Hindawi
Scientific Programming
Volume 2020, Article ID 6385281, 7 pages
https://doi.org/10.1155/2020/6385281

Research Article

Identifying Ethnics of People through Face Recognition: A Deep CNN Approach

Ahmed Jawad A. AlBdairi,[1,2] **Zhu Xiao**,[1] **and Mohammed Alghaili**[1]

[1]*College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China*
[2]*Computer Center, University of Babylon, Hillah, Babil, Iraq*

Correspondence should be addressed to Ahmed Jawad A. AlBdairi; ahmed_albdairi@hnu.edu.cn and Zhu Xiao; zhxiao@hnu.edu.cn

Received 11 February 2020; Revised 29 April 2020; Accepted 15 June 2020; Published 14 July 2020

Academic Editor: Chenxi Huang

The interest in face recognition studies has grown rapidly in the last decade. One of the most important problems in face recognition is the identification of ethnics of people. In this study, a new deep learning convolutional neural network is designed to create a new model that can recognize the ethnics of people through their facial features. The new dataset for ethnics of people consists of 3141 images collected from three different nationalities. To the best of our knowledge, this is the first image dataset collected for the ethnics of people and that dataset will be available for the research community. The new model was compared with two state-of-the-art models, VGG and Inception V3, and the validation accuracy was calculated for each convolutional neural network. The generated models have been tested through several images of people, and the results show that the best performance was achieved by our model with a verification accuracy of 96.9%.

1. Introduction

The scope of face recognition field has been increased recently. Face recognition refers to the ability of identifying any person from an image or a video frame. Many techniques have been used in face recognition. One of the first techniques used is using a 2D pattern recognition problem in which a distance between the important points in an image is used to recognize the face [1], like calculating the distance between eyes and distance between other important points.

Another technique is called holistic matching technique in which complete face region is taken into account as an input data into the catch face system. The most important studies that used this technique are eigenfaces [2], principal component analysis, and linear discriminant analysis [3].

Feature-based structural technique is another technique used in face recognition where the local features of the face are extracted first and their locations and local statistics are fed into a structural classifier.

The holistic and feature extraction techniques are used together to make a new technique called hybrid technique in which 3D images are used. The person's face image is caught in 3D; the system after that will note the important features such as curves or shapes in the face. The system after that detects the image whether it is a photograph or real time, determines the location of the face, and measures the curves and shapes of the important features in that face, converting the face into a numerical representation and matching this numerical representation with a dataset of faces.

The most important technique in face recognition that has been emerged recently is using the convolutional neural network (CNN) [4]. Although a lot of studies used CNN in face recognition, none of these studies has proposed a robust model to identify ethnics of people through their faces with high classification accuracy for people who have some similarities with different ethics.

Motivated by this, we propose two new models for face recognition with regularization and without regularization, in which they have the ability to recognize the ethnics and origins of people through their faces' facial. To specify, the primary contribution of this paper is proposing a face recognition model that can detect the detailed features of the

faces and differentiate between them using RGB images or a real-time face recognition. The ethics of different people can be recognized using this model through extracting the most detailed features of the peoples' faces. A new dataset has been collected for that purpose with high resolution from three different regions in Asia. These images were collected from social media like Facebook and VK (Russian social media website). Finally, we achieved a promising performance on another dataset collected for the test purpose.

The remainder of this paper is organized as follows. Section 2 presents the related works. Section 3 shows the designed network for face recognition. The experiments and results of the new models are given in Section 4. Section 5 concludes the paper.

2. Related Work

A face recognition method has been presented based on dense grid histograms of oriented gradients (HOG) [5]. In that study, the face image has been divided into many dense grids from which the HOG features have been extracted. After that, all these HOG feature grids vectors are composed to realize the feature expression of the whole face, and the k-nearest neighbor classifier is used for recognition. The authors used face dataset in the training stage with complex changes in illumination, time and environment, to test the gamma illumination correction, the spatial gradient direction, the size of the block, the standardization, and the face image resolution to find and analyze the optimal HOG parameters for face recognition. The FERET database is a dataset used for facial recognition system evaluation.

There are many methods in face recognition with high recognition accuracy, which are based on deep learning. One of these methods have a good effect in a restricted environment as well as in the natural environment [6]. The authors improved the method of multipatches by using 4 areas' patches in the face. In order to have a higher performance, they also used a Joint Bayesian (JB) measure in face verification. The model has been trained by the set of CASIA WebFace and tested in the Labeled Faces in the Wild (LFW).

Learning for face recognition has been proposed in another study [7]. The authors argued that the DeepID can be effectively learned through challenging multiclass face identification tasks. Furthermore, the generalization capability of DeepID increases as more face classes are to be predicted at training. They have used about 10,000 face identifications in the training set. The generated model achieved 79.45% verification accuracy on LFW dataset. The deep ConvNet contains 4 convolutional layers with Maxpooling to extract features hierarchically followed by the fully connected DeepID layer and the softmax output layer indicating identity classes.

The developing of an effective feature representations for reducing intrapersonal variations while enlarging interpersonal differences in face recognition has been solved in another study [8] using the deep learning and using both face identification and verification signals as supervision. The Deep IDentification-verification features (DeepID2) are learned by a deep convolutional network. The face identification task increases the interpersonal variations by drawing DeepID2 features extracted from different identities apart, and the face verification task reduces the intrapersonal variations by pulling DeepID2 features extracted from the same identity. The face verification accuracy that has been achieved by testing the method on LFW dataset [9] was 99.15% and this accuracy is different from the validation accuracy. The error rate has been significantly reduced by 67% as compared to the best previously deep learning results [7].

Another approach for face recognition was presented in which the convolutional neural network (CNN) and a logistic regression classifier (LRC) are combined [4]. The CNN used to extract the features in order to detect and recognize the face images and LRC [10, 11] is used to classify the features learned by the convolutional neural network. The structure of the CNN used in this study is composed of four layers: input layer, two convolutional layers, and one subsampling layer. The first layer is considered as 64×64; therefore, the dataset was resized to that size to be compatible with the proposed structure and the output layer is a fully connected layer with 15 feature maps with the size of 1×1.

In ours study, we build two models, with dropout and without dropout layers to find out the effect of this layer in the training. This study is concerned with the recognition of the ethnics of people through their facial features through these two models. We used a new CNN with regularization like dropout layers and without regularization to find out the most accurate performance. During training, we used Adam optimizer [12] with a learning rate of 0.001 and categorical cross-entropy loss function. The generated models can detect the detailed features of the faces from RGB images or through a camera.

3. Convolutional Deep Learning for Face Recognition

3.1. Ethnics Identification Using Deep Learning. Our deep learning layers consist of twelve layers. Four of these layers are Conv layers, each followed by the Maxpooling layer, and some of these Conv layers are also followed by the dropout layer after the Maxpooling layer to extract the facial features. A drop connect layer is placed after the four Conv layers as a separator between them and the two fully connected layers. The output of the drop connect layers is passed to a flatten layer to flatten the output before they pass to the first fully connected layer. Between the two fully connected layers, another dropout layer is used. The softmax output layer is used to identify the classes. The purpose of using dropout layers is to get rid of the overfitting during the training. Figure 1 shows the whole structure of the network layers that predict n classes (e.g., n is 3). The number of predicted classes n can be extended to contain as many nationalities as possible.

The input to this network is an image of $128 \times 128 \times 3$ size (e.g., 3 feature maps). The patch size is 3×3 with the same padding in every Conv layer and stride is 1 which make the output of the Conv layer roughly the same size as the input. The output of each Conv layer is passed to Max-

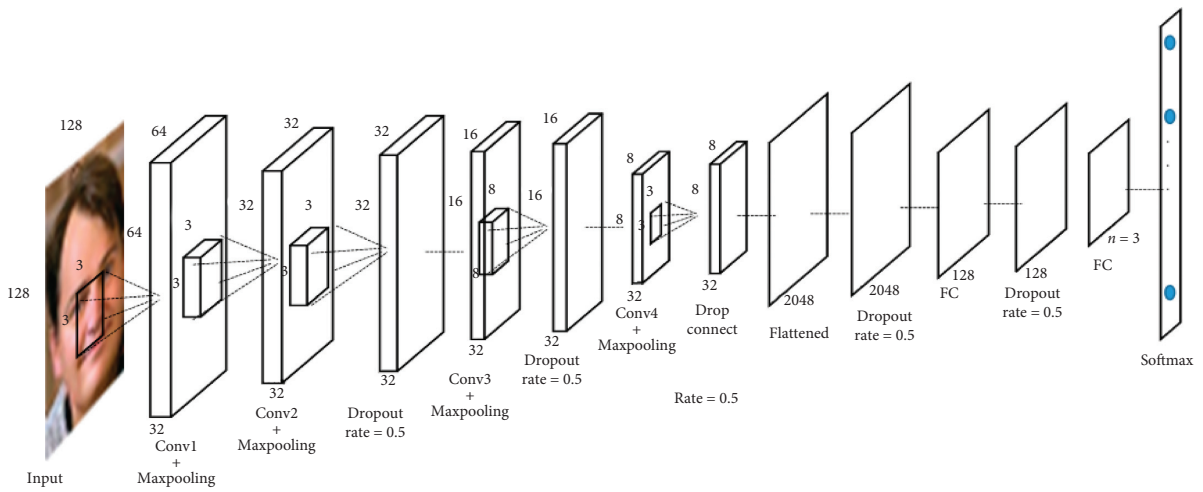

Figure 1: . The ConvNet layers. The small cuboid inside a square denotes the window map size of each Conv layer.

Pooling layer to minimize the input size. After that, the output of each Maxpooling layer is fed to ReLU activation function. The Conv layer with feature map equation is

$$f(x)^{j(r)} = \max\left(0, b^{j(r)} \sum_{n=1}^{\infty} k^{ij(r)} * x^{i(r)}\right), \tag{1}$$

where $f(x)^{j(r)}$ is the j^{th} output patch of the convolutional layer in a particular region r and $x^{i(r)}$ is the i^{th} input patch in a particular region r to the convolutional layer. The input of the first convolutional layer is an image of the size 128×128 divided into regions according to the size of window patch which is 3×3, as it is shown in Figure 1. $b^{j(r)}$ is the bias of the j^{th} output patch in the same particular region r. $k^{ij(r)}$ is the convolution kernel between the i^{th} input patch and the j^{th} output patch, whereas the multiplication of $k^{ij(r)}$ and $x^{i(r)}$ denotes the convolution.

The output of each convolutional layer is passed to the Maxpooling. The formula of the Maxpooling layer is as follows:

$$f(x)^{i}_{jk} = \max_{0 \le m,n < sz} \left(x^{i}_{j \cdot sz+m, k \cdot sz+n}\right). \tag{2}$$

The neurons in i^{th} the output patch $f(x)^{i}$ pool over $sz \times sz$ local region in the i^{th} input patch x^{i}. The output of the Maxpooling layer in each Conv layer is passed to ReLU nonlinearity $f(x) = \max(0, x)$. The ReLU sets all negative values in the input x to zero and all other values are kept constant, and it shows better fitting abilities than the sigmoid function [13].

Some of the Conv output is passed to a dropout to prevent the overfitting in the network. The number of dropout layers used is three where two of them are used after the second and the third Conv layer, and the third one is used between the last two fully connected layers.

The last layers are the two fully connected layers with dropout layer between them. The equation can be represented as follows:

$$fc = \max\left(0, \sum_{i} x^{i-1} \cdot w^{i-1, j-1}\right) + \max\left(0, \sum_{i} \text{DOut}_{\text{rate}}\left(x^{i} \cdot w^{i,j}\right)\right), \tag{3}$$

where x^{i-1} and $w^{i-1,\,j-1}$ denote the neurons and the weights of the previous layer, respectively. The output of the first fully connected layer is passed to $\text{DOut}_{\text{rate}}$ where the rate is 0.5 and the output of $\text{DOut}_{\text{rate}}$ is passed to the last fully connected layer. x^{i} and $w^{i,j}$ denote the neurons and the weights of the first fully connected layer before passing them to the $\text{DOut}_{\text{rate}}$ layer.

The output of the ConvNet is n-way softmax to predict the ethic of the face among n different ethics. The softmax works as follows:

$$y_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp\left(x_j\right)}, \tag{4}$$

where x_i is a vector of the inputs to the output layer and it denotes the most important features used to recognize the face. The output of that vector is calculated in x_j where x is the index of the output in n, e.g., number of classes.

3.2. Dropout Layers in the Network. Sometimes in the testing phase, the results are not accurate due to the training error. Researchers argue that because of overfitting [14], strong regularization like dropout [15] is used to overcome this problem. The idea of dropout is to drop out some neurons in a neural network wherein neurons are chosen randomly with probability $q = 1 - p$. When the neuron is dropped out, that means its input and output connection will be ignored and that will allow each neuron to learn something useful on its own without relying too much on other neurons to correct its shortcomings [16, 17]. Figure 2 illustrates the idea of dropout.

The input and output of each patch are computed as follows before we apply dropout:

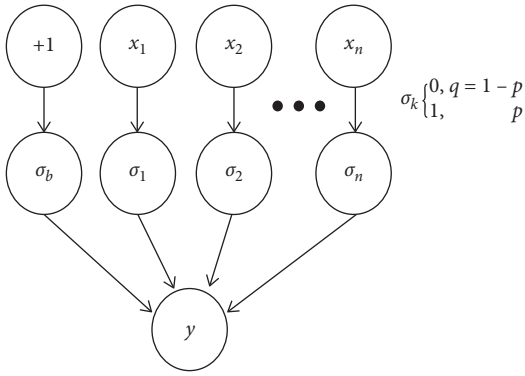

Figure 2: Neurons' training with dropout. The hidden neurons are randomly dropped out with Bernoulli's distribution p.

$$x^{l+1} = w^{l+1} y^{l} + b^{l+1}, \tag{5}$$

$$y^{l+1} = \mathrm{AF}\left(x^{l+1}\right), \tag{6}$$

where l denotes the index of the network layer. x^{l+1} is the input patch and y^{l+1} is the output patch at a hidden layer $l = 1, \ldots, l-2$, the layer being l. w^{l+1} is the weight and b^{l+1} is the bias. AF denotes the activation function. The following operations occur when the dropout is performed:

$$\sigma_i^l \approx \mathrm{Bernoulli}\,(p), \tag{7}$$

$$y'^{l} = \sigma^l \oplus y^l, \tag{8}$$

$$x^{l+1} = w^{l+1} y'^{l} + b^{l+1}, \tag{9}$$

$$y^{l+1} = \mathrm{AF}\left(x^{l+1}\right), \tag{10}$$

where $\oplus$ is the multiplication of an element by element and σ_i^l is a Bernoulli random variable of the i^{th} neuron at layer l with probability being 1.

3.3. Training Two Networks. The first network is the layers consisting of twelve layers including dropout layers. The training accuracy rate of this network is 96.9% and the validation accuracy rate is 96.9% with a validation loss of 0.221 which means the overfitting has been drastically eliminated as it is shown in Figure 3. In the second network, all the dropout layers are omitted, and the training accuracy is checked. The training accuracy rate in that network is 100%, the validation accuracy rate is 96.9%, and the least validation loss is 0.525. That means the overfitting is very high, and accordingly, the error rate of the created model from that network is more than that in the first network. Figures 3 and 4 show the training accuracy and validation accuracy for each network. The training accuracy in Figure 4 in epoch number 18 is 100% and that accuracy rate did not change until the end of the training which means the overfitting is very high and consequently the error rate is more than the error rate in the first network.

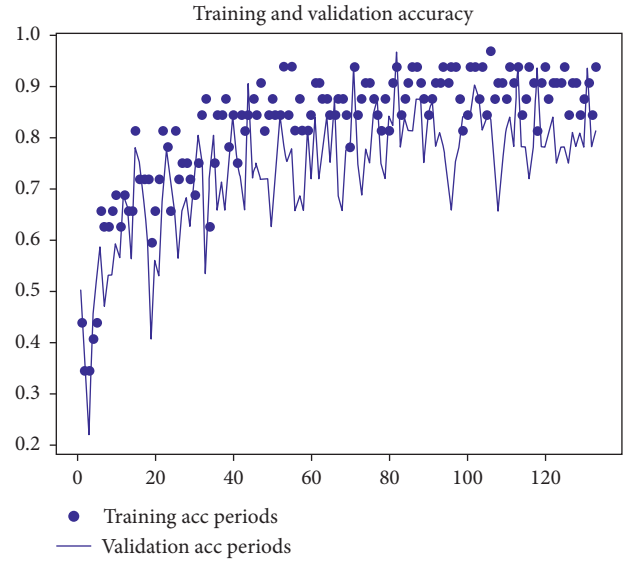

Figure 3: Training and validation accuracy for the first network.

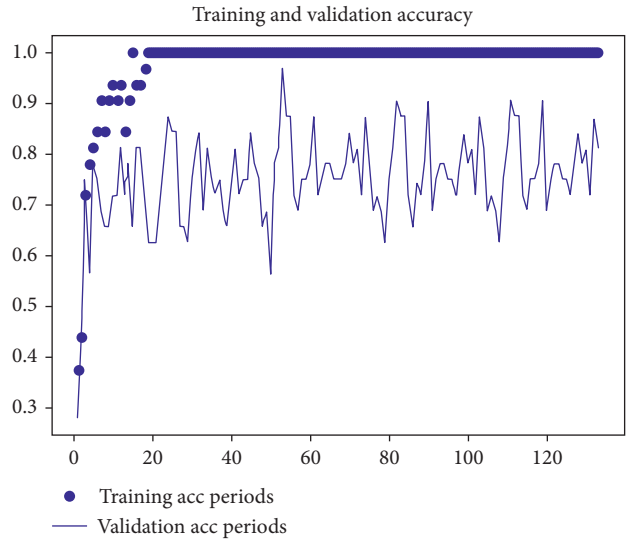

Figure 4: Training and validation accuracy for the second network.

4. Experiment

4.1. Experimental Training Dataset. Although there are many large-scale facial image databases available online, but all these databases are not appropriate to meet the objective of this study. Therefore, we manually collected 3141 photos from different resources. We collected 1081 Chinese facial images, 1021 Pakistani facial images, and 1039 Russian facial images. After collecting the images, they were processed to extract the faces from the whole images. The total images after that were divided into two sets; the first set was used for training stage and we took 70% of the whole images and the other 30% of the images as the second set for validation stage. Figure 5 shows a subset of the new dataset.

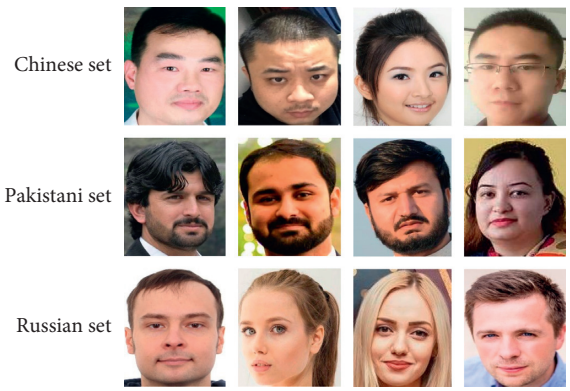

Figure 5: Three different subset images were collected from three different regions.

4.2. Comparison with the State-of-the-Art Approaches. Two state-of-the-art approaches were selected, and the last four layers for each approach have been frozen and used our fully connected layers to determine the number of output according to the number of classes in the dataset. These approaches are VGG [18] and Inception V3 [19]. The training was made in Tesla K80 GPU which is freely provided by Google Colaboratory. The results show that our approach has the highest validation accuracy and the least validation loss. Table 1 shows the results of training of our network and the two state-of-the-art approaches.

The comparison between our approach and the two state-of-the-art approaches VGG and Inception V3 is shown in Table 1 where it was observed that our approach has the highest validation accuracy (96.6%) with the less validation loss (0.22) as shown in Figure 3 with regularization. Figure 4 shows that our approach without regularization has the same validation accuracy (e.g., 96.6%), but the loss function value is different (0.525) which indicates that there is an overfitting problem, whereas the validation accuracy of VGG and Inception V3 are (91.48%) and (61.92%) with validation loss of (0.23) and (0.81), respectively, as shown in Figures 6 and 7.

Tables 2 and 3 summarize the total number of images for each category, the number of images that are predicted correctly and the number of images that are predicted incorrectly for the two models. The confusion matrix for both models is calculated to visualize the performance of each model.

The performance metrics that were widely used to evaluate the predicting results of the models were precision and recall. The results are summarized in Table 4.

Furthermore, a statistical significance test was conducted to compare the results of the two models. From the evaluation, the first model with the dropout layers has the highest accuracy rate with (90.65%), while the second model without the dropout layer has the lowest accuracy rate with (76.70%).

In this study, we need to insert some dropout layers into some specific places on our CNN to overcome the overfitting barrier and get high results. It is difficult to use some CNNs architecture like ResNet or SENet because they are heavy and take long time in training, and it is difficult to control the overfitting problem easily in such architecture due to the

Table 1: A comparison between our approach and the two state-of-the-art approaches over the validation rate and the validation loss values.

Approach	Validation acc. (%)	Loss
VGG	91.48	0.23
Inception V3	61.92	0.81
Our network	96.9	0.22

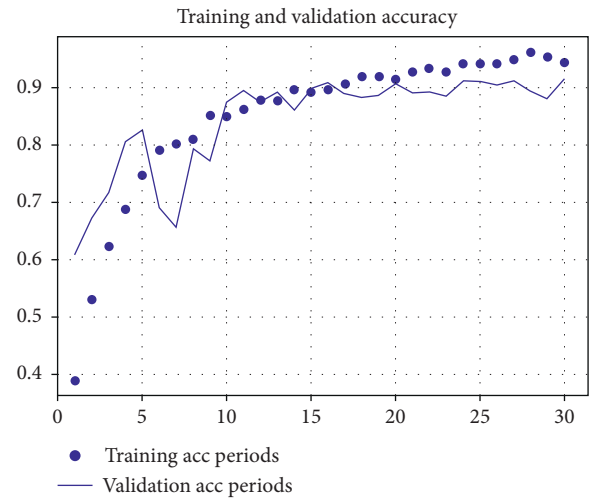

Figure 6: Training and validation accuracy for VGG.

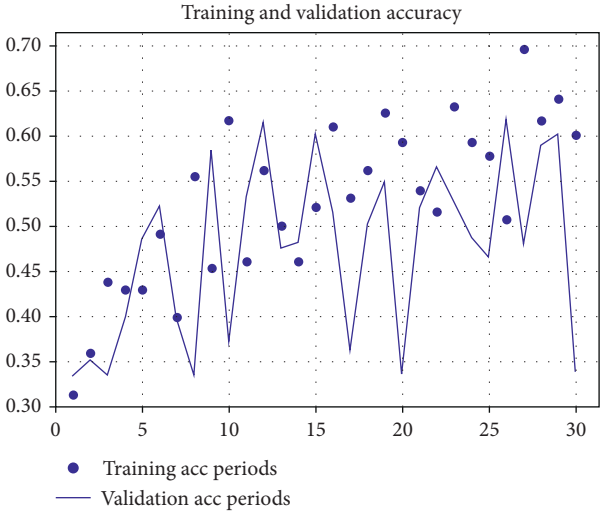

Figure 7: Training and validation accuracy for Inception V3.

Table 2: Number of images predicted correctly and incorrectly for the network with dropout layers.

Nationality	Total images	Correctly predicted	Incorrectly predicted
Chinese	540	511	29
Russian	642	561	81
Pakistani	582	527	55

Table 3: Number of images predicted correctly and incorrectly for the network without dropout layers.

Nationality	Total images	Correctly predicted	Incorrectly predicted
Chinese	540	388	152
Russian	642	467	115
Pakistani	582	498	114

Table 4: Statistical significance test for each model.

	Model with dropout	Model without dropout
TP	511	388
FP	29	152
TN	1088	968
FN	81	144
Recall (FP rate)	0.863176	0.729323
FP rate	0.025962	0.136079
Kappa	0.904659	0.762491
Accuracy rate	90.64626%	76.70068%
Precision	0.946296	0.718519

difficulties in changing their architecture. VGG and Inception V3 are also very heavy networks in training and it is difficult to change their architecture to control the overfitting problem too.

This paper is based on Cohen's methods [20]. Cohen's methods measure the degree of agreements amongst the assigned labels correcting for agreement by chance. In the evaluation, the number of unseen images is 1764, which is not included in the training dataset to evaluate the performance of each model. We found that the number of errors in the image predicting using a second model without dropout layers is larger than the number of errors in the first model with dropout layers.

5. Conclusions

In this paper, we propose a new deep learning convolutional neural network designed to create a new model that can recognize the ethnics of people through their facial features. The new model is compared with two state-of-the-art models, VGG and Inception V3, and the validation accuracy is calculated for each convolutional neural network. Two models from the proposed convolutional neural network are created with dropout layers and without dropout layers to discover the effect of the regularization in the performance of the models.

A new dataset is collected to use in the training phase to identify the ethnics of people through images from three different regions. This dataset is considered as the first dataset collected for ethnics of people and that will be available for the research community. Another unseen dataset is collected to evaluate the performance of our two models, and a statistical significance test is conducted to evaluate the performance of the two models.

Data Availability

The collected dataset has been uploaded to the following ULR: https://drive.google.com/file/d/1brRMSh7XDR7h5awgXudQXBqxAIiYSHy_/view?usp=sharing.

Disclosure

The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; and in the decision to publish the results.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant nos. 61836009 and 61702175), the Fund of State Key Laboratory of Geo-Information Engineering (no. SKLGIE2018-M-4-3), the Open Fund of Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education (no. IPIU2019007), the fund of Hubei Key Laboratory of Transportation Internet of Things (no. WHUTIOT-2019004), and the Natural Resources Scientific Research Project of Department of the Natural Resources of Hunan Province (no. 201910).

References

[1] C. A. Hansen, *Face Recognition*, Institute for Computer Science University of Tromso, Tromsø, Norway, 2009.

[2] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 586–591, IEEE, Maui, HI, USA, August 1991.

[3] S. Satonkar Suhas, B. Kurhe Ajay, and B. Prakash Khanale, "Face recognition using principal component analysis and linear discriminant analysis on holistic approach in facial images database," *IOSR Journal of Engineering*, vol. 2, no. 12, pp. 15–23, 2012.

[4] H. Khalajzadeh, M. Mansouri, and M. Teshnehlab, "Face recognition using convolutional neural network and simple logistic classifier," in *Soft Computing in Industrial Applications*, pp. 197–207, Springer, Berlin, Germany, 2014.

[5] Z. Xiang, H. Tan, and W. Ye, "The excellent properties of a dense grid-based HOG feature on face recognition compared to gabor and LBP," *IEEE Access*, vol. 6, 2018.

[6] J. Yan, L. Zhang, Y. Wu et al., "Research on face recognition method based on deep learning in natural environment," in *Proceedings of the IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, pp. 501–506, Taichung, Taiwan, November 2017.

[7] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1891–1898, Columbus, OH, USA, June 2014.

[8] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1988–1996, Montreal, Canada, December 2014.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, MA, USA, 2007.

[10] S. K. Palei and S. K. Das, "Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: an approach," *Safety Science*, vol. 47, no. 1, pp. 88–96, 2009.

[11] S. Permission, "Generative and discriminative classifiers: naive bayes and logistic regression," 2005.

[12] D. Kingma and J. Ba, "A method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097–1105, 2012.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[15] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, https://arxiv.org/abs/1207.0580.

[16] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26, pp. 351–359, MIT Press, Cambridge, MA, USA, 2013.

[17] P. Baldi and P. J. Sadowski, "Understanding dropout," in *Advances in Neural Information Processing Systems*, vol. 26, pp. 2814–2822, MIT Press, Cambridge, MA, USA, 2013.

[18] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730–734, Kuala Lumpur, Malaysia, November 2015.

[19] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015.

[20] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.