Identity-Aware Automatic Emotion Detection And Analysis System

Xinchen Wang, Xiaowen You, Jiakai Zhang and Chuang Liu *Instructor*

School of Information Science and Technology, Shanghaitech University

Abstract

Recently, in the field of computer vision, the algorithms for Facial Recognition have already been mature, while Facial Expression Recognition is a problem explored by many. Our design combined those two kinds of algorithms, which enabled Up2 to recognize and memorize emotions by people's facial expressions and give real time feedbacks. We resorted to convolution neural network to train our emotional recognition model. The results were presented by an interactive emotion management interface.

Key words: real-time system, face recognition, emotion recognition, convolution neural network

1 Introduction

An Identity-Aware Automatic Emotion Detection and Analysis System is a system which can automatically recognize and distinguish between different person's faces, assess certain emotions on the faces and finally record the corresponding emotion data. It is a real-time system which not only gives instant response to image inputs but also have long-term memory. The design has applied machine learning, a recently hot topic, to the realm of human behavioral science. It aims to help people comprehend interpersonal behaviors and their implicit meanings through technological methods. Furthermore, understanding and responding to people's emotions would be the next milestone for AI to communicate with people. Thus, the importance of emotion recognition and analysis is self-evident, and the prospect of this topic is brilliant.

2 System Design

The hardware structure of this system is relatively simple, which is composed of

the IntelUp2 board, a 2 million pixel OV2710 distortion-free camera, a random screen, a mouse and a keyboard. The operation system of Up2 is Linux. This section will elaborate on the software structure.

As is shown in Fig. 2, the system includes a sensing system, a real-time information processing system, a pre-training system and an information storage system. There are some other peripherals such as the user interface which asks the user whether to memorize the appearance of a certain person.

Currently, the sensing system only consists of a camera. But there is potential to add more sensors in order to better monitor people's emotion.

The pre-training system locates outside the board. It trains the convolution neuro-network model for the emotion recognition algorithm. Since the calculation is fairly complex, the training procedure had to be completed on a remote host unless a neural computing stick was added to the board.

The real-time signal processing system is the core of our design. It can be divided into the character identification component and the emotion recognition component. It is fed with real-time images, displays identity and emotion recognition results dynamically on a screen, interacts with the user and delivers emotion data and pictures of specified persons to storage system. The character identification utilizes a python face-recognition library while the emotional recognition is based on the trained neuro-network.

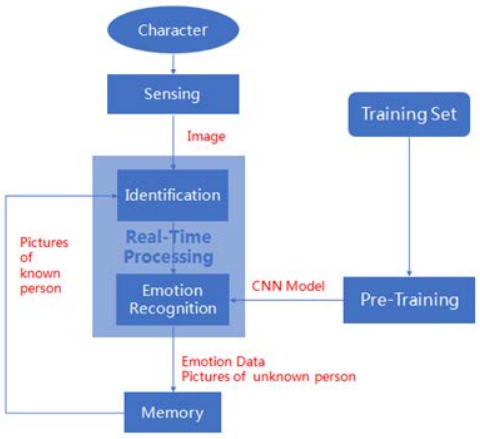

Figure 2 Software Structure

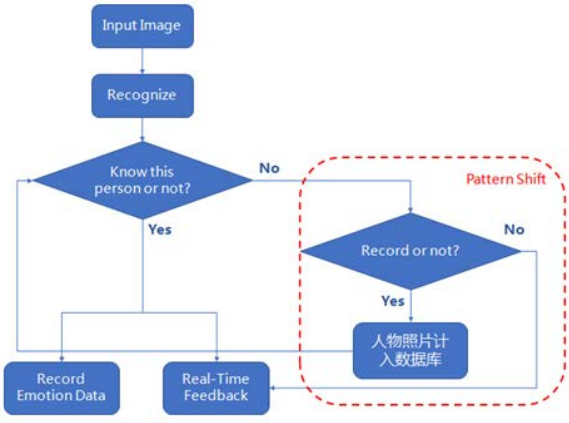

Figure 3 Operation Flow

3 Functions and Operation Flow

Fig. 3 depicts how the system operates.

First, after a frame of input image was pre-processed (normalize grayscale, filter noises etc.) , the character identification system would locate people's faces and judge whether the features of a face matches the photo of a known person in memory. Once any unknown face shows up, it would ask the user whether to add the person to database, and to name the person to be added through command line. Then folders with pictures of personnel of interest were created in memory. Next, the emotion recognition component would classify the emotion for all the input faces. However, only the emotion data for faces of interest was recorded. At the same time, the video of input image was displayed on screen, with people's faces boxed by a rectangle and their names and emotions labeled nearby. If an unknown person appeared, the name would be exactly 'unknown'.

On the top of that, the system has two modes. The first modes was described in the above paragraph while the second mode was exactly the same as the first mode except that it does not interrupt and ask the user the add unknown person to memory. It only records the emotion data for personnel that already exist in database.

4 Algorithm Depiction

Emotion recognition is the primary technical difficulty in our design therefore we will focus on the corresponding algorithm – convolution neuro-network in this section. The overall structure is of our CNN model is in Fig. 4-2.

CNN is a deep feed-forward neural network. A CNN has multiple convolution kernels which convolve the input image, extracting the crucial features for classification. After an image has passed through the convolution layers, several feature maps were extracted:

$$FeatureMap(i,j) = \sum_{m=1}\sum_{n=1} K(m,n) * Input(i-m+1, j-n+1)$$

where K(m,n) means a kernel, m n are kernel's index in two dimentions.

In the classic CNNs, diffusion and explosion of gradients frequently happens when training, spoiling the results. A batch normalization layer can solve the problem. The position of BN layer depends. In this model, it's put before the activation layer.

The following formulas summarize the function of BN layer:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad mean\ of\ mini-batch$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2 \qquad variance\ of\ mini-batch$$

$$\hat{x}_i = \frac{x_i-\mu_B}{\sqrt{\sigma_B+\varepsilon}} \qquad normalize$$

$$y_i = \gamma\hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i) \qquad scale\ and\ shift$$

where y_i are the resolutions on feature maps, γ $and\beta$ are used for reforming the feature map after convolution[1]

Next, the extracted feature maps enter activation layers for further feature extraction. ReLU function, a function useful for simulating the activation process of the creature neural, is used as the activation function for accuracy:

$$ReLU(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

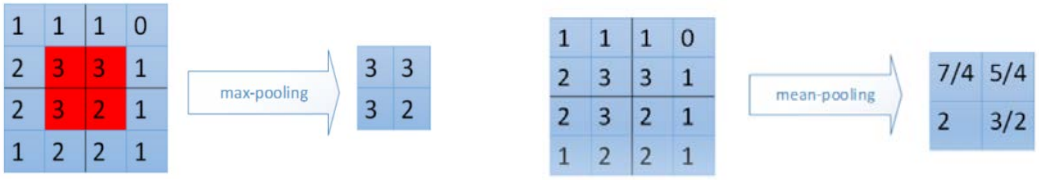

Figure 4-1 Max Pooling and Average Pooling

Then, the outputs are compressed by pooling layers to reduce feature redundancy. Two common kinds pooling methods are Max Pooling and Average Pooling.

After that, a global average pooling layer is used to classify the emotions. Global average pooling is a latest breakthrough. It is a better alternative for fully-connected layers, which extracts features from the overall image. It alleviates overfitting and boost the speed. The output of each neuron is mapped to a value between 0 and 1 by the Softmax function[2]:

$$S_i = \frac{e^i}{\sum_{j=0} e^j},$$

where S_i is the probability of the class i.

Finally, to train the network, iterations of the above procedures are carried out. The goal is to quickly reduce the output of loss function. Adam's optimization algorithm is applied to automatically adjusting the learning speed during iterations[3].

Our model resorted to Xception method and Global Pooling Layer (GAP) to shrink the number of training arguments. It not only reduced calculation complexity but also increased accuracy because redundant parameters were given up.

FER-2013 database was used to train our network. It contains 38000+ figures

labeled with the six fundamental emotions: happy, sad, angry, fear, disgust and surprise. The accuracy for our emotion detection modal was 65%.

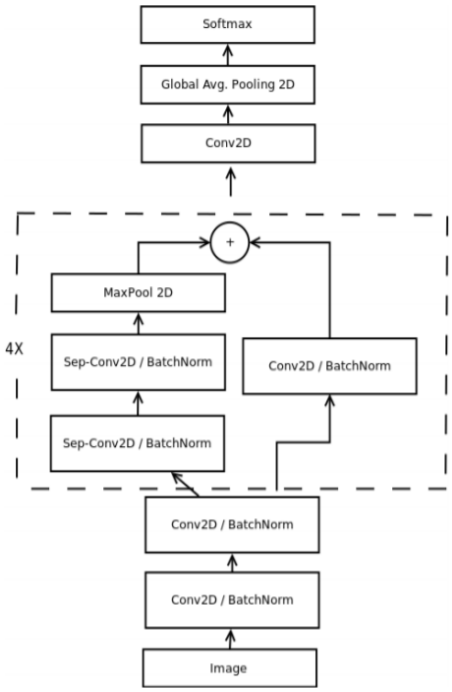

Figure 4-2 Model we used for training with Xception and GAP[4]

5 Test Results

Figure 5-1 *The Truman Show* movie clips

Figure 5-2 Truman and Lora in database

Movie clips from 'the Truman show' were chosen to test our system because expression in movies changes more dramatically and vividly than in real life. In this film, Truman fell in love with Lora at first sight. He wondered whether Lora is single. He became disappointed after Lora said she was neither allowed to speak to him nor to date with him next week. Nonetheless, she invited him to have a pizza with her right away. As a result, Truman immediately turned from sad to joy.

Emotion data was first extracted from each frame the video and then recorded in a six-dimensional vector for the sake of calculation. If any emotion showed up in the frame, the corresponding component in the vector would be one, otherwise it would

be zero. The average vector every two seconds were calculated for trend of change of emotion intensity. The results are shown in Fig. 5-3.

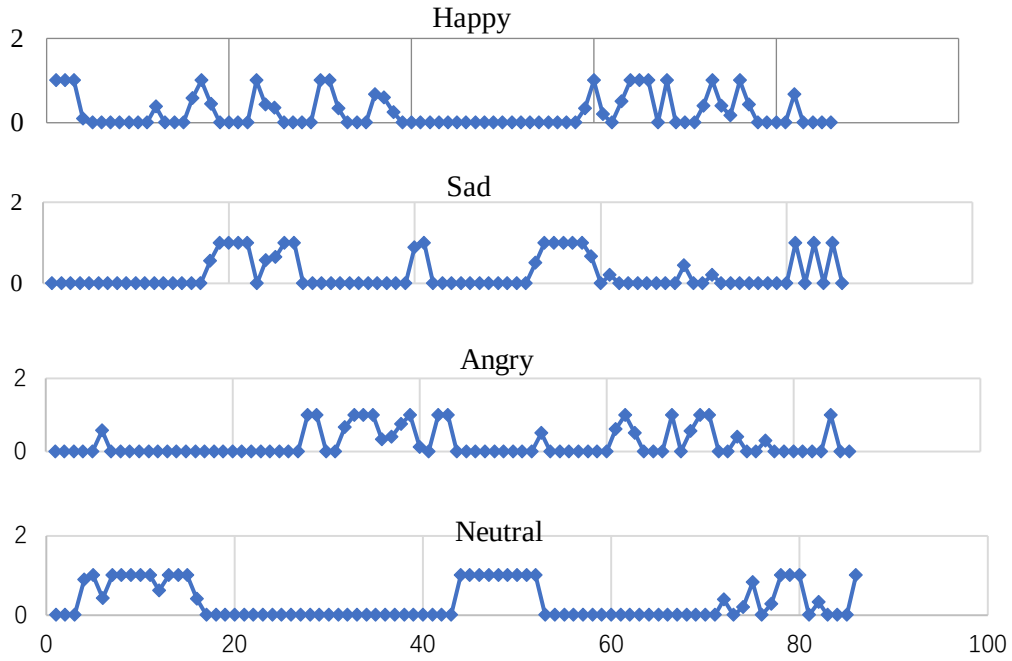

Figure 5-3 Emotion change of *Trueman* Clip

The horizontal axis is in seconds while the vertical axis signifies emotion intensity.

Some inevitable errors exit in the results. One possible cause is the shooting angle. Faces in training sets were shot from the front but in the tests there were always deviations from the front face position.

References

[1] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. 2015:448-456.

[2]Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.

[3]Kingma D, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.

[4] Octavio Arriaga, Paul G. Plöger, Matias Valdenegro. Real-time Convolutional Neural Networks for Emotion and Gender Classification[J/OL], 2017, https://github.com/oarriaga/face_classification/blob/master/report.pdf