Hybrid Supervised Deep Learning for Ethnicity Classification using Face Images

Zhao Heng*, Manandhar Dipu†, Kim-Hui Yap‡
School of Electrical and Electronic Engineering
Nanyang Technological University
Email:*zhao0248@e.ntu.edu.sg,†dipu002@e.ntu.edu.sg, ‡ekhyap@ntu.edu.sg

***Abstract*—Ethnicity information is an integral part of human identity, and a useful identifier for various applications ranging from video surveillance, targeted advertisement to social media profiling. In recent years, Convolutional Neural Networks (CNNs) have shown state-of-the-art performance in many visual recognition problems. Currently, there are a few CNN-based approaches on ethnicity classification [1], [2]. However, the approaches suffer from the following limitations: (i) most face datasets do not include ethnicity information, and those with ethnicity information are typically small to medium in size, thereby they do not provide sufficient samples for training of CNNs from the scratch, and (ii) the CNN methods often treat ethnicity classification as a multi-class classification where the likelihood of each class label is generated. However, it does not utilize the intermediate activation functions of CNNs which provide rich hierarchical features to assist in ethnicity classification. In view of this, this paper proposes a new hybrid supervised learning method to perform ethnicity classification that uses both the strength of CNN as well as the rich features obtained from the network. The method combines the soft likelihood of CNN classification output with an image ranking engine that leverages on matching of the hierarchical features between the query and dataset images. A supervised Support Vector Machine (SVM) hybrid learning is developed to train the combined feature vectors to perform ethnicity classification. The performance of the proposed method is evaluated using a dataset consisting of Bangladeshi, Chinese and Indian ethnicity groups, and it outperforms the state-of-the-art methods [2], [3] by up to 3% in recognition accuracy.**

***Index Terms*—Deep-learning, CNNs, ethnicity classification, face images**

I. Introduction

Ethnicity plays a vital role in biometric recognition. Classification of people according to ethnicity, nationality and race has great application impact on surveillance, advertisement and social media profiling. The rapidly changing populations constantly re-define the collective identities of regions, nations and races. Although challenging as it is, the practice of classification of people according to race, ethnicity has made great contributions in many countries' national censuses and national security [4]. The categorization process is also a significant topic in social science. It is proven to be useful in health care, educational and socioeconomic status study [5]. It also has commerical values in market research, especially in a multi-ethnic nation.

In the past few years, several works [6]–[10] have tried to use conventional visual recognition methods to perform ethnicity classification on different datasets. Most of the methods are based on face images, which provide the quickest and most direct way to evaluate on a persons ethnicity or demographic information. Siyao Fu et al. [6] conducted a thorough survey on some state-of-art advances in face-race perception, principles, algorithms and applications, as well as some feature representation models. Xiaoguang Lu et al. [7] studied classification between Asian and non-Asian using linear discriminant analysis (LDA) on a database with 2,630 face images from 263 subjects. M.A. Borgi et al. [8] used a new approach called multi-regularized learning (MRL), which derived from multi-stage learning (MSL) and multi-task features learning (MTFL) to apply on race recognition problem. S. Hosoi et al. [9] tried using Gabor wavelets transformation and retina sampling to train a Support Vector Machine (SVM) classifier to construct human-friendly machine interfaces for ethnicity classification. Zhiguang Yang et al. [10] applied local binary pattern (LBP) operator to classify age, gender and ethnicity. However, these conventional approaches use low-level feature representation which may not effective enough to achieve strong performance.

The last few years have witnessed dramatic development and success in computer vision through deep learning. Based on the deep architecture, Convolutional Neural Networks (CNNs) have been proved to achieve superior performance in many vision-related tasks, including object detection [11], image classification [12], or semantic segmentation [11]. The features learned through these deep networks provide a robust image representation, and are shown to achieve good performance for vision related-tasks which even human may find as difficult.

Some recent works have exploited the use of deep learning techniques and architectures for ethnicity classification. Amogh Gudi [13] presented to recognize semantic facial features using deep learning. Recently, Haoxuan Chen et al. [1] implemented several approaches including a K-nearest neighbour algorithm, a SVM classifier, a two-layer neural network and a CNN to train a classifier to predict Chinese, Japanese and Korean. It achieved an overall prediction of 89.2% in the 3-class classification. Wei Wang et al. [2] performed their work on ethnicity classification using CIFAR-10 network, which focused on classification of black and white people, Chinese and non-Chinese people, and Han Chinese and Uyghurs people. The experiment used both public and self-collected datasets. Masood et al. [14] applied a CNN to predict three ethnicities: Mongolian, Caucasian and Negro consisting of 447 images collected from the FERET database.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

The method takes advantages of the geometric features and colour information obtained from the neural network. The result showed improvement compared to a multiplayer perceptron (MLP) network. However, their categories are rather distinctive, such as differentiation between black and white people, European and East Asian people. The results are less meaningful since each category can be easily differentiated by human/machine without much difficulty. It is generally more meaningful to classify people with close geographical relations.

The CNN-based approaches [1], [2], [13] mentioned above perform the classification based on the last fully-connected layer with softmax classifier. These methods, however, do not make use of the rich intermediate features learned from the network, and hence the information available may not be used. Moreover, for cases where only small datasets are available for training, the classifier may not learn sufficiently and this can give rise to poor fitting. As a result, the classification results obtained using such classifiers are not robust .

In order to mitigate the issues mentioned above, we propose a hybrid supervised learning system that takes advantages of rich intermediate features from CNN activations to enhance the classification accuracy. The proposed framework consists of two components: a CNN-based classifier and an image ranking engine that makes use of features extracted from the intermediate activations of CNNs. A new hybrid feature is formed that exploits both the strength of CNN classification as well as the rich features of the networks. The experiments are performed on a dataset of close demographic ethnicity. It shows that the proposed method outperforms other competitive method, which demonstrates the effectiveness of the proposed hybrid supervised learning.

The rest of the paper is structured as follows. Section II outlines the overall framework of the proposed method. Section III explains the methodology of the proposed method in details, Section IV presents the experimental results and discussion, which is followed by conclusion in Section V.

II. Overview of the proposed framework

The overview of the proposed hybrid supervised learning for ethnicity classification is given in Fig. 1. The proposed framework is a two-path image classification system using information from both CNN based classifier and image ranking engine to realize the hybrid supervised learning for ethnicity classification. The CNN-based classifier is trained using a deep network with a final softmax layer. On the other hand, the image ranking engine extracts the features from the intermediate layer and top similar images are retrieved. Next, a majority voting histogram are combined to form a new aggregated feature. A support vector machine (SVM) classifier is trained using the new hybrid features. This strategy is effective as it incorporates the rich hierarchical features from intermediate CNN activations and soft probabilistic outputs of CNN to enhance the classification result. The experimental results clearly show the proposed method outperforms other competitive methods

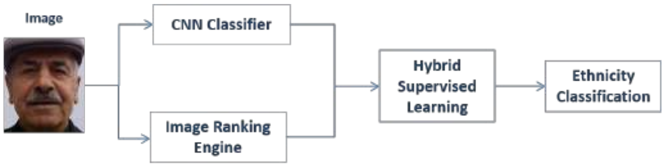

Fig. 1: Pipeline of the proposed method

III. Methodology

The proposed model consists of three main modules, that are (a) CNN-based classifier, (b) Image ranking engine, and (c) Hybrid supervised learning that uses both classification and ranking engine results to predict final ethnicity label. The following sections describe each module in details.

A. CNN-based classifier

CNNs have shown good performance in various classification tasks. A typical CNN consists of several convolutional layers, followed by a few fully connected layers and a softmax layer which producing a distribution over the trained categories. Fig. 2 shows the architecture of the network used for CNN classifier, which takes a 224×224 face image as input and predicts its ethnicity category using VGG-16 network architecture. The network consists of 13 convolutional layers and 3 fully-connected layers.

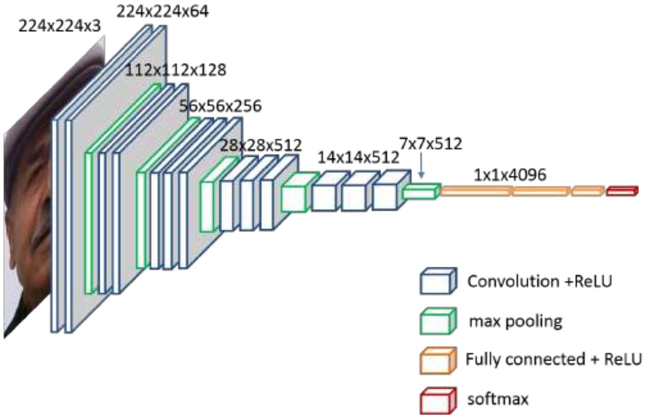

Fig. 2: The architecture of the proposed CNN classifier

The score y'_k for k^{th} class is computed using the last fully-connected layer (4096D for VGG16-Net) feature vector, the equation is shown below (1).

$$y'_k = \sum_{m=1}^{4096} x_k \cdot w_{m,k} + b_m \tag{1}$$

x_k is the output of the last fully-connected layer for k^{th}. The weights $w_{m,k}$ and biases b_m are obtained by fine-tuning the network. Next, the output softmax layer produces the probability distribution for all N classes. The normalized probability score for each class is calculated using (2) as follows:

$$y_k = \frac{exp(y'_k)}{\sum_{i=1}^{N} exp(y'_i)} \tag{2}$$

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

Therefore, we can obtain the class score y_k for k_{th} class using (2), which consists of the likelihood of the target image belongs to the k_{th} class. From the output of the softmax layer, we can extract a score vector $1 \times N$, denoted as $\mathbf{S_{cnn}} = [s_1, s_2, s_k \cdots, s_N]$, where s_k represents the probability score for k_{th} class and N is the total number of classes.

B. Image ranking engine

Ideally, CNN-based classifier shoud provide good classification decision after sufficient training. However, in some cases, when the dataset is small and contains similar images, the result may not be that robust. In light of this, we propose an image ranking engine based on extracted features to improve the prediction accuracy. Image features generated from various Content-Based Image Retrieval (CBIR) models are efficient tools to represent an image. Some popular approaches like Bag-of-Words model [15] have achieved good performance. In this paper, we propose to extract rich features from intermediate activation of the convolutional layer to strengthen the training of classifier.

Specifically, we pool features from the activation of the intermediate convolution layers which are 3D tensors. Following the recent works in instance search based on CNN features [16]–[19], we utilize the 3D features with size of $W \times H \times K$ dimensions, where K is the number of output channels and $W \times H$ represents the width and the height of feature map. This 3D tensor is represented as stack of 2D feature map R_i where $i = \{1, 2, ..., K\}$. The 2D map R_i is the filter response over a set of all valid spatial locations ($p \in L$) for the i^{th} feature channel.

$$F_{L,i} = \max_{p \in L} R_i(p), \qquad (3)$$
$$\text{where,} \;\; i = 1, ..., K$$

p denotes a particular position of image, and $R_i(p)$ is the feature response at p^{th} position . Therefore, a feature vector $\mathbf{F}_L$ can be constructed by a spatial max-pooling over all locations which represents maximum activations of convolutions (MAC) [20]. These features are further normalized using L_2-norm and used for searching visually similar images. Given a query image, we retrieve the top-k most similar images from the database images using cosine similarity measure.

$$Sim(\mathbf{Q}, \mathbf{D}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2}\sqrt{\sum_{i=1}^{n} d_i^2}} \qquad (4)$$

where q_i and d_i are components of the two image feature vectors $\mathbf{Q}$ and $\mathbf{D}$.

The top retrieved list of images is used to predict the class label of the test image. A simplest way is to pass the label of top-1 retrieved to the test image. However, this simple strategy is too harsh and the prediction may not be always correct. In view of this, we make use of the top-k retrieved images to make the prediction more accurate. Based on each retrieved image in top-k, we cast a vote for the particular class, which is weighted by a weight inversely proportional to the rank of the retrieved image. We refer this strategy as weighted majority voting. Hence, by using ranked images and their class labels, a histogram is generated which forms a new feature vector for the test image.

Mathematically, let $\mathbf{C} = \{C_1, \cdots, C_i, \cdots, C_N\}$ be the ethnicity classes, where N is the total number of the classes. Let $\mathbf{I_r} = \{I_{r1}, I_{r2}, \cdots, I_{rm}, \cdots, I_{rk}\}$ be the top-k retrieved images using the ranking engine. The weighted histogram with majority voting, $\mathbf{H_{ranking}} = \{h_1, \cdots, h_i \cdots, h_N\}$ is constructed, where each component h_i is computed as follows.

$$h_i = \sum_{m=1}^{k} w(m) \cdot 1 \cdot [I_{rm} \in C_i] \qquad (5)$$

where, $1 \cdot [I_{rm} \in C_i]$ is an indicator function which equals to 1 when m^{th} retrieved image falls into the class C_i. The parameter w in (5) is the weight given to retrieved images based on their ranks. We experiment various weighting strategies for w_i in the experiments. Note that the intermediate feature based retrieval essentially encodes the rich intermediate activations into the images feature via majority voting strategy, which in turn is used for the prediction of the test image's ethnicity.

C. Hybrid supervised deep learning

This section describes a hybrid learning strategy for enhanced ethnicity classification. The information obtained using CNN-based classifier and the ranking engine are combined to learn the class label of the query image. In particular, the vectors from class probability scores $\mathbf{S_{cnn}}$ and weighted histogram $\mathbf{H_{ranking}}$ are concatenated to form a new hybrid feature vector $\mathbf{I} = [\mathbf{S_{cnn}} \quad \mathbf{H_{ranking}}]_{1 \times 2N}$. This hybrid features fuses information from CNN-based classifier and rich information from intermediate CNN activations. The hybrid features are used to train a SVM classifier. SVM employs an iterative training process to construct the optimal hyperplane, it uses the basic classifier $w^T\mathbf{I} + b$ to find the largest possible distance $\frac{1}{2}w^T w$ to separate data. Take outliers into consideration, the purpose of SVM is to minimize the following objective function:

$$\frac{1}{2}w^T w + C \sum_{i=1}^{N} \xi_i \qquad (6)$$

subject to the constraints:

$$y_i(w^T\mathbf{I_i} + b) \geq 1 - \xi_i \; and \; \xi \geq 0, i = 1, ..., N \qquad (7)$$

where w is the vector of coefficients, y_i is the ethnicity label for data point $\mathbf{I_i}$. $\xi_i \geqslant 0$ is the slack variable, it defines the functional margin, which is the allowable deviation of $\mathbf{I_i}$, however, ξ_i is also subjected to have a minimized sum for all data points in order to have an optimal hyperplane. C is the capacity constant which ensures the objective function will have the hyperplane with largest margin and smallest data deviation. By implementing this primal formulation of SVM, the hyperplane which separates different ethnicity classes can be constructed. The advantage of integrated information in the hybrid features is demonstrated via experiments in the following section.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

IV. Experiments

We perform experiments on a newly collected dataset for evaluating the proposed approach. It consists of three geographically close ethnicity. Each image is taken under good daylight with centered upright face. The dataset contains 1000 images of 1000 Bangladeshi people, 1520 images of 1042 Chinese people and 1078 images of 1009 Indian people. All images have similar resolution and are cropped to keep the face region only.

A. Experimental Setup

To test the classification accuracy, the total 3598 images are separated into 3 different subsets before training the classifier: Training set: 46%, Validation set: 16% and Testing set: 38%. The proposed method uses VGG-16 as the CNN network architecture model for classification. The same architecture is also used for image ranking engine to extract intermediate convolutional layer features where the $conv5\_3$ layer, the last convolutional layer with 512 dimensions of feature, is used to extract MAC descriptor. The experiments are carried out with Nvidia Tesla K40m GPU during training and testing. Fig. 3 shows some sample training and testing images from the dataset.

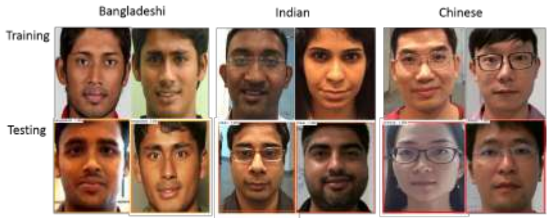

Fig. 3: Sample images

B. Model evaluation on ethnicity classification

The model is trained using the training and validation sets on VGG16 network which is initialized with pre-trained weights using ImageNet. The trained CNN model are then used to obtain the classification result along with the score vectors $\mathbf{S}_{cnn}$ for all training and testing images. During the testing, the intermediate features are computed on-the-fly which are used for image ranking. We experiment on different configurations of top-k retrieved images and weighting strategies $w(m)$. Specifically, we employ equal weighting scheme and linear weighting scheme and vary the parameter k in [2,20]. We use the top-5 retrieved images with equal weighting, which achieves good accuracy among all combinations. Hence, we use these values in the following experiments. The concatenated hybrid features are then used to determine the final class label.

C. Results

The hybrid supervised learning result on the total 1355 testing images achieves an overall accuracy of 95.2%. From the result, it can be seen that the proposed method of hybrid supervised deep learning achieves a better performance. Performance results of different methods are summarized in the following table for comparison:

Recognition Accuracy				
	Bangladeshi	Chinese	Indian	Overall
Faster R-CNN [3]	86%	99.5%	90.6%	93.5%
Wang's method [2]	84.7%	99.2%	88.5%	92.4%
Proposed method	**88.4%**	**100%**	**93.1%**	**95.2%**

Table 1: Performance Comparison of different methods

The results in Table 1 show that the proposed method outperforms that of Faster R-CNN [3] and Wang's method [2]. Specifically, the proposed method can recognize all images of Chinese ethnicity correctly and yield a 2-3% improvement for Bangladeshi and Indian over the Faster R-CNN [3] and 3-4% improvement over the Wang's method [2]. The superior results of the proposed method show that the intermediate features extracted from the convolutional layers of the CNN carry rich information that can be used to complement classification result to enhance the performance. The results clearly show the effectiveness of the proposed method in ethnicity classification.

V. Conclusion

In this paper, we have presented a new approach on ethnicity classification based on hybrid supervised learning method, and evaluated the method on a newly collected dataset. The proposed method utilizes rich CNN intermediate features to enhance the performance of CNN classification. Experimental results show that the accuracy of the classification is improved when compared to other state-of-the-art methods by 2-4%. It shows the effectiveness of the proposed method in handling ethnicity classification.

VI. Acknowledgement

This research was carried out at the Rapid-Rich Object SEarch (ROSE) Lab at Nanyang Technological University, Singapore. The ROSE Lab is supported by Infocomm Media Development Authority, Singapore. We gratefully acknowledge the support of NVIDIA AI Technology Center for their dontation of a Titan X and a GeForce GTX 1080 GPUs used for our research at the ROSE Lab.

References

[1] H. Chen, Y. Deng, and S. Zhang, "Where am i from? -east asian ethnicity classification from facial recogition," *Project study in Stanford University*, 2016.

[2] W. Wang, F. He, and Q. Zhao, "Facial ethnicity classification with deep convolutional neural networks," in *Chinese Conference on Biometric Recognition*, pp. 176–185, Springer, 2016.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[4] D. Howard and P. E. Hopkins, "race, religion and the census," *Population, space and place*, vol. 11, no. 2, pp. 69–74, 2005.

[5] D. R. Williams, N. Priest, and N. B. Anderson, "Understanding associations among race, socioeconomic status, and health: Patterns and prospects.," *Health Psychology*, vol. 35, no. 4, p. 407, 2016.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

[6] S. Fu, H. He, and Z.-G. Hou, "Learning race from face: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2483–2509, 2014.

[7] X. Lu, A. K. Jain, *et al.*, "Ethnicity identification from face images," in *Proceedings of SPIE*, vol. 5404, pp. 114–123, 2004.

[8] M. A. Borgi, M. El'Arbi, D. Labate, and C. B. Amar, "Face, gender and race classification using multi-regularized features learning," in *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 5277–5281, IEEE, 2014.

[9] S. Hosoi, E. Takikawa, and M. Kawade, "Ethnicity estimation with facial images," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pp. 195–200, IEEE, 2004.

[10] Z. Yang and H. Ai, "Demographic classification with local binary patterns," *Advances in Biometrics*, pp. 464–473, 2007.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

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

[13] A. Gudi, "Recognizing semantic features in faces using deep learning," *arXiv preprint arXiv:1512.00743*, 2015.

[14] S. Masood, S. Gupta, A. Wajid, S. Gupta, and M. Ahmed, "Prediction of human ethnicity from facial images using neural networks," in *Data Engineering and Intelligent Computing*, pp. 217–226, Springer, 2018.

[15] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*, p. 1470, IEEE, 2003.

[16] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[17] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European conference on computer vision*, pp. 584–599, Springer, 2014.

[18] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.

[19] A. Salvador, X. Giró-i Nieto, F. Marqués, and S. Satoh, "Faster r-cnn features for instance search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 9–16, 2016.

[20] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic convnet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1790–1802, 2016.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE