Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks

Xuan-Phung Huynh
Sejong University
South Korea
phunghx@gmail.com

Yong-Guk Kim*
Sejong University
South Korea
*ykim@sejong.ac.kr

Abstract

Discriminating between genuine and fake emotion is a new challenge because it is in contrast to the typical facial expression recognition that aims to classify the emotional state of a given facial stimulus. Fake emotion detection could be useful in telling how good an actor is in the movie or in judging a suspect tells the truth or not. To tackle this issue, we propose a new model by combining a mirror neuron modeling and deep recurrent networks, called long-short term memory (LSTM) with parametric bias (PB), by which features are extracted in the spatial-temporal domain from the facial landmarks, and then boil down to two PB vectors: one for genuine and other for fake one. Additionally, a binary classifier based on a gradient boosting is used to enhance discrimination capability between two PB vectors. The highest score from our system was 66.7 % in accuracy, suggesting that this approach could have a potential for useful applications.

1. Introduction

The primary sources for nonverbal communication among human are facial expression and hand (or body) gesture. Among them, human facial expression provides a fast and subtle way of communication. Recognition of human facial expression using a computer has been a popular research topic partly because it has diverse applications, such as smart surveillance, human-robot interaction, and video search [15]. Although humans can express many different facial expressions, it is known that there are six basic facial expressions [4, 5]. Indeed, most researchers on recognition of facial expression deal with these six emotions: how well human or an algorithm recognizes the facial expression? On the other hand, discriminating between genuine and fake emotion is a rather different issue. Historically, investigators in psychology, neurology, and psychiatry have discriminated between deliberate and spontaneous facial expression [6]. Recently, determining emotional authenticity has attracted attention among some researchers. Such investigation can be useful and has the certain applications, such as determining deceiving behavior in police investigations or judging how real the actor is in the movie. However, these applications suffer from some limitations due to lack of experience on an automatic program to discriminate the authenticity of the given expression, and existing methods mainly focus on genuine smiles. Because of such reason, we believe that the present data corpus containing sincere and deceptive universal facial expression of emotion will become a milestone in this area. The Chalearn LAP challenge for real versus fake recognition first introduced the SASE-FE database that analyzed cues of deception while controlling for the emotional status of a subject as shown in Fig. 1. [16, 20]

Figure 1. An example of genuine and fake facial expressions for six emotions in the challenge dataset.

The dataset given to the present challenge has the video format, and each video contains a single facial expression. In other words, each video has only one label: either fake or genuine expression. Our initial guess was that Recurrent Neural network (RNN) could be a good candidate in dealing with such circumstance. Soon, it is found that a modified version of RNN, called RNN-PB (Parametric Bias), is a better option since the given facial expression as a video is crystallized as a parametric bias in this network. RNN-PB has managed to demonstrate that a robot imitates human's gesture after watching it: mirroring other action using the mirror neuron modeling, although hand gestures or

2473-9944/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCVW.2017.362

IEEE computer society

upper body gestures are extensively used in these studies. Such mirror neuron modeling has been primarily inspired by Rizzolatti's finding that a group of neurons in area F5 area of the monkey fires when he just watches other's hand movement as well as during the execution of his own hand movement: the neurons that imitate other's hand movement are called mirror neurons [18]. The implication of mirror neurons has been very influential and persuasive such as the social brain, autism, and neural modeling. Here, a noticeable fact is that the same area is also activated with facial expression [18], suggesting that hand gesture and facial expression may share the same functional role: imitating (or recognizing) other actions and expressing his actions. In any case, the present study is to utilize the mirror neuron modeling in discriminating between fake and genuine emotion.

Among several mirror neuron models, the architecture of RNN-PB is adopted in this study. It is based on a vanilla RNN and yet has parametric bias [9, 19], with which both bottom-up and top-down interaction processes are possible. The bias neurons gain individual value for each pattern during the learning phase. It is shown that it recognizes other's hand gesture and generates its own hand action by learning different time series patterns, although the performance deteriorates when the relational structure of the patterns becomes complex and/or the difference between patterns is subtle like the present fake emotion discrimination case.

LSTM is an updated version of RNN by solving the vanishing gradient problem. It has been successful in a variety of applications, such as speech recognition, handwriting generation and image captioning [8]. There are diverse variations of LSTM. The important component of it is the state unit that has a linear self-loop since LSTM network can learn long-term dependency more easily than the simple RNN. It allows us to store information for the extended time interval case.

The core architecture of the proposed system shown in Fig. 3 is designed to integrate advantages of RNN-PB and LSTM: Both top-down (learning and generation mode) and bottom-up (recognition mode) interactions are possible; It keeps the long term dependency and yet exploits powerful discrimination capability of LSTM against the subtle difference between fake and genuine facial expressions.

Our system detects the face in the first frame of test video, and then tracks this face on remaining frames, following extracts facial landmarks and then learns parametric bias (PB) vectors from a testing stream. Then, these PB vectors are classified using the gradient boosting machine. Fig. 3 illustrates the pipeline of our framework. LSTM-PB and the gradient boosting machine are trained using the challenge dataset, consisting of six facial expressions such as happiness, sadness, disgust, anger, contempt, and surprise. Our result is placed as the 1st place, sharing it with another team in ChaLearn LAP 2017 challenge.

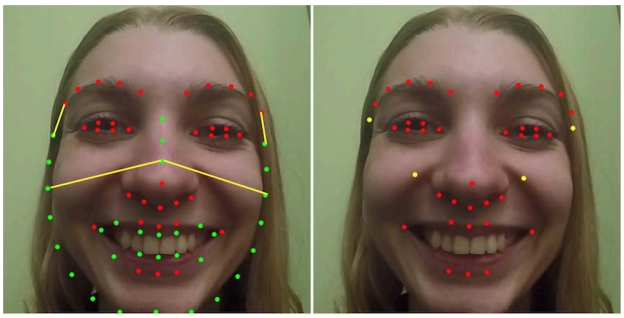

Figure 2. Facial landmarks used for the present study. (A) 64 facial landmarks detected by DLib. Among them, the red facial landmarks are adopted, whereas the green facial landmarks are removed. (B) 40 landmarks that are used in our study. Note that four yellow landmarks are added, and each yellow landmark is located in the center of each yellow line in (A).

Our contributions are as follows:

- LSTM-PB is proposed by combining mirror neuron modeling, *i.e.* RNN-PB, with Deep Recurrent Neural Network, *i.e.* LSTM.
- This would be the first mirror neuron modeling by which we solve a classification problem, recognizing a fake facial expression.
- To enhance discrimination capability of LSTM-PB, a strong binary classifier, *i.e.* the gradient boosting machine, is added.

The remainder of the paper is as follows. In Section 2 we introduce our method by describing facial landmarks and LSTM-PB components. In Section 3 we present the data set and experiments and discuss our framework and its performance on the challenge. Section 4 concludes our work on fake/real emotion recognition.

2. Proposed method

2.1. Face detection and tracking of ROI

The present database contains many frames wherein a subject expresses her emotions in front of a high-speed camera. Here, the upper part of a torso is seen continuously during a session. Since we need only the facial area within a frame, the face detection is carried out by combining a Haar-feature face detector [14] and an MOSSE-based object tracker within the OpenCV environment [2]. MOSSE (Minimum Output Sum of Squared Error) tracker is known to be robust against variations of illumination, scale and pose while operating at high speed [1]. When the face detector detects a region of interested (ROI) in the first frame, and

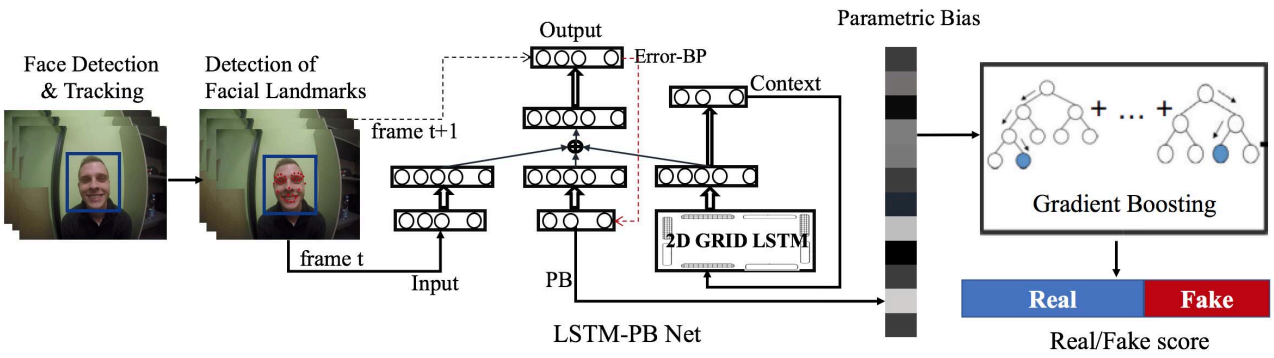

Figure 3. The pipeline of our framework for real versus fake expression recognition.

then MOSSE tracks this ROI from the second frame to the final one.

2.2. Design of facial landmarks

After the face detection process, the facial landmarks are detected using DLib library [13], that implements an ensemble of regression trees for detecting landmarks [12]. For each face, although the default option is to extract 68 landmarks, some landmarks located along the chin and inner mouth are removed to reduce the complexity of dataset as shown in Fig. 2. Additionally, landmarks along the inner lip are also removed since they are corresponding with the remained points on the lip. On the other hand, two landmarks at cheeks and two at the end of eyebrows are added because it is found that movement around cheeks and landmark at the end of eyebrows provide relevant information while expressing an emotion. There are four yellow lines, and each yellow landmark is located in the center of each yellow line in Fig. 2.

2.3. Long-short term memory with Parametric bias

The basic architecture of RNN-PB is a Jordan-type recurrent feed forward neural network [10]. The difference is that it has PB nodes in the input layer. Unlike the other input nodes, these PB nodes take a particular constant vector at each time sequence. They have a mapping between the fixed-length vectors and time sequences. In other words, PB nodes encode the time sequences throughout the self-organizing process. Like RNN-PB, LSTM-PB learns time sequences in a supervised manner. The only difference is that the 2D-Grid LSTM is used to store memory information during the learning process. The backpropagation through time (BPTT) algorithm is utilized in training the structural properties of the training time sequences. Meanwhile, PB vectors encode the specific properties of each time sequence simultaneously. As a consequence of the learning process, the LSTM-PB self-organizes a mapping between PB vectors and time sequences.

To learn the PB vectors, we utilize a variant of the BPTT algorithm. This procedure adjusts the intrinsic values of the PB layer and holds the weights of their outgoing connection fixed to update PB vectors; we accumulate the back propagated errors concerning the PB nodes for all time steps. Formally, the PB vector p_{x_i} encoding the i-th training time sequence x_i is updated as follow

$$\delta p_{x_i} = \frac{1}{l_i} \sum_{t=0}^{l_i - 1} error_{p_{x_i}}(t)) \tag{1}$$

$$p_{x_i} = p_{x_i}^{old} + \delta p_{x_i} \tag{2}$$

In equation 1, the average back propagated error concerning a PB node through all time steps via BPTT algorithm, and the vector is the update of PB values as equation 2. As the original RNN-PB, the LSTM-PB can generate a time sequence of its corresponding PB vectors. This generation process utilizes the LSTM-PB with the appropriate PB vector, a fixed initial context vector, and input vectors. In our experiment, the external information, facial landmarks, is employed as input vectors.

Also, the LSTM-PB can be used not only for sequence generation process but also for recognition processes. Using these equations 1-2, we obtain the corresponding PB vectors for a given sequence. The learning process extracts the relational structure, the most important characteristic nature of the LSTM-PB, among the training time sequences in the PB space. In another word, LSTM-PB is the representative of the dynamical system approaches as the original RNN-PB since the properties of the mirror neurons [17]. LSTM-PB also has three operational models: learning, generation, and the recognition mode.

2.3.1 Learning mode

Training of LSTM-PB is carried out using labeled facial expression videos, consisting of genuine and fake emotions. As a result, two parametric biases are created: one is for the genuine emotion and the other for the fake one. The

goal of training is to update weight sets such that the network becomes a time series predictor for the facial stimuli, and to create two PB vectors, corresponding to the genuine and fake emotion, respectively. Since the learning process is based on the prediction error (or mean square error), the BPTT method is thoroughly used in adjusting weights of the network for all training patterns. Similarly, PB vector is updated for each training pattern to reduce the prediction error, and yet the variation of PB is kept slow to obtain a constant PB value in the end.

2.3.2 Generation mode

When the learning is completed, the network in the generation mode can produce a stream of facial landmarks corresponding to either a genuine or a fake facial expression, depending on the given PB vector. Since the command to the whole network is given by PB vector, the generation mode operates in the top-down interaction. Note that there is no change in weights of the network during the generation mode.

2.3.3 Recognition mode

In the recognition mode, LSTM-PB observes the given stream of facial landmarks and computes a PB vector that matches with pre-trained one. When the network makes an initial prediction, the error between the prediction and the target is generated at the output layer. Then, the prediction error is back-propagated to the PB vector in term of mean square error. If pre-learned facial landmark movement patterns are perceived, the PB values tend to converge to the values that have been determined in the learning phase [9].

2.3.4 2D Grid LSTM

LSTM is based on the gated recurrent unit, and it is one of the most efficient sequence models used for the practical applications. Original LSTM model uses self-loops to produce paths so that the gradient can flow for a long duration. The Grid LSTM is designed to utilize the long-short term memory storage efficiently, and it can be created by stacking LSTM cells in a multidimensional grid [11]. One of the major goals of Grid LSTM is to create a unified way of using LSTM for both in-depth and sequential computations so that it can be a better solution for this problem by locating cells as the multidimensional blocks including the depth of the network. A N-dimensional block has input that consists of N-hidden vectors and N-memory vectors, and it has output that also consists of N-hidden vectors and N-memory vectors shown in Fig. 4. Each block has separate weight matrices. Each grid has incoming N-hidden and N-memory vectors and outgoing N-hidden and N-memory vectors. In the present study, a 2D Grid LSTM is used as shown in Fig. 4.

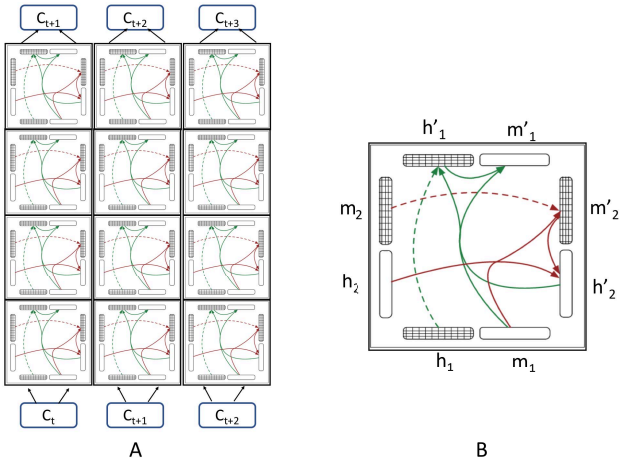

Figure 4. 2D Grid LSTM adopted in our LSTM-PB. (A) The layout and time sequence of the 2D Grid LSTM, consisting of four blocks. This 2D Grid LSTM is inserted between the context out and the context input in Fig. 3. (B) The inner structure of a 2D block where four hidden and memory vectors are positioning along four sidewalls.

2.3.5 The architecture of LSTM-PB

Fig. 3 illustrates the schematic architecture of our system wherein LSTM-PB plays a central role. It has three inputs: facial landmarks, parametric bias and context input. The facial landmark input has 80 nodes since they correspond to (x,y) coordinates for 40 landmark points. Parametric bias, as well as context input, has 64 nodes, respectively. Both facial landmarks input and parametric bias are fully-connected to its hidden layer while context input goes into 2D Grid LSTM as a vector, and output of 2D Grid LSTM is fully-connected to the hidden layer. In this depth level, each hidden vector has 512 nodes. The output of LSTM-PB is computed by fully-connecting the last hidden layer. The mean square error between the prediction and the value from the facial landmarks on the next frame is propagated through the time over the network. During the learning mode, the modification of weights in the network is carried out including the 2D Grid LSTM. In addition, the internal state of PB layer changes depending on the error back-propagation from the output.

2.3.6 Details of learning

Our LSTM-PB network is trained for six different facial expressions separately simply because the challenge database consists of six separate emotions. The weights of network and the PB vector are updated as follows: Stochastic Gradient Decent method is used in optimizing the weights with

a learning rate of 0.01 for 100 epochs; Similarly, PB node is updated with the learning rate of 0.9. Both learning rates are decreased by a factor of 10 after ten epochs when the accuracy of the network does not improve on the training set. The weights are initialized in the form of uniform distribution. Our 2D-Grid LSTM is also trained according to the setup of [11]. Both initial values for parametric bias and context vector are set to zero during the learning and recognition mode.

2.4. Classification with Gradient Boosting

Gradient Boosting Machine (GBM) is a statistical framework which includes AdaBoost and related algorithms [7]. In machine learning, boosting utilizes a group of weak learners to enhance and complement ability of the learning model. It carries out such task by proposing a hypothesis in succession by revising the previous hypothesis that has been unsuccessful. GBM minimizes the loss of the model using a gradient descent like procedure, and it is a stage-wise addition model. Therefore, it can handle arbitrary differential loss function. Although the initial purpose of this framework is for regression, it can be used for binary and multiclass classification. GBM contains three components: (1) a loss function that needs optimization, (2) a weak learner to predict, and (3) an additive model to combine weak learners to minimize the loss function. The first component, *i.e.* loss function, should be differentiable and it varies depending on the given problem. GBM greedily constructs a forest of trees by selecting the best split points based on scores of the loss function, and yet the existing trees in the model are not changed while adding the trees. Normally GBM quickly overfits a training data set because it is a greedy algorithm. Therefore, it needs to regularize the basic gradient learning by decreasing the learning rate.

The final goal of this study is to discriminate between a fake and the genuine facial expression. Although LSTM-PB provides a useful framework by which a group of facial landmarks of a facial expression is transformed into a PB vector, we found that determining whether a PB vector from the LSTM-PB belongs to a fake or the genuine facial expression is not a trivial problem. Since it is known that GBM is a powerful classifier, we adopt it for the present binary classification task. Among several publically available gradient boosting libraries, one of them is used [3], where the weak learners are implemented as a decision tree. For each emotion, 5000 boosted trees are generated. Although the training set contains 40 subjects, data augmentation is required for training LSTM-PB and gradient boosting machine since we need more data for training them. Augmented data are created by generating facial landmarks on the different cropping face. One way to do it is to choose a random number, then use it in removing that amount of pixels on edge for a detected face by our face detection system, and generate the corresponding facial landmarks on this image in a repetitive fashion.

Data	Number of labels	Number of videos	Number of subjects	Labels provided
Training	12	480	40	Yes
Validation	12	60	5	No
Testing	12	60	5	No

Table 1. The summary of the dataset.

3. Dataset and system implementation

3.1. ChaLearn dataset for real versus fake emotion

The main challenge is, of course, to discriminate between fake and genuine emotion. As far as we know, this is the first kind of challenge where genuinity of the given emotion has to be determined. Fifty subjects participate in the video recording, and 60% of them are males and 40% females, respectively, with age of 19-36. A variety of ethnic ancestries such as African, Asian and Caucasian are included. Each subject performs six facial expressions: angry, happy, sad, disgust, contempt, and surprise. A subject watches a video which is meant to induce a specific emotional state in her mind, and then she expresses it into a facial expression accordingly. In each video, subject starts with a neutral emotion and then expresses her either fake or genuine facial expression. The challenge dataset contains 600 videos since 50 subjects express both fake and genuine facial expressions. A GoPro-Hero camera is used for the recording, and yet the length of each video varies. Experimental psychologists have supervised the process closely to achieve a realistic and reliable dataset. Table 1 shows the summary of the dataset. While preparing the database, external factors such as personality or mood of the subjects have been ignored [16].

3.2. Implementation of our system

Face detection and object tracking are implemented within the OpenCV environment using Python. Training of six LSTM-PBs is carried out with Nvidia Titan X within the Torch framework. Training the network takes about five hours and training a gradient boosting machine about 0.5 hours for each expression, respectively. The software is written using Lua and Python. The source codes and preprocessing data are publicly available at https://github.com/phunghx/Real_Fake_Expression.

4. Results

The challenge consisted of two phase: development (or validation) and testing phase.

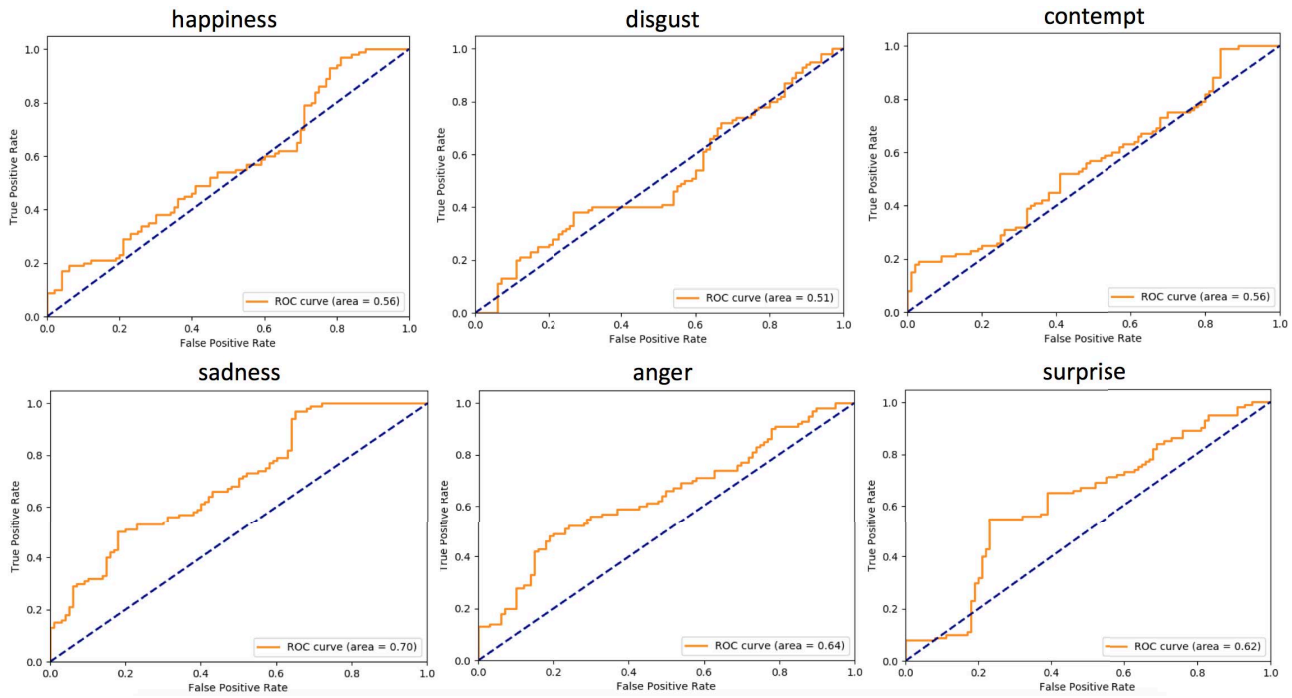

Figure 5. The ROC curves for six emotions.

Rank	Team	accuracy (%)
1	NIT-OVGU	76
2	**HCILab (ours)**	71
3	innovwelt	63
4	TUBITAK UZAY-METU	61
5	faceall Xlabs	58
6	ICV Team	53
7	BNU CIST	53

Table 2. Development results. Ours is second one.

Rank	Team	accuracy (%)
1	**HCILab (ours)**	**66.7**
1	NIT-OVGU	66.7
3	TUBITAK UZAY-METU	65
4	BNU CIST	61.7
5	faceall Xlabs	51.7

Table 3. Final results. Ours is the fist place with another team.

4.1. Development phase

During the validation phase, the teams developed their algorithms and submitted the results to challenge for validating. There are unlimited times of submission before the deadline. Table 2 shows the results of this process. We are the second performance at the end of this phase with 71 % accuracy.

4.2. Test phase

For the test phase, the organizers released the validation labels and granted the access to the test videos but without labels. The teams submitted their results on the test videos to the competition server. The scores were updated after a submission. Each team allowed 12 times of submission during this phase. We used our framework to obtain our results. Table 3 depicts the final results for this challenge. NIT-OVGU team and ours are same performance and shared the first place with 66.7 % accuracy. Compared to the validation phase, our method is more robust than the NIT-OVGU team since our performance has small decrease with validation set. We also compute the ROC curve to compare the performance between emotions as Fig 5. In the literature of the recognition of facial expressions, accuracies for three emotions such as happy, sad, surprise are generally higher than others emotions such as anger, contempt, and disgust, presumably because the former emotions are bigger than the latter ones. However, for the discrimination between fake and genuine emotion in the present case, accuracies for three emotions such as sadness, anger, and surprise are higher, suggesting that it is easier to detect fakeness for three cases than other cases. A possible interpretation would be that different facial muscles are used in expressing three emotions between the fake and genuine facial expression cases.

5. Conclusions

Telling whether the subject in the present challenge dataset makes a fake emotion or not is a tough task even for human beings. Many different approaches are possible in tackling this problem such as the conventional handcraft feature extracting method, multi-layered deep neural network, and other machine learning techniques. However, we thought that the brain-inspired neural modeling could provide a useful stepping stone in solving such difficult issue. Mirror neuron system has been a major issue in neuroscience as well as in the diverse academic areas, and recently some researchers in robotics have been using it for the interesting demonstrations. As far as we know, the present study is the first attempt where mirror neuron modeling method is mainly used for the recognition problem. RNN-PB has been successful in modeling hand or body gesture and the corresponding action by a robot, and yet it is found that it has some limitation in dealing with subtle and complex input such as facial landmarks from a facial video. One more difficulty comes from the fact that each video has only one label throughout the whole frames. The proposed LSTM-PB can deal with longer and complex input well because a 2D Grid LSTM is adapted with RNN-PB and the unit number of our network is increased very much to afford such a unique case. Given that the present system needs to classify PB vector into either a fake emotion or not for a higher accuracy, a powerful classifier such as the gradient boosting machine is necessary, but any other decent classifier can do the job. Additionally, although a facial landmark detection scheme is adopted for the present work, we plan to test other options such as a high performing face model like DB-ASM or a face modeling based on the deep neural network. Fake emotion detection will be a new research area because it is hard to do it but has many interesting applications. It is believed that the present ChaLearn dataset shall be an interesting testbed with which various experiments on recognition of deceptive facial expressions can be done.

Acknowledgement

This work was supported by Institute for information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT)(No.2016-0-00498, User behavior pattern analysis based authentification and abnormally detection within the system using deep learning techniques) and (No.2017-0-00731, Personalized Advertisement Platform based on Viewers Attention and Emotion using Deep-Learning Method)

References

[1] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2544–2550. IEEE, 2010.

[2] G. Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[3] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[4] P. Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, 1992.

[5] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3/4):169–200, 1992.

[6] P. Ekman, J. C. Hager, and W. V. Friesen. The symmetry of emotional and deliberate facial actions. *Psychophysiology*, 18(2):101–106, 1981.

[7] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[8] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[9] M. Ito and J. Tani. On-line imitative interaction with a humanoid robot using a dynamic neural network model of a mirror system. *Adaptive Behavior*, 12(2):93–115, 2004.

[10] M. I. Jordan and D. E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive science*, 16(3):307–354, 1992.

[11] N. Kalchbrenner, I. Danihelka, and A. Graves. Grid long short-term memory. *arXiv preprint arXiv:1507.01526*, 2015.

[12] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[13] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[14] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–I. IEEE, 2002.

[15] I. Lillo, J. Carlos Niebles, and A. Soto. A hierarchical pose-based approach to complex action understanding using dictionaries of actionlets and motion poselets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1981–1990, 2016.

[16] I. Ofodile, K. Kulkarni, C. A. Corneanu, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari. Automatic recognition of deceptive facial expressions of emotion. *arXiv preprint arXiv:1707.04061*, 2017.

[17] E. Oztop, M. Kawato, and M. Arbib. Mirror neurons and imitation: A computationally guided review. *Neural Networks*, 19(3):254–271, 2006.

[18] G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27:169–192, 2004.

[19] J. Tani. Learning to generate articulated behavior through the bottom-up and the top-down interaction processes. *Neural Networks*, 16(1):11–23, 2003.

[20] J. Wan, S. Escalera, X. Baro, H. J. Escalante, I. Guyon, M. Madadi, J. Allik, J. Gorbova, and G. Anbarjafari. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenge. *ChaLearn LaP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV, 2017*.