Face Recognition with Age, Gender and Emotion Estimations

Dongqin Gu [a,b*], Miaomiao Zhu [b], Lifeng Zhang [b], Seiichi Serikawa [b]

[a]Yangzhou University, Yangzhou, 225000, China
[b]Kyushu Institute of Technology, Kitakyushu, 8048550, Japan

*Corresponding Author: gu.dongqin580@mail.kyutech.jp

Abstract

In this paper, we propose an application that can be used to help retail industries adjust purchase lists and operation modes to increase sales and improve service quality. The convolutional neural network and residual network structures are used to estimate the age, gender, and emotions of the face, which can help merchants to build their own databases, analyze users' shopping preferences, and track sentiment to the feedback shopping experience.

keywords: Face detection, age estimation, gender estimation, emotion estimation.

1 Introduction

A face contains much information, such as age, gender, and emotions. Age and gender have features of personal identity, which play important roles in social life. Artificial intelligence's prediction of age and gender can be applied in many areas such as intelligent human-machine interface development, security, cosmetics, and e-commerce. And the recognition of emotions can help us understand the psychological state of the participants.The estimations of age, gender and emotion can make the feedback more comprehensive.

We propose a system to estimate the age, gender, and emotions of the face by using convolutional neural network and residual network structures. This system can be applied in retail industries, helping merchants build their own databases to analyze shopping preferences and shopping experiences of users with different ages and genders, and then adjust purchase lists or operating models in a timely manner. The advantages are been shown as follows.

On the one hand, it can track customers' face vectors and collect useful information such as age and gender. Through this data, merchants can categorize customers, and count their shopping preferences, so that to design targeted marketing strategies and in-store activities.

On the other hand, it can catch customer sentiment. This system can provide valuable feedback for in-store promotions by tracking customer emotional responses, enabling retailers to improve product categorization and improve service quality.

We create a face recognition system based on three pre-trained convolutional neural networks (CNN) models that recognize faces and predict their age, gender, and emotion from images or videos.

The IMDB Database contains the facial image with the gender and age tags were selected to train age and gender models, and the FER-2013 Database with the emotion tags was used to obtain the emotion model. The position of the face is obtained by using the Haar cascade classifier model. And then the age, gender and emotion models are separately trained in conjunction with the residual network. Finally, the age, gender, and emotion information of faces in the video can be estimated in real-time.

2 Database

In the process of interaction with the user, the machine can infer the user's age, gender, and emotion by relying only on facial information. This task is highly complex for different face samples. Using machine learning techniques, models with millions of parameters are trained under tens of thousands of samples.(1) This section introduced the databases used for real-time estimation with age, gender, and emotion.

2.1 IMDB-WIKI Database

In this paper, we use the IMDB-WIKI Database to train age and gender features and use the corresponding

DOI: 10.12792/iciae2020.055

Table 1: IMDB-WIKI Database and its partitions sizes in number of images.

Database	number of images
IMDB-WIKI	524,230
IMDB	461,871
Wikipedia	62,359
For CNN training	260,282

Fig. 1: Samples of the IMDB database

models for age and gender estimations.

IMDB-WIKI is a database of facial images containing the corresponding age and gender labels. A total of 524,230 images containing age and gender information were obtained from the IMDB and WIKI websites. Table 1 shows the distribution of pictures obtained in IMDB and WIKI. Part of faces in the IMDB Database and the age distribution in IMDB-WIKI are shown in Fig. 1 and Fig. 2, respectively.

2.2 FER-2013 Database

Similar to IMDB-WIKI, the FER-2013 Database is used to train facial emotions, and the corresponding model is used for facial emotion recognition.

The FER-2013 facial emotion database consists of 35,886 facial emotion images. It contains 28,708 test pictures (Training), 3,589 public test charts (PublicTest) and private verification pictures (PrivateTest). Each picture is composed of grayscale images with a fixed size of 48×48. There are 7 kinds of emotions, which are represented by the numbers 0-6. The labels and corresponding emotions are shown in Table 2.

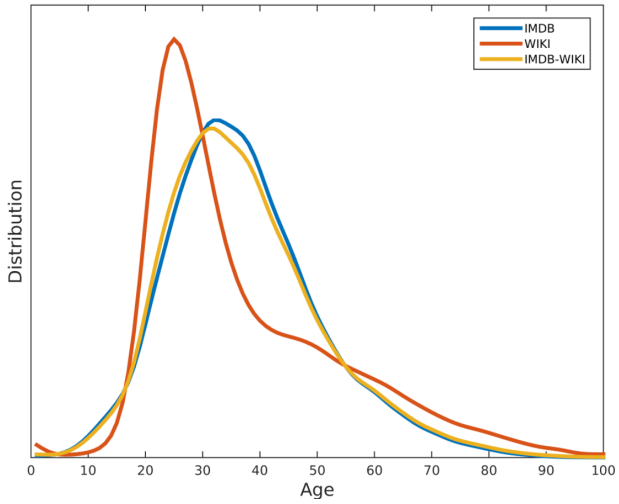

Fig. 2: The distribution of age in the IMDB-WIKI Database

Table 2: The corresponding labels and emotions in FER-2013 Database

Label Number	Emotion
0	anger
1	disgust
2	fear
3	happy
4	sad
5	surprised
6	normal

3 Model

Different target requires different detection and recognition model. In our system, one detection model, the face detection model, and three estimation models, age, gender and emotion estimation models, need to be considered.

3.1 Face Detection

For training and testing images, we use the Haar cascade classifier in OpenCV to get the position of the face in the face detector. The earliest Haar features were proposed by Papageorgiou C. *et al.*[4] Then, Paul Viola and Michal Jones proposed a method for quickly calculating Haar features using the integral image method.[5] Later, Rainer Lienhart and Jochen Maydt extended the Haar signature library with diagonal features.[6] The Haar cascade classifier in OpenCV is based on the extended feature library.

The Haar-like feature is a feature for real-time face

tracking. Each class of Haar feature describes the contrast mode of adjacent image regions. For a given image, the features may vary depending on the size of the region, which may also be referred to as the window size. However, only two images that differ in scale should have similar features. Therefore, it is useful to be able to generate features for windows of different sizes. These feature sets are called cascading. And the Haar cascade has scale invariance.

3.2 Age and Gender Estimation

The age and gender of face images were trained using the IMDB-WIKI Database. The literature[2,3] used a pre-trained VGG network and proposed a novel regression algorithm for the classification of ages. The essence is that after the 101 categories between 0-100, the scores obtained are multiplied by 0-100, and the final results are summed to obtain the final identified age. The process is shown in Fig. 3. Through investigation, we can know that using the IMDB Database, the correct rate for testing gender predictions can reach 96%.

Different from,[2,3] this paper uses a wide residual network to start from scratch. On the basis of,[7] the extensive residual network portion was modified by adding two classification layers at the top of the extensive residual network for age and gender estimates. In,[2,3] age and gender were independently estimated by two different CNNs, while a single CNN simultaneous estimation was used in this paper.

Firstly, download the IMDB-WIKI Database, filter out the noisy data and serialize the images and tags for training into a type of ".mat" file. Then, train the network with the training data above. If the verification loss becomes minimum in the previous period, the training age and gender weight files are stored as "weights.*.hdf5" and "gender_mini_XCEPTION.*.hdf5", respectively. The workflow is shown in Fig. 4.

In the training network section, we used a residual network.[8] The structure of its internal residual block is shown in Fig. 5. The residual module modifies the required mapping between subsequent layers so that the learned features become the difference between the original feature map and the desired feature. Therefore, the desired feature $H(x)$ is modified to solve the easier learning problem $F(x)$. That is, satisfy the equation (1).

$$H(x) = F(x) + x. \tag{1}$$

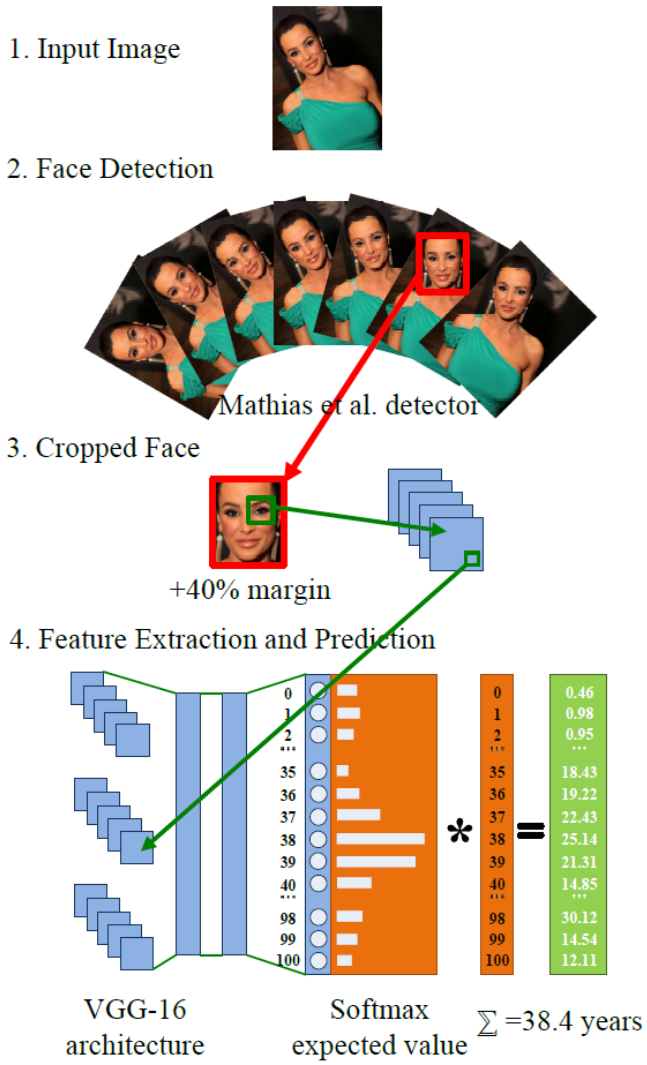

Fig. 3: Pipeline of DEX method (with one CNN) for apparent age estimation

3.3 Emotion Estimation

The FER-2013 Database is used to train the emotions of the face. Firstly, download the FER-2013 Database. There are several versions of the FER-2013 Database. This article uses the "imdb_crop.tar" file (faces only 7G). After decompressing, perform the expression training classification operation to obtain the "fer2013_mini_XCEPTION.*.hdf5" file. The next steps are similar to the workflow for age and gender estimation.

Emotions, images and the purposes of usage are stored in a csv file in the form of data, not pictures. As shown in Fig. 6, the first line is the header, which means the meaning of each column of data. The first column represents the emotional tag, contains seven numbers from 0-6. The second column is the original image data, which

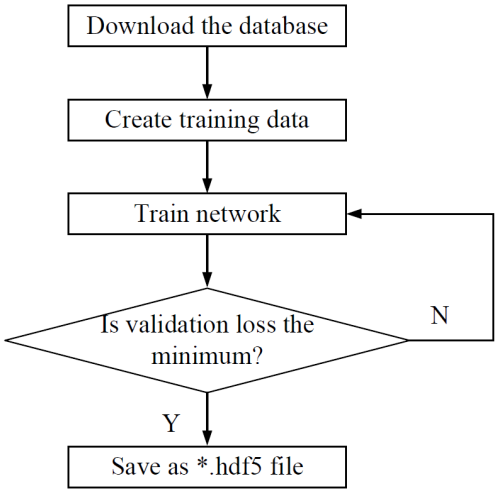

Fig. 4: The workflow of the age and gender estimation

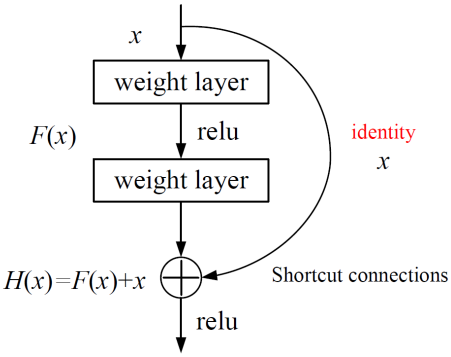

Fig. 5: The internal residual block

contains 2304 data per line. And the last column is the purpose of usage.

Use the pandas library to parse the csv file, and then save the original image data as a jpg file. Sort it according to the purpose and label. Then, save it to the corresponding folder. The running result is shown in Fig. 7. Subfolders with corresponding emotions in the PrivateTest, PublicTest, and Training folders, are shown in Fig. 8. A face image of the corresponding mood is stored under each emotion folder. Part of the emotion pictures in the FER-2013 Database is shown in Fig. 9. Through investigation, the accuracy can obtain 66% in the estimation of the emotion classification tasks.

4 Experiments

In order to verify the correctness, some experiments are carried out. This section describes the environment configurations and part experimental results.

emotion	pixels	Usage
0	70 80 82 72 58 58 60 63 54 58 6	Training
0	151 150 147 155 148 133 111 140	Training
2	231 212 156 164 174 138 161 173	Training
4	24 32 36 30 32 23 19 20 30 41 2	Training
6	4 0 0 0 0 0 0 0 0 0 0 3 15 23	Training
2	55 55 55 55 55 54 60 68 54 85 1	Training
4	20 17 19 21 25 38 42 42 46 54 5	Training
3	77 78 79 79 78 75 60 55 47 48 5	Training
3	85 84 90 121 101 102 133 153 15	Training
2	255 254 255 254 254 179 122 107	Training

Fig. 6: The FER-2013 Database in the csv file

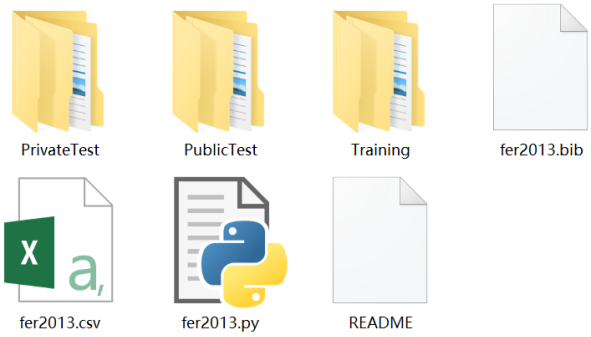

Fig. 7: Parsed image folders

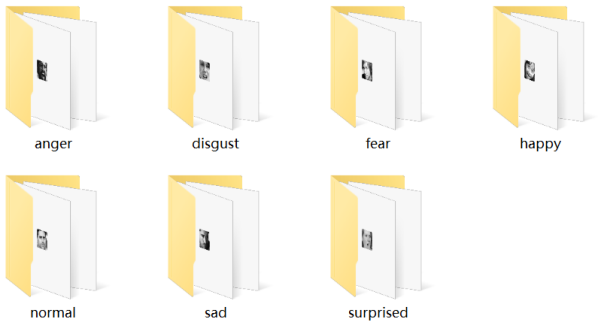

Fig. 8: Subfolder of the corresponding emotions

4.1 Software Configuration

This experiment is carried out in the following environment and libraries: Anaconda3, python3.5, OpenCV3.4.2, TensorFlow1.14.0, Keras2.2.5, scipy1.2.1, numpy1.17.0, pandas0.25.1, tqdm, tables, and h5py2.9.0.

Anaconda is an open source tool. The conda toolkit and virtual environment management system for Windows, MacOS and Linux, which can be supported for sklearn, TensorFlow and sciPy. This experiment is based on the Windows version of the Anaconda environment.

TensorFlow, a Python library for numerical calculations, is widely used in the programming of various ma-

Fig. 9: Samples of the FER-2013 Database

Table 3: The correct rate of age, gender and emotion estimations

Test Object	Age	Gender	Emotion
Person 1	89%	100%	75%
Person 2	91%	96.66%	68%
Person 3	88.4%	95.83%	72%
Person 4	86.3%	100%	71%

chine learning algorithms. The session in TensorFlow connects it to the C++ backend, assigns computing devices (CPU or GPU) to it, and provides computational methods to iteratively train the model. Due to limited conditions, this experiment uses the CPU version of TensorFlow.

Keras is an open-source artificial neural network library written in Python that can be used as a high-level application interface for Tensorflow, Microsoft-CNTK, and Theano for the design, debugging, evaluation, application, and visualization of deep learning models.

4.2 Real-Time Estimate Results

For the age, gender, and emotion of the face, we used a webcam for real-time detection. And the results of the estimations are shown in Fig. 10 and Fig. 11. Above the faces, the predicted age, gender and emotion are displayed, respectively. For the prediction of age, it is assumed that the error within plus or minus 2 years is considered to be correct. We did some experiments, and the results are shown in Table 3.

From Table 3, it can be seen that the prediction accuracy rate for gender is the highest, which is more than 95%. Followed by gender, the accuracy rate can reach more than 86% in age estimation. The recognition ac-

Fig. 10: Real-time estimate results (woman)

Fig. 11: Real-time estimate results (man)

Table 4: The correct rate of different emotion recognition

Emotion	Correct Rate
anger	89%
disgust	26%
fear	45%
happy	95%
sad	86%
surprised	62%
normal	53%

curacy of these two features can well realize user classification, feedback shopping preferences, so that to help merchants adjust the purchase list timely. However, the accuracy of emotion estimation is low, not more than 75%. Therefore, we made statistics on the accuracy of individual emotion estimation, and the results are shown in Table 4. We found that the correct rate for angry, happy, and sad emotions are high. And it is possible to focus on detecting these three emotions to feedback user's shopping experience.

5 Conclusion and Future Works

Through the face information, the system we proposed can correctly detect faces in real-time, and estimate the age, gender, and emotion. Applied in retail industries, it can help merchants build their own databases and analyze the shopping preference so that to adjust the purchase list and operation mode timely. Moreover, through emotion feedback, retailers can catch user experience to improve service quality.

At the same time, there are several considerations for future work. First of all, the prediction accuracy needs to be improved, specifically the emotion estimation. Secondly, we want to build a user interface for easy operation and data management. Finally, enter member information to improve personalized service.

Acknowledgment

This work was supported by Yangzhou University International Academic Exchange Fund.

References

(1) Dario Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and mandarin," CoRR, abs/1512.02595, 2015.

(2) R. Rothe, R. Timofte and L. V. Gool, "DEX: Deep EXpectation of apparent age from a single image," ICCV, 2015.

(3) R. Rothe, R. Timofte and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," IJCV, 2016.

(4) C. P. Papageorgiou, M. Oren and T. Poggio, "A general framework for object detection," Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271), pp: 555-562, 1998.

(5) P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, 2001.

(6) R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," Proceedings. International Conference on Image Processing, 2002.

(7) S. Zagoruyko and N. Komodakis, "Wide Residual Networks," https://arxiv.org/pdf/1605.07146v4.pdf

(8) 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, 2016.