Comparison of Ear AWE and UTKFace datasets using convolution neural network models

Assignment #3

Image Based Biometrics 2021/1, Faculty of Computer and Information Science, University of Ljubljana

Robert Košir

***Abstract*—Nowadays, security is the most appreciated asset in any environment or situation. Most of the recognition systems use face to determine the properties of a criminal. One would think that face recognition is the most accurate system in the means of determining the gender and race of human being. However, significant contributions have been made in the field over recent years, but the major problem is still the same how to get additional information out of a criminal's image. This paper presents a way to construct a deep neural model using two different datasets where one uses ear and the other one uses face recognition to determine the properties of a human being. In addition to a comprehensive way for construction, the paper also shows the difference between two datasets that were used during the development of the models.**

***Index Terms* – biometry, dataset, convolution neural network, ear recognition, gender, ethnicity**

I. Introduction

Nowadays, most recognition systems use face to determine the properties of a subject. These systems are useful to find out the gender, age, ethnicity, mood and many more. However, it is known that data from the face is sometimes not accurate and there is a need to create another recognition system which will be used to determine the properties of a human being. This paper describes the way the CNN (convolution neural network) model is created and show results of the CNN that uses subject's ear to recognize its properties and is later compared with the results of another model that uses different dataset and face to recognize its properties. Face recognition is also used for many security purposes and therefore, we must not omit the fact that usually a criminal hides his face but not ears, as he knows that many can recognize him by his face. Thus, having a system that can give us some additional data would be beneficial.

A. Contributions and Paper Organization

In this paper, we show the results of two different trained models that try to determine the age and ethnicity of a subject. The first model will be created from UTKFace dataset [1] and will use the face as recognition type to out the properties whereas the second one will use AWE (Annotated Web Ears) dataset [2] and ear recognition type. The rest of the paper is structured as follows. In Section II we present the steps that were needed to train the models from the provided datasets. In Section III the training part and results from the models are discussed and compared. The paper concludes with some final comments in Section IV where we try to discuss possible improvements for our models

B. Dataset

Both databases presented below consist of many annotations. However, we will only need two, gender (man/women) and race information. We will define a person to one of the following races: white, Asian, south Asian, black, south American, other.

- UTKFace dataset UTKFace dataset consists of the cropped face image, which is annotated in the file name in the following format. {age}_{gender}_{race}_{date}.jpg. This database is already cropped is ready to use.
- AWE dataset AWE dataset was firstly present in the paper [2]. It consists of images collected from the web and is the first dataset for ear recognition gather on the internet. However, the images in this dataset are not of the same size and must be pre-processed in to be used with the same methodology as UTKFace dataset.

II. Methodology

A. Preparing the dataset

Firstly, we have to map image information (gender, race, etc.) to each image in the dataset. The fact that AWE does not consist of images of the same size, as mentioned in section I-B, we have to resize/rescale our images. After we have successfully rescaled our images, we split prepared data into a random train and test subsets which will be later used to define our model. Afterwards, we start with the model. First we *flatten* the data. Flattening transforms a two-dimensional matrix of features into a vector that can be fed into a fully connected neural network. Above steps are required for the input layer to be ready. Afterwards, we want to create a hidden layer. Here we add *Danse* layers, where we provide the number of units (neurons) in the layer and activation function. Activation function (sigmoid, relu, softmax) is a function that will make a neuron fire. In most cases, relu (Rectified Linear Unit) function is used. Finally, we have to define the output layer which in our case will also be Dense layer. The layer will have as many neurons as there are classifications (e.g. gender=2). We want the output to be a probability distribution where each classification will be mapped with the probability. In the end, we just have to define some parameters for the training of the model. We uses *adam* (Adaptive Moment Estimation)

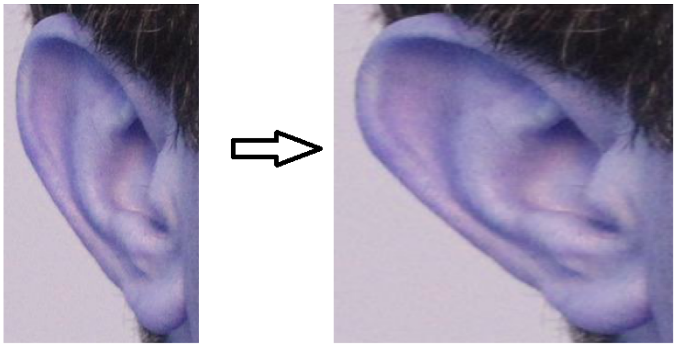

Figure 1. Resizing and rescaling images to 200px*200px

optimizer. The *adam* optimizer is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. [3] In a nutshell, it converges faster than regular gradient descent as it does not take into account the intermediate weights.

III. Results

A. UTKFace dataset

While the model was training, we logged how gender and race accuracy changed. It looks like a market price, sometimes it gets better sometimes falls down but in the end, it has some final "price" that nobody can change. Figure 2 shows how our model converged using UTKFace dataset, there are two lines present. The blue one presents how gender accuracy converged after every Epoch step. In the end, the accuracy for determining whether it is a man or woman was around $52,3\%$. With only $42,4\%$, we could say, that subject belongs to on of the races we defined in Chapter I-B.

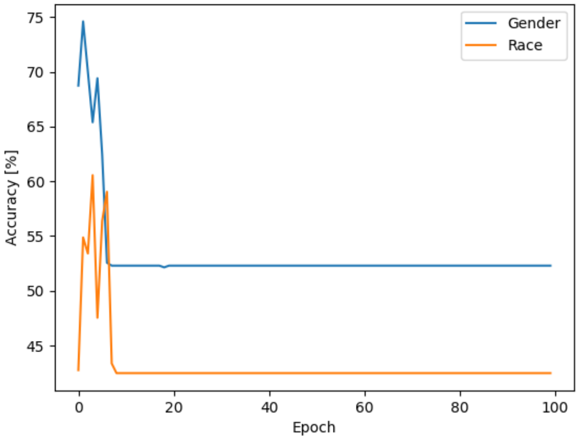

Figure 2. Accuracy of gender and race on UKTFace dataset.

B. AWE dataset

As we can see in the previous section, the results were not as good as we want them to be. Therefore, we trained the model on Ear AWE dataset and results, as we can see in Figure 3, are much better. Gender accuracy peaked at $epoch = 90$ with $99,2\%$. Using this dataset also increased race accuracy which peaked at $epoch = 89$ with $94,1\%$.

After we trained the model, we can use it to see how well it performs. The model was tested on several images and we can say that it is quite accurate. Figure 4 shows the output that uses ear detection and ear model to determine gender and race.

IV. Conclusion

In this paper, we explained the way we performed a comparison between two very different datasets. One using face and another using ear. One would think that to predict gender and race would be more accurate using face then ear but we have shown that AWE dataset is well prepared to train the model and gives us better results compared to UTKFace dataset. However, we have to take into account that our model may not be constructed in the best possible way.

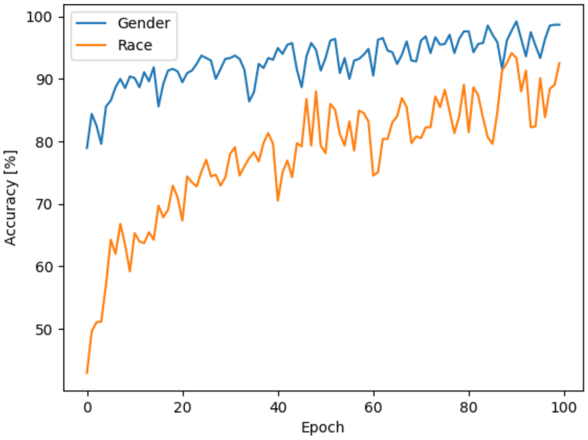

Figure 3. Accuracy of gender and race on AWE dataset.

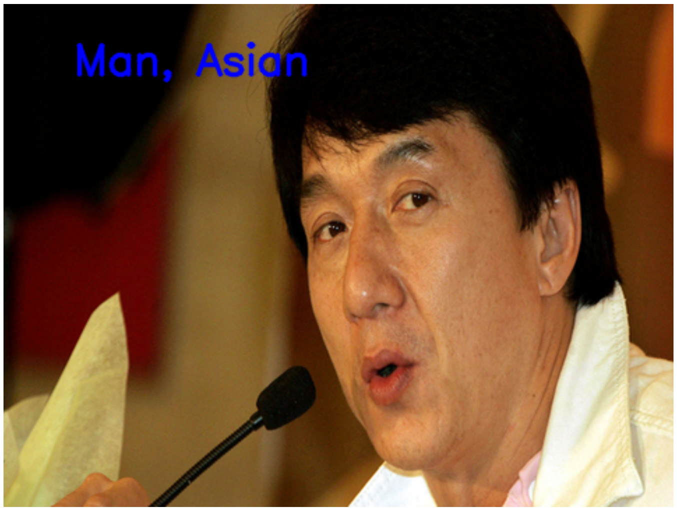

Figure 4. Ear detection and ear model on real data.

References

[1] S. Y. Zhang, Zhifei and H. Qi, “Age progression/regression by conditional adversarial autoencoder,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[2] Ž. Emeršič, V. Štruc, and P. Peer, “Ear recognition: More than a survey,” *Neurocomputing*, vol. 255, pp. 26–39, 2017.

[3] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.