification is applied. The outcome of the classification is used to identify mouth images of Pakistani origin.

6. Experimental Results

Human based Discrimination Ability

The study measured face ethnicity discrimination ability in participants of both Pakistani and non-Pakistani ethnicity. Fig. 2 presents ethnicity classification accuracy, measured for Pakistani (dark bars) and non-Pakistani (light bars) participants for the tested mouth images. Pakistani and non-Pakistani participants demonstrated comparable levels of ethnicity discrimination accuracy for the mouth region. The data indicates that human observers can classify ethnicity, based on information from the mouth region alone, at a level which exceeds chance-level performance.

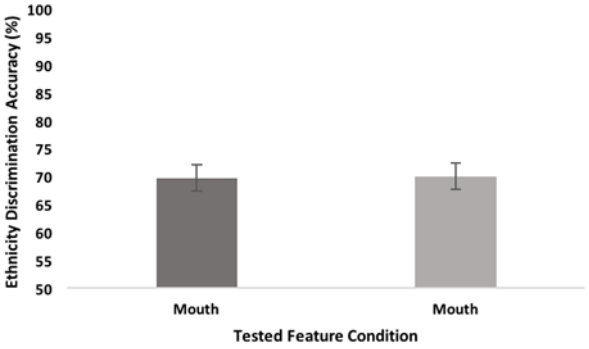

Fig. 2: Ethnicity discrimination accuracy (%) for Pakistani (dark grey column) and non-Pakistani (light column) observers for the classification of ethnicity from mouth images.

To investigate if differences in display (i.e., 2-AFC or single image option) affect ethnicity discrimination ability; 11 participants (8 Pakistani, 3 non-Pakistani, 7 females and 4 males in total) were recalled partaking in the same ethnicity classification task, 12 months after the first session. The reason for participant recall was because the presentation of 2 images gave the participant the opportunity to decide by a process of elimination. The single image condition, on the other hand, challenges the participant to make an ethnicity classification decision on a single image. Moreover, the single option task allows for a fairer comparison with the Machine Learning experiments. This is since the Machine Learning framework processes the input data i.e., mouth images, one at a time.

One reasonable hypothesis was that the single image display would reduce cognitive load by allowing the observer to remain fixated on a single image, instead of looking between two. By doing so, the task could be relatively simpler and thus there would be the possibility of a higher accuracy rate. However, the observed effects are not in accordance with this hypothesis: observers performed less accurately in the single, compared to the 2-AFC condition.

A reduction in sensitivity indicates that in trials where a target mouth is presented alongside a distractor, the participant is more likely to select the correct choice, compared to when a single image is shown. This may be explained by the process of elimination, whereby an observer may make comparisons between the two images and select the one which they consider appearing most like a mouth of Pakistani heritage; something which is not possible when a single image is shown. While ethnicity discrimination of the mouth by the participants was averaged at 71%, for the 2-AFC condition, a decline in performance (64%) was reported when a single image was shown without a distractor. A repeated measures ANOVA confirmed that this difference was significant ($p<0.05$).

Machine based Discrimination Ability

By extracting deep features from the mouth using ResNet-50, ResNet-101 and ResNet-152, we achieved results above 90% for the binary classification of a Pakistani mouth using a challenging dataset of 2,200 images, see Table I.

TABLE I: Performance Accuracy (%) for the Binary Classification of Pakistani Mouth Images Using Residual Learning and ConvNet-Networks

Feature Extraction Model	Linear SVM Classier	Feature Extraction Model	Linear SVM Classifier
ResNet-50	90.0%	**VGG-Face**	79.5%
ResNet-101	**91.5%**	**VGG-16**	**86.2%**
ResNet-152	90.6%	**VGG-19**	84.4%

Based on the assessment of Table I, ResNet-101 marginally outperformed ResNet-50 and ResNet-152, when classifying the mouth as a hallmark of Pakistani ethnicity. In contrast, the ConvNet-based models overall had a lower performance accuracy, though VGG-16 did outperform VGG-F and VGG-19. A potential explanation for the reduction in performance (even though the same data was used), may be linked to the finding that the ResNet models are classed as deep learning models, which are architecturally different and naturally report higher accuracies.

The strength of the 3 ResNet models is further evident especially considering that a lower number of dimensions were used (2048) compared to the ConvNet model which extracted 4096 dimensions. The results show that the mouth can provide reliable, discriminatory information within a machine learning platform, to distinguish between the Pakistani and non-Pakistani class. Also, the deep layered architecture of ResNet-101 is proficient at capturing ethnic traits from a limited source of data. It is surprising that the ConvNet-based models exhibited a decline in performance. Especially considering that they are trained specifically on a dataset of faces. Taking the results into consideration, it is apparent that they are in line with literature proposing that the features of the face are important in determining ethnicity in humans (Hosoi et al.; Masood et al. 2018).

To determine the efficiency of the results reported, performance metrics such as sensitivity and specificity were calculated. Sensitivity is the value of positive cases classified correctly (TPR), and specificity is the value of negative cases correctly rejected (TNR) (Powers 2011). By combining sensitivity and specificity the overall performance accuracy of the classification algorithm can be shown in the form of a confusion matrix. A confusion matrix suggests that for any given experiment, there can be 1 of 4 outcomes. Fig. 3 shows the confusion matrix for the highest achieving Residual learning model (ResNet-101) and ConvNet-based model (VGG-16), in addition to visual representations of all the models, in the form of a Receiver Operating Characteristics (ROC) Graph.

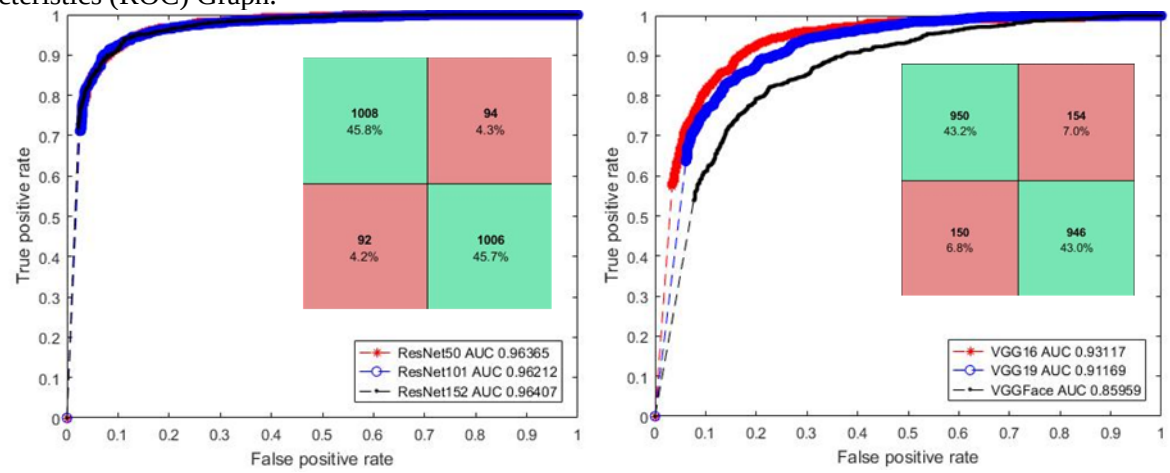

Fig. 3: (Right) Confusion matrix for ResNet-101 (There was a total of 1,100 images per class (Pakistani and non-Pakistani)). The top-left green box denotes True-Positive (TP) classification for the first class whereas the lower-right green box denotes the True Negative (TN) for the second class. The ROC depicts the classification of the Pakistani mouth, using ResNet-50/101/152 features. The True Positive Rate (TPR) is the value of cases where the data has been correctly identified and the False Positive Rate (FPR) is the value of cases where data has been incorrectly identified as a positive. (Left) Confusion matrix for VGG-16 (There was a total of 1,100 images per class (Pakistani and non-Pakistani)). The top-left green box denotes TP classification for the first class, whereas the lower-right green box denotes the TN for the second class. The

ROC depicts the classification of the Pakistani mouth, using VGG-F/16/19 features. TPR is the value of cases where the data has been correctly identified and FPR is the value of cases where data has been incorrectly identified as a positive.

7. Discussion

The discriminative nature of the mouth as a marker of ethnicity for the Pakistani population, was investigated. From the results, it is apparent that deep learning algorithms outperformed ConvNet-based models, on the classification of isolated mouth images. Importantly, the results reported within this paper expand on the current understanding of ethnicity determination since, to the best our knowledge, the application of deep learning algorithms for the ethnic classification of Pakistani ethnicity based specifically on the mouth, has not been attempted previously. Moreover, while there is no obvious literature which reports on the informative nature of the mouth as an ethnic trait, within a two-tier machine learning framework. There is some literature available on the incorporation of the mouth as an enriched feature for ethnicity verification.

Given the novelty of our experimental stimuli and the specificity of our approach (ethnicity determination based on mouth features only), the opportunity to compare our results with previous reports is extremely limited. Importantly, the results are in support of the proposal that the mouth is an enriched face feature and an advantageous feature for determining Pakistani ethnicity.

The human ethnicity discrimination task reveals the importance of the mouth as a feature which enables humans to make ethnicity judgements. In brief, the results highlight that while there is no difference in ethnicity discrimination accuracy (based on mouth information alone) for Pakistani and non-Pakistani participants (i.e., no 'own-race bias'), an intermediate performance accuracy was found; approximately 70%, which is considerably higher than chance level (50%). With regards to the difference in performance accuracy in the single vs 2-AFC condition, it is evident that the results are significantly higher for the 2-AFC condition, relative to the single image display option.

In contrast, when analysing the machine learning results ResNet-101 achieved a performance accuracy of 91.5%, followed closely by ResNet-50 and ResNet-152 with classification at 91.0% and 90.46% respectively. The closeness in performance between the deep learning models is consistent with results presented by He et al., (He et al.) The 50/101/152-layer ResNets report high and accurate results given the considerably increased depth of the architectures. While the results for the ConvNet-based models are not as significant as those achieved by Residual learning, they demonstrate the robustness of machine learning algorithms for ethnicity verification from restrictive date. To the best of our knowledge, the mouth has not been previously investigated for a binary ethnicity classification task, using either ResNet or ConvNet-features.

It may seem surprising that the mouth is so informative, considering it undergoes significant changes in shape and size during portraying facial expressions, however subjects were requested to keep a neutral facial expression during image photography. Moreover, the human discrimination task data clearly demonstrates that despite the lack of colour information (since all the images were presented in grayscale), observers were able to perform ethnicity categorisation at a level significantly better than would be expected, based on chance performance.

8. Conclusion

The process of ethnic classification from the mouth as an isolated face feature has been conducted from a two-tier approach, which incorporates machine learning algorithms and then tests human performance. The binary classification task used a challenging yet novel dataset of mouth images belonging to an ethnically diverse group of participants. ResNet and ConvNet-features were extracted, and the weights were forwarded to a Linear SVM classifier, for binary classification. A performance accuracy of 91.5% was concluded for ResNet-101, while VGG-16 achieved an accuracy of 86.2%. Interestingly, human performance was not on-par with the machine learning algorithms and an average classification of 70% was concluded. Human discrimination ability is robust mainly when there is the need to discriminate between two images, in contrast to when a single image is shown (which is reported to be cognitively challenging). In conclusion, the findings of the current research demonstrate the richness of the mouth as a hallmark for the Pakistani ethnicity, and that the ability of machine learning algorithms exceeds human discrimination ability.

References

[1] Aworinde, H. O. and Onifade, O. F. W. (2019) A Soft Computing Model of Soft Biometric Traits for Gender and Ethnicity Classification. *International Journal of Engineering and Manufacturing* 9 (2), 54.
https://doi.org/10.5815/ijem.2019.02.05

[2] Belcar, D., Grd, P. and Tomičić, I., 2022, February. Automatic Ethnicity Classification from Middle Part of the Face Using Convolutional Neural Networks. In *Informatics* (Vol. 9, No. 1, p. 18). MDPI.
https://doi.org/10.3390/informatics9010018

[3] Brainard, D. H. (1997) The psychophysics toolbox. *Spatial vision* 10, 433-436.
https://doi.org/10.1163/156856897X00357

[4] Brooks, K. R. and Gwinn, O. S. (2010) No role for lightness in the perception of black and white? Simultaneous contrast affects perceived skin tone, but not perceived race. *Perception* 39 (8), 1142-1145.
https://doi.org/10.1068/p6703

[5] Bruce, V. and Young, A. (1986) Understanding face recognition. *British journal of psychology* 77 (3), 305-327.
https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

[6] Bukhari, A. A. (2011) The distinguishing anthropometric features of the Saudi Arabian eyes. *Saudi Journal of Ophthalmology* 25 (4), 417-420.
https://doi.org/10.1016/j.sjopt.2011.05.004

[7] Burns, E. J., Tree, J., Chan, A. H. D. and Xu, H. (2019) Bilingualism shapes the other race effect. *Vision research* 157, 192-201.
https://doi.org/10.1016/j.visres.2018.07.004

[8] Chang, K. I., Bowyer, K. W. and Flynn, P. J. (2006) Multiple nose region matching for 3D face recognition under varying facial expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (10), 1695-1700.
https://doi.org/10.1109/TPAMI.2006.210

[9] Chierchia, G., Pustelnik, N. and Pesquet, J. C. Random primal-dual proximal iterations for sparse multiclass SVM. 2016. IEEE.
https://doi.org/10.1109/MLSP.2016.7738833

[10] Dua, M., Gupta, R., Khari, M. and Crespo, R. G. (2019) Biometric iris recognition using radial basis function neural network. *Soft Computing*, 1-15.
https://doi.org/10.1007/s00500-018-03731-4

[11] Ellis, H. D., Shepherd, J. W. and Davies, G. M. (1979) Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception* 8 (4), 431-439.
https://doi.org/10.1068/p080431

[12] Elsamny, T. A., Rabie, A. N., Abdelhamid, A. N. and Sobhi, E. A. (2018) Anthropometric Analysis of the External Nose of the Egyptian Males. *Aesthetic plastic surgery* 42 (5), 1343-1356.
https://doi.org/10.1007/s00266-018-1197-8

[13] Farkas, L. G., Katic, M. J. and Forrest, C. R. (2005) International anthropometric study of facial morphology in various ethnic groups/races. *Journal of Craniofacial Surgery* 16 (4), 615-646.
https://doi.org/10.1097/01.scs.0000171847.58031.9e

[14] Fu, S., He, H. and Hou, Z.-G. (2014) Learning race from face: A survey. *IEEE transactions on pattern analysis and machine intelligence* 36 (12), 24832509.
https://doi.org/10.1109/TPAMI.2014.2321570

[15] Gangwar, A. and Joshi, A. DeepIrisNet: Deep iris representation with applications in iris recognition and cross-sensor iris recognition. 2016. IEEE.

https://doi.org/10.1109/ICIP.2016.7532769

[16] Greenwell, C., Spurlock, S., Souvenir, R. and Jacobs, N. GeoFaceExplorer: Exploring the geo-dependence of facial attributes. 2014. ACM.

https://doi.org/10.1145/2676440.2676443

[17] Haller, J. S. (1995) *Outcasts from evolution: Scientific attitudes of racial inferiority, 1859-1900.* SIU Press.

[18] Hayward, W. G., Rhodes, G. and Schwaninger, A. (2008) An own-race advantage for components as well as configurations in face recognition. *Cognition* 106 (2), 1017-1027.

https://doi.org/10.1016/j.cognition.2007.04.002

[19] He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition. 2016.

https://doi.org/10.1109/CVPR.2016.90

[20] Heng, Z., Dipu, M. and Yap, K.-H. Hybrid supervised deep learning for ethnicity classification using face images. 2018. IEEE.

https://doi.org/10.1109/ISCAS.2018.8351370

[21] Hosoi, S., Takikawa, E. and Kawade, M. Ethnicity estimation with facial images. 2004. IEEE.

[22] Huh, J.-H. (2018) PLC-integrated sensing technology in mountain regions for drone landing sites: focusing on software technology. *Sensors* 18 (8), 2693.

https://doi.org/10.3390/s18082693

[23] Islam, M. T., Workman, S., Wu, H., Jacobs, N. and Souvenir, R. Exploring the geo-dependence of human face appearance. 2014. IEEE.

https://doi.org/10.1109/WACV.2014.6835989

[24] Jain, A. K., Dass, S. C. and Nandakumar, K. Can soft biometric traits assist user recognition? 2004. Vol. 5404. International Society for Optics and Photonics.

[25] Jenkins, R. and Burton, A. M. (2011) Stable face representations. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366 (1571), 1671-1683.

https://doi.org/10.1098/rstb.2010.0379

[26] Jilani, S. K., Ugail, H., Bukar, A. M. and Logan, A. On the Ethnic Classification of Pakistani Face using Deep Learning. 2019. IEEE.

https://doi.org/10.1109/CW.2019.00039

[27] Jilani, S. K., Ugail, H., Bukar, A. M., Logan, A. and Munshi, T. A Machine Learning Approach for Ethnic Classification: The British Pakistani Face. 2017. IEEE.

https://doi.org/10.1109/CW.2017.27

[28] Jilani, S. K., Ugail, H., Cole, S. and Logan, A. J. (2018) Standardising the Capture and Processing of Custody Images.

https://doi.org/10.9734/CJAST/2018/44481

[29] Jilani, S. K., Ugail, H. and Logan, A. (2019) Inter-Ethnic and Demic-Group Variations in Craniofacial Anthropometry: A Review. *PSM Biological Research* 4 (1), 6-16.

[30] Johnston, R. A. and Edmonds, A. J. (2009) Familiar and unfamiliar face recognition: A review. *Memory* 17 (5), 577-596.

https://doi.org/10.1080/09658210902976969

[31] Kanwisher, N., McDermott, J. and Chun, M. M. (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience* 17 (11), 4302-4311.

https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

[32] Kanwisher, N., Stanley, D. and Harris, A. (1999) The fusiform face area is selective for faces not animals. *Neuroreport* 10 (1), 183-187.

https://doi.org/10.1097/00001756-199901180-00035

[33] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. Large-scale video classification with convolutional neural networks. 2014.

https://doi.org/10.1109/CVPR.2014.223

[34] Krumhuber, E., Küster, D., Namba, S., Shah, D. and Calvo, M. (2019) Emotion recognition from posed and spontaneous dynamic expressions: Human observers vs. machine analysis.

https://doi.org/10.31234/osf.io/x5uht

[35] Lin, F., Wu, Y., Zhuang, Y., Long, X. and Xu, W. (2016) Human gender classification: a review. *International Journal of Biometrics* 8 (3-4), 275-300.

https://doi.org/10.1504/IJBM.2016.082604

[36] Lu, X. and Jain, A. K. Ethnicity identification from face images. 2004. International Society for Optics and Photonics.

[37] Lv, C., Wu, Z., Wang, X., Dan, Z. and Zhou, M. (2020) Ethnicity classification by the 3D Discrete Landmarks Model measure in Kendall shape space. *Pattern Recognition Letters* 129, 26-32.

https://doi.org/10.1016/j.patrec.2019.10.035

[38] Masood, S., Gupta, S., Wajid, A., Gupta, S. and Ahmed, M. (2018) Prediction of human ethnicity from facial images using neural networks. *Data Engineering and Intelligent Computing.* Springer. 217-226.

https://doi.org/10.1007/978-981-10-3223-3_20

[39] Mohammad, A. S. and Al-Ani, J. A. Convolutional Neural Network for Ethnicity Classification using Ocular Region in Mobile Environment. 2018. IEEE.

https://doi.org/10.1109/CEEC.2018.8674194

[40] Mohammed, I., Mokhtari, T., Ijaz, S., Ngaski, A. A., Milanifard, M. and Hassanzadeh, G. (2018) Anthropometric study of nasal index in Hausa ethnic population of northwestern Nigeria. *Journal of Contemporary Medical Sciences* 4 (1).

https://doi.org/10.22317/jcms.03201806

[41] Nalavade, K. and Meshram, B. B. Data Classification Using Support Vector Machine. 2012. Vol. 2.

[42] Ozdemir, F. and Uzun, A. (2015) Anthropometric analysis of the nose in young Turkish men and women. *Journal of Cranio-Maxillofacial Surgery* 43 (7), 1244-1247.

https://doi.org/10.1016/j.jcms.2015.05.010

[43] Pickering, C. and Hall, J. C. (1854) *The races of man: and their geographical distribution.* George Bell.

[44] Powers, D. M. (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

[45] Qiu, X., Sun, Z. and Tan, T. Global texture analysis of iris images for ethnic classification. 2006. Springer.

https://doi.org/10.1109/ICIP.2007.4379178

[46] Riccio, D., Tortora, G., De Marsico, M. and Wechsler, H. EGA—Ethnicity, gender and age, a pre-annotated face database. 2012. IEEE.

https://doi.org/10.1109/BIOMS.2012.6345776

[47] Schroff, F., Kalenichenko, D. and Philbin, J. Facenet: A unified embedding for face recognition and clustering. 2015.

https://doi.org/10.1109/CVPR.2015.7298682

[48] Simonyan, K. and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[49] Song, S., Ohnuma, K., Liu, Z., Mei, L., Kawada, A. and Monma, T. (2009) Novel biometrics based on nose pore recognition. *Optical Engineering* 48 (5), 057204.

https://doi.org/10.1117/1.3130242

[50] Srinivas, N., Atwal, H., Rose, D. C., Mahalingam, G., Ricanek, K. and Bolme, D. S. Age, gender, and fine-grained ethnicity prediction using convolutional neural networks for the East Asian face dataset. 2017. IEEE. https://doi.org/10.1109/FG.2017.118

[51] Tariq, U., Hu, Y. and Huang, T. S. Gender and ethnicity identification from silhouetted face profiles. 2009. IEEE. https://doi.org/10.1109/ICIP.2009.5414117

[52] Todorov, A., Pakrashi, M. and Oosterhof, N. N. (2009) Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition* 27 (6), 813-833. https://doi.org/10.1521/soco.2009.27.6.813

[53] Ugochukwu, E. G., Ijomone, O. M., Ude, R. A. and Nandi, E. M. (2014) Anthropometric analysis of the nose of the Ejagham ethnic group in Cross River State, Nigeria. *Annals of Bioanthropology* 2 (1), 13. https://doi.org/10.4103/2315-7992.143401.

[54] Wang, W., He, F. and Zhao, Q. Facial ethnicity classification with deep convolutional neural networks. 2016. Springer. https://doi.org/10.1007/978-3-319-46654-5_20

[55] Wu, L., Du, X. and Fu, X. (2014) Security threats to mobile multimedia applications: Camera-based attacks on mobile phones. *IEEE Communications Magazine* 52 (3), 80-87. https://doi.org/10.1109/MCOM.2014.6766089

[56] Young, A. W. and Burton, A. M. (2018) Are we face experts? *Trends in cognitive sciences* 22 (2), 100-110. https://doi.org/10.1016/j.tics.2017.11.007

[57] Öztürk, F., Yavas, G. and Inan, U. U. (2006) Normal periocular anthropometric measurements in the Turkish population. *Ophthalmic epidemiology* 13 (2), 145-149. https://doi.org/10.1080/09286580500507220