Facial Expression Recognition for User Response Information Extraction

Genki Yamaguchi^{a,*}, Naoya Yamaguchi^b, and Kenji Matsui^{a,*}

^a Graduate School of Robotics & Design Engineering, Osaka Institute of Technology, Japan

^b Graduate School of Engineering, Osaka Institute of Technology, Japan

*Corresponding Author: kenji.matsui@oit.ac.jp

Abstract

This paper describes our preliminary study of facial expression recognition in order to extract user response information. We used Kinect to recognize facial expressions of the user in real time, we extract 6 facial expression categories (neutral, happiness, disgust, surprise, sadness, closed-eye). As for the recognition process, we applied a multi-layer-perceptron to classify the face expressions. A total of 1,912 facial expression data sets were collected from 17 subjects. We performed holdout test using 80% of training data and 20% of test data. The recognition rate without “sadness” feature was around 90%, and the rate using every categories was around 80%. Then, we focused on detecting the 0 to 5 level of “smile” face, such as “smiling”→“laughing”→“loud laugh”. The experimental result showed that the average correct answer rate was 61.3%. However, if we consider the nearest neighboring results of the expected result as a correct answer, the score goes up to the average of 93.5%.

Keywords: Facial recognition, 3D sensor, machine learning.

1. Introduction

In our life, facial expression plays an important role in communication. Among human non-verbal behaviors such as gesture, contact, posture, walking, expressions appearing on the face are said to be particularly large in information volume and can be said to play a central role in interpersonal communication. If the computer can infer the partner's emotions from the expression of the face like a human being, various kinds of information deliberately or unconsciously transmitting by human facial expressions, like a human being, it will be able to acquire naturally through the channel of vision⁽¹⁾. In this research, we investigate a facial expression recognition performance

using Kinect V2 sensor with machine learning based classification algorithm. We first apply the facial expression recognition for 5 types of facial expressions. This is useful in case of understanding user's over all feedback. Next, we focus on detecting the intensity level of “smile” face, such as “smiling” → “laughing” → “loud laugh”. This approach is also useful in case of comparing people's performance or finding the best part of total performance, almost automatically.

Face detection has been studied by many researchers, and latest powerful approaches are based on 3D images. However, face detection in video frames has not been extensively studied. Vinnetha et al. ⁽²⁾ described facial expression recognition using Kinect 3D features. A. Youssef et al. ⁽³⁾ described their attempt to recognize facial expressions using a 3D Kinect sensor. They constructed a training data set containing time dimension. For individuals who did not participate in training the classifiers, the best accuracy levels were 38.8% (SVM) and 34.0% (k-NN). However, in case of closed test, the best accuracy levels that they obtained raised to 78.6% (SVM) and 81.8% (k-NN).

Puica et al. ⁽⁴⁾ showed emotion recognition from facial expressions using Kinect sensor with the Face Tracking SDK. The accuracy of emotion recognition with data outside the training set was around 80%. However, the results of the systems described showed that the expression of sadness and disgust were more difficult to recognize than others. In order to solve this problem, a neural network with Kinect sensor output was tested for the facial expression detection. Since Kinect has been the tool to recognize the faces in video frames, using a multilayer perceptron to classify the different expressions detected in the users' faces is a simple but powerful combination. Using such recognition framework to detect sad or disgust expressions, improved results were observed.

Likewise, to evaluate some systems based on art, like

music or visual art, an empirical test is usually performed. This test consists of a list of questions about several audio files or digital images where they give a punctuation of the quality of the art shown. However, in such situations the results obtained can differ from what listeners really feel when the sounds are played. Also, detailed subjective evaluation requires a lot of effort and time. Thus, a facial recognition application could be helpful to validate this kind of results more efficiently.

Therefore, the contribution of this work is twofold. One thing is to detect several important sentiments expressed by the users using a neural net. The other thing is to apply this recognition result for various evaluation, such as musical listenings or work of art evaluations.

This article is structured as follows. Section 2 contains the overall description of the system. Section 3 describes the experiments performed for the facial recognition. Section 4 explains the preliminary results obtained and Section 5 presents the conclusions and future work.

2. Facial expression recognition method

2.1 KinectV2Sensor

An automatic face expression recognition system falls into the following steps: face detection and the location in a practical scene, facial feature analysis, and facial expression pattern matching. The application requires the user to be seated in front of the Kinect V2 Sensor. Then, the device detects the human body and estimates the position of his or her head, drawing a face on this position. Once the face is isolated, 17 features are extracted to be the input of the multilayer perceptron. This classifier properly trained gives the facial expression of the user. The main tool used to extract the feature values of the main 17 points of the user's face in real time is Kinect V2 Sensor⁽⁵⁾.

We estimate the position of the face of the user using facetracking, detect and track it, and use the face API, HDface, FaceShapeAnimation Enum to acquire the feature values of 17 major points of the user's face. This Enum property has a shape unit (SU) and an animation unit (AU). SU represents the shape feature of the detected 3D face model, and AU represents the feature of the movement of the face. In this study, all values of AU value were used. The API property has 17 AUs as shown in Fig. 1.

0)JawOpen, 1)LipPucker, 2)JawSlide, 3)LipStretcherRight, 4)LipStretcherLeft,5)LipCornerPullerRight,

6)LipCornerPullerLeft,7)LipCornerDepressorLeft, 8)LipCornerDepressorRight,9)LeftCheekPuff, 10)RightCheekPuff,11)LeftEyeClosed, 12)RightEyeClosed,13)RightEyebrowLowerer, 14)LeftEyebrowLowerer,15)LowerLipDepressorLeft, 16)LowerLipDepressorRight

Most of the AUs are expressed as a numeric weight varying between 0 and 1, three of them, JawSlide, RightEyebrowLowerer, and LeftEyebrowLowerer, vary between -1 and +1. Kinect updates those AU values every frame.

Figure 1. The 17 Animation Units

2.2 Preliminary recognition experiment

Weka's multilayer perceptron was used for facial expression recognition and identification in this study. Weka (Waikato Environment for Knowledge Analysis) is a machine learning software developed at Waikato University in New Zealand. Weka is a collection of visualization tools and algorithms for data analysis and prediction modeling and has a graphical user interface that can easily handle its functions. 17 AU values obtained from Kinect V 2 sensor are used as the inputs. The output is the type of facial expression identified in this research. The types of facial expressions to be identified are six types; Neutral, Happy, Disgust, Surprise, Sad, Eye-closed. Previous research has a problem that it is difficult to distinguish between Sad and Disgust, and in this research we also confirmed the recognition rate with five types of facial expressions that surpassed Sad. In the setting of Multilayer perceptron, basically the default setting was used, the number of nodes in the hidden layer and the number of times of learning were changed, and the recognition rate was confirmed. Figure 2 shows the flowchart of the recognition system using the multilayer perceptron. The training data consists of the 17 AU values for each facial expression acquired from each subject by Kinect V2 Sensor. The subjects were 10 adult males and 7 females, asked them to make 6 kinds

of expressions, and used Kinect V2 Sensor to store 20 AU patterns for each facial expression. The data which cannot be used are deleted, and the total number of data is 1920.

Figure 2. Flow chart of facial expression recognition

Learning was performed with 80% of the learning data, and the recognition rate was evaluated in the remaining 20%. In evaluation, we changed the number of nodes in the hidden layer and confirmed the change of recognition rate iteratively in order to decide the appropriate setting. Weka's default node is represented by the following equation.

$$N = \frac{a + c}{2} \quad (1)$$

a represents the number of inputs, and c represents the number of outputs. In this study, input is 17, output is 5 or 6, so N is 11. Also, in order to measure the recognition rate with 80% (1536) of the learning data and 20% (384) of the evaluation data, the following formula is applied ⁽²⁾.

$$A_{cc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Where TP : True Positive values, TN : True Negative rates, FP : False Positive values and FN : False Negative Values.

2.3 Experimental results

The results are plotted in Fig. 6 and 7. Horizontal axes represent the number of hidden layers applied in each execution. Vertical axes represent the recognition rate (accuracy) in percentages. Color lines represent the number of iterations used. Blue line indicates 500 iterations for each number of hidden layers. Likewise, orange lines corresponds to 1000 iterations, whereas grey lines mean 1500 iterations are applied to the system. During the system test operation, we noticed that the sadness facial expression seems confusable compare with other categories. Therefore, we decided to test both cases, i.e. 5 categories and 6 categories. Figure 6 shows the recognition results in the case of 5 facial expression categories, excluding “Sadness”. Figure 7 shows the ones in the case of 6 facial expression

categories. The results showed that the classification accuracy is higher when the sad faces are excluded, although both cases have good recognition rates (above the 85% or recognition). As we can see, the classifier gives nice results when the number of hidden layers is above eleven. In the case of six faces classification, 11 layers is the best result obtained. In the first case, the best one achieved has 14 hidden layers. Thus, we can state that the classifier using between 11 and 13 layers seems good performance in any case.

It is to note that the results show quite similar results independently of the number of training iterations. We only obtain some small deviations (more than 5% in the recognition rate) in the case of fourteen hidden layers with six categories and in the case of eight hidden layers with five categories. Thus, we can conclude that the number of

Figure 3. Recognition results in the case of 5 facial expression categories

training iterations is not sensitive in terms of the

Figure 4. Recognition results in the case of 6 facial expression categories

recognition accuracy.

Although our preliminary face expression classification experiment is a sort of initial step, we were able to confirm that the proposed system setting has a potential to get user feedbacks from the facial expressions. KinectV2 sensor with HD face API seems very powerful tool to make such facial expression recognition system in a short development period. In this study, we did not use the facial motion, however, facial motion information seems very important to be able to detect the detailed expression information.

2.4 Music evaluation experiment

Based on the result of Experiment I, we asked six subjects (adult male) to listen to five classical songs for 30 seconds and got their feedbacks. At the same time, the facial expression recognition was performed. The first song was a gentle song, the second song and the fourth song were sad songs, and the third and fifth songs were fun songs. Eye-closed face was detected very frequently in all the results. From the results of the first song, Neutral is the most frequent result of the facial expression recognition, which is similar to the feedback. Happy and Sad were detected reasonably, however, it is conceivable that some misrecognition has been done because many Disgusts are detected.

In case of the second and fourth song, the result was reasonably good. From the results of the third song, the feedback shows happiness and surprise, however, we could not get the same result. From the results of the fifth song, all the feedbacks were happy, but Happy was hardly detected.

From this fact, in order to actually use a face recognition system for evaluation experiments, it is conceivable that it is necessary to construct a face recognition system with higher accuracy. Also, during the evaluation experiments, subjects listening to music do not change facial expressions so much.

3. Measuring degree of fun and excitement

3.1 Data classification

In order to understand how much is our customer enjoying and having fun and excitement, we focused on one facial expression (Happy), and detect 0 to 5 level of “happy” face, such as “smiling” → “laughing” → “loud laugh”. The subject was asked to watch a 5 minutes TV comedy show and while the subject is watching the show,

AU data was saved using Kinect V2 sensor. Then we classified the data into six clusters by Weka's k-means clustering. We confirmed that the clusters are related to the intensity of happy face. Before the clustering, most of the Neutral face images were discarded. The total number of AU vectors are 43126 after eliminating the Neutral frames. After the clustering operation, we picked up 8 important AU parameters out of 17.

3)LipStretcherRight, 4)LipStorcherLeft,
5)LipCornerPullerRight, 6)LipCornerPullerLeft,
7)LipCornerDepressorLeft, 8)LipCornerDepressorRight,
11)LeftEyeClosed, 12)RightEyeCloed

In order to confirm whether ideal classification is done by clustering, data of each cluster was extracted and examined with the actual 2D image. Also, we asked questionnaire to which subject the subject fit. The images used for the questionnaire are 12 images.

3.2 Data collection

The procedure of data collection is as follows.

- (1) Launch user's AU value extraction system with Kinect V2 sensor.
- (2) Run a video of TV comedy show to the users. (Video of Junichi Davidson of 2015 R1 Grand Prix championship)
- (3) Activate the AU value extraction system and at the same time, record subject's face with the video camera.

The number of subjects was 15 (14 adult males, 1 female), and the video was approximately 5 minutes long. Because Kinect V2 sensor is updated with FPS 30, data collected per 5 minutes for data will be 9000, but the actual number of data obtained by Kinect V2 sensor per person is about 6300, the total number of data is 87416 met. It can be thought that this is due to the fact that the operation has been delayed on the PC side. Changes in facial expression of Happy mainly are seen in 5) LipCornerPullerRight, 6) LipCornerPullerLeft. Therefore, it is ideal that this parameter is divided into six equal parts between 0 and 1. From this result, looking at the values of the centroid value of each classified result 5) LipCornerPullerRight, 6) LipCornerPullerLeft

3.3 Evaluation of clusters

Good results could be obtained by clustering with AU value of 8. Next, it is necessary to investigate whether the ideal classification is really done by comparing the result with the data taken by the actual video camera. Therefore, two parameters closer to the center of gravity value were extracted from the clustering result, and an image corresponding to that value was used from the data photographed by the video camera. A total of 12 images were prepared and asked 14 subjects to respond by questionnaire method to which intensity of each facial expression corresponds which image. The intensity is expressed by 0 to 5, where 0 means Neutral, and 5 means Happy. The questionnaire results are as follows;

Table 1. Results from the questionnaire

Image number	Correct answer number	Number of correct answers	Correct answer rate	Consider ± 1	Correct answer rate
1	5	14	100%	14	100%
2	0	7	50%	13	93%
3	4	6	42%	13	93%
4	3	8	57%	14	100%
5	1	7	50%	14	100%
6	3	7	50%	11	78%
7	2	6	42%	13	93%
8	0	13	93%	14	100%
9	5	10	71%	12	85%
10	2	8	57%	11	78%
11	1	9	64%	14	100%
12	4	8	57%	14	100%

From this result, the average correct answer rate was 61.3%. For the intensities 0 and 5 it was a relatively high percentage of correct answers, but variations were observed in the intensities 1 to 4. However, when looking at the questionnaire results, it was found that the subjects often responded with ± 1 intensity level error. In the case of discriminating the intensity divided into six stages from the image, it is considered that it is difficult to accurately discriminate since the intermediate intensity. For example, strength 3 is set to 2 or 4, etc. Considering such error, the correct answer rate was much higher, and the average score is 93.5%. From the overall evaluation, it can be considered that the six groups classified by clustering by K-means are reasonable.

4. Conclusions

In this paper, we have discussed research on facial expression recognition method for user reaction information extraction.

First of all, we constructed a system that extracts AU values from user's face information using Kinect V2 sensor, and asked the subjects to perform six types of facial expressions (Neutral, Happy, Disgust, Surprise, Sad, Eye-closed). We evaluated the recognition rate using Weka's multilayer perceptron. As a result, recognition rate of 80% ~ 95% could be obtained.

Next, we conducted music evaluation experiments using the facial expression recognition system. We asked the subjects to listen to five songs of classical music for 30 seconds each and took a questionnaire. As a result, many parameters different from the questionnaire result were detected. Therefore, it was thought that it was necessary to improve recognition rate and realize accurate facial expression recognition.

Then, we focused on the "Happy" face expression. The subject was asked to watch a video of comedy about 5 minutes, and data was collected using Kinect V 2 Sensor. In order to adjust the number of data, most of the natural data was deleted as preprocessing, and clustering was performed using Weka's K-means. Six clusters were generated based on the intensity of "Happy". Then the clustering was performed again using 8 AU values out of 17 AU values in order to perform better classification. subjects were asked to respond by questionnaire in terms of intensity level. As a result, the correct answer rate was 61.3%. However, if we consider the nearest neighboring results of the expected result as a correct answer, the score went up to 93.5%. Therefore, we think our approach seems promising.

References

- (1) Shigeru Akamatsu : "Recognition of Facial Expressions by Human and Computer", The Journal of the Institute of Electronics, Information and Communication Engineers, Vol.85, No.10, pp766-771, 2002
- (2) Vineetha, G. R., Sreeji, C., & Lentin., J. (2012). Face Expression Detection Using Microsoft Kinect with the Help of Artificial NN. *Trends in Innovative Computing* .
- (3) Youssef, A. E., Aly, S. F., Ibrahim, A. S., & Abbott, A. L. (2013). Auto-Optimized Multimodal Expression

Recognition Framework Using 3D Kinect Data for ASD Therapeutic Aid. *International Journal of Modeling and Optimization* , 3 (2), 112.

- (4) Puica, M., & Florea, A. M. (2013). Towards a computational model of emotions for enhanced agent performance. *Phd Thesis* . Romania: Univ. of Bucharest.
- (5) Zhang Z : “Microsoft kinect sensor and its effect”, IEEE MultiMedia , Vol.19, No.2, pp4-10, 2012.