Emotion in Robots Using Convolutional Neural Networks

Mehdi Ghayoumi[✉] and Arvind K. Bansal

Artificial Intelligence Lab Computer Science Department,
Kent State University, Kent, USA
{mghayoum, akbansal}@kent.edu

Abstract. These years, emotion recognition has been one of the hot topics in computer science and especially in Human-Robot Interaction (HRI) and Robot-Robot Interaction (RRI). By emotion (recognition and expression), robots can recognize human behavior and emotion better and can communicate in a more human way. On that point are some research for unimodal emotion system for robots, but because, in the real world, Human emotions are multimodal then multimodal systems can work better for the recognition. Yet, beside this multimodality feature of human emotion, using a flexible and reliable learning method can help robots to recognize better and makes more beneficial interaction. Deep learning showed its force in this area and here our model is a multimodal method which use 3 main traits (Facial Expression, Speech and gesture) for emotion (recognition and expression) in robots. We implemented the model for six basic emotion states and there are some other states of emotion, such as mix emotions, which are really laborious to be picked out by robots. Our experiments show that a significant improvement of identification accuracy is accomplished when we use convolutional Neural Network (CNN) and multimodal information system, from 91 % reported in the previous research [27] to 98.8 %.

1 Introduction

Emotions have main role and are affected in developing any type of social setting and humans are social and live socially and most of the actions are emotional. Human emotional states (expression and recognition) have been the focal point of attention in several areas of neuroscience and psychology to cognitive and computer science. For the acceptance of robots by humans the application of emotions for Human Robot Interaction (HRI) purpose are very significant. A robot that is able to realize and express emotions can pass on in a lifelike way. The observation of different modalities, such as facial expression, gesture, and speech, improves the emotional state recognition. Moreover, recognizing the emotion is a complicated process and there are some researches which looking for recognizing real emotion. In our previous work, we deployed group theory concept of recognizing real emotion by detecting symmetry patterns in face [15].

GU and et al. [2] analyzed and explored the importance and the use of the information in each trait which are efficient in human emotional states. They found out when

we would wish to recognize emotional states, non-verbal communication, facial expressions and body posture/motion complement each other. Adolphs [3] showed how the human brain correlates past experiences, motion information in the visual stimuli, and face expressions. The brain is able to integrate this multimodal information and generate a theatrical performance of the visual stimuli based on all of them in concert. The pretense of this operation in computer systems can be achieved by neural models, with a specific social system that has different type of feature representations such as Convolutional Neural Networks (CNN).

CNN were introduced formally by Lecun, et al. [4]. They are prompted by the hierarchical process of simple and complex cells in the human learning ability to extract and learn different information from visual stimuli. Each layer of a CNN has the capability to react to different information, and when stacked together the layers can create a complex representation of the optical input.

In our recent works we presented a multimodal architecture for emotion in robot and we broke down what it has in mind for a robot to have emotion and distinguishing emotional state for communication from an emotional state as a mechanism for the formation of its behavior with humans and robots by (CNN) [5, 6]. In this clause, we plan to implement the given model and compare the results with our previous works.

This paper is coordinated as follows: The next section explains the related works. Section 3 describes human and robot emotion. The relation between deep learning and emotion is given in Sect. 4. In Sect. 5 we demonstrate the integrated model. In Sect. 6 we present experimental results and stopping points and future works are shown in the final part.

2 Related Works

The research study by Mehrabian [14] has indicated that 7 % of the communication data is transferred by linguistic language, 38 % by paralanguage, and 55 % by facial expressions in human face-to-face communication. Some models of multimodal databases can be found in [7–9] and most studies, have looked at the integration of facial looks and speech information and there have been a few efforts to fuse data from body movement and motions in a multimodal framework. Sun et al. [25] designed hidden identity features with deep convolutional networks to realize approximately 1000 false identities on LFW database and achieved 97.45 % verification accuracy with only weakly aligned faces. El Kaliouby and Robinson [11] offered a model to make head movements and facial expressions state information. Susskind et al. [23] took advantage of learning deep belief nets to classify facial action units in realistic face images. Krizhevsky et al. [24] used the deep convolutional neural network to classify the 1.2 million images in the ImageNet LSVRC-2010 contest in 1000 different categories and achieved the inconceivably higher accuracy than the temporal state-of-the-art. Gunes and Piccardi [10] fused facial expressions and body gestures information for bimodal emotion recognition. For identification purposes, almost, all types of machine learning techniques have been used in emotion recognition approaches [12, 13]. For many reasons and mainly for our final goal of creating an emotion in robots as much as similar to human emotion, we are looking for learning method which

can satisfy these parameters. Lately, CNN showed up good results in biometrics, particularly in facial expression and speech recognition. We decided to use it do some preprocessing of data before feeding to the algorithm, such as LSH to prune the database data space [22, 26]. We present a multimodal CNN-based model for automatic emotion recognition and expression. Our model deploys the CNN method, and uses it for multimodal emotional state recognition using facial expression, gesture and speech recognition. This information indicates that, the facial expressions give a great amount of data in human communication. Deploying different modalities and multimodal systems, such as body position, gestures and speech, improved the determination of the emotional state.

3   Human and Robot Emotion

The Human emotional state causes the focus of attention in several areas from biology, neuroscience and psychology to cognitive and computer science due to its importance in human communication, interaction and social dealings. Here, we explain a little about Neuromodulation and Cognitive parameters and their relation with emotion.

3.1   Neuromodulation

Neuromodulation refers to the action on nerve cells of endogenous substances called neuromodulators. Three main neuromodulator systems involved in emotion are:

- **Dopamine** based communication and motor activation,
- **Serotonin** based regulation of conduct,
- **Opioid** based regulation and relaxation [16, 17].

Emotion can be regarded as continuous patterns of neuromodulation of certain lots of brain structure. All EE and ER functionalities are related to the special activities in the brain, for example for facial expression, the smiles are initiated in the motor cortex and routed via the pyramidal motor system. If we would like to simulate the EE in the robots, knowing about these parameters in details and their weights on the emotional state types for simulating the human emotion in robots can assist us. In the following study, in the future, we plan to utilize these parameters and their weights for making the model more flexible.

3.2   Cognition

Robot learning process steps (here, EE and ER) should be very similar to human and it needs to include cognition. In that respect are several integrated cognitive architectures trying to develop all aspects of conduct as a single system while remaining constant across different domains [18]. More or less of these cognitive architectures are biologically inspired, while some others are inspired by psychological theories, in which some of them also contain the concept of effect in their intent. There are the interplay of

affect (value), motivation (action tendencies), cognition (meaning), and behavior at three levels of information processing:

- **Reactive:** a hard-wired release of fixed action patterns and an interrupt generator.
- **Routine:** the locus of unconscious well-learned automatized activity and primitive and unconscious emotions.
- **Reflective:** the home of higher-order cognitive functions.

Based on the traditional approaches, cognition emphasizes on information processing which normally has excluded emotion. On the other hand, new growth of cognitive neuroscience as an inspiration for understanding human cognition has highlighted its interaction with emotion. Probes into the neural systems underlying human behavior demonstrate that the mechanisms of emotion and knowledge are intertwined from early perception to abstract thought. These findings suggest that the classic division between the subject of emotion and knowledge may be unrealistic and that an apprehension of human cognition involves the consideration of emotion. Emotions influence fundamental processes mediating high level cognition such as:

- Attention speed, duration and capacity,
- Working memory speed and capacity,
- Long term memory recall and encoding.

It is also apparent that cognition divided functions into different domains, such as memory, attention, and reasoning. The concept of emotion causes a structural architecture that may be similarly diverse and complex.

4   CNN and Emotion

Deep learning can be employed in robots and build the robot emotions more realistic and HRI & RRI better. Deploying different modalities and multimodal systems, such as facial expression, gestures and speech, improved the determination of the emotional state.

4.1   Facial Expression Recognition

Studies on facial expression recognition have been lasting for three decades since 1970s. Paul Ekman et al. [1] postulated six cross-cultural, basic emotions (anger, disgust, fear, happiness, sadness, and surprise) from a psychological view, and developed Facial Action Coding System (FACS) to describe facial micro-expression [19]. Our work also selects the six basic emotions and neutral emotion as our measure of facial expression classification. In general, for facial expression recognition system, there are three basic parts:

- **Face detection:** Most of face detection methods can detect only frontal and near-frontal views of the fount. Viola and Jones [20, 21] utilized a lot of rectangular features to find facial expressions in real time.

- **Facial feature extraction:** Sorts of features (geometric features, show features and hybrid features of geometric and appearance features) are drawn out for recognizing facial expression.
- **Facial expression recognition:** In facial expression recognition, there are dissimilar methods. Due to lack robust features, most of facial expression recognition models work poorly in the complex environment [22].

In recent years, deep learning arouses academia and industrial attentions due to its magic in computer vision. Our work is taking advantage of deep models to extract robust facial features and translate them to recognize facial emotions. FACS system analysis [26] has been employed to derive the features-details that are important during the formulation of a specific facial expression. There are 13 moving-points (11 active points and 2 passive points) and 6 non-moving reference points. The FAUs have been rendered to the corresponding feature-level movements as given in Table 1. We denote vertical-up motion by ↑, vertical-down motion by ↓, horizontally stretched outwards by '⟷', horizontally compressed inwards by '←⟨', oblique-stretched downwards by '↘', oblique-stretched upwards by '↗'. If the emotion is symmetric, then the superscripts L (left) and R (right) have been excluded. If the move is optional or shows a higher intensity increase and so it has been ranked inside the square brackets '['… ']'. Junction is shown using concatenation. Disjunction is shown using vertical bar '|'. Essential feature-point are within parenthesis '('… ')' separated by ','. The details presented in our previous research [27].

Table 1. Feature Point Displacements (FDP)

Facial Expressions	Major Feature-points displacements
Anger::	$(e_1 \leftarrow\!\langle\ el\uparrow, e_3\uparrow) + [e_2\uparrow] + [m^T\uparrow\ m^B\uparrow]$
Disgust::	$(m^T\uparrow ch\uparrow) + [m^L\downarrow|\ m^R\downarrow] + [m^B\uparrow]$
Fear::	$(e_1\uparrow, m^L\downarrow\ m^R\downarrow) + [m^T\downarrow] + [e1\leftarrow\!\langle]$
Happiness::	$(m^L\nearrow m^R\nearrow, m^T\uparrow m^B\downarrow\ ch\downarrow m^L\!\longleftrightarrow\! m^R\!\longleftrightarrow\!)$
Sadness::	$(el\downarrow\ m^L\!\longleftrightarrow\! m^R\!\longleftrightarrow\!) + [ch\downarrow]$
Surprised::	$(e^1\uparrow e^2\uparrow e^3\uparrow el\uparrow ch\downarrow) + [m^T\uparrow m^B\downarrow]$

4.2 Speech Recognition

Human language encodes emotional information in two different ways:

- What is said? And
- How it is said?

And then a spoken message can be split down into two sections:

- A semantic and
- A paralinguistic one.

Several approaches to recognize emotions from speech have been reported [28–30]. Voice communication systems should be able to treat the non-linguistic information such as emotions, along with the message. For instance, words associated with happiness are

characterized by longer utterance duration, shorter inter-word silence, and higher pitch and energy values with more extensive scopes. In sad sentences, the vitality and the pitch are usually held at the same point. Thus, these emotions are hard to be separated. We possess three important speech characteristics to model emotional speech:

- The standard deviations and ranges;
- Maximum, minimum and median values of the pitch; and
- Energy.

The deep neural network trained itself and resolves the complex problems based on the knowledge available. Resolutions of the individual groups, as considerably as a combined set, have led to the following assumptions: among acoustic features duration and energy appear to be most relevant, while voice quality showed less impact. However, no single group outperformed the pool of all acoustic features. In our experiments we restricted the set of features to those that can be extracted in real time and in a fully automatic mode.

4.3   Gesture Recognition

Gestures are expressive and meaningful questions, involving hands, face, head, shoulders, and/or the complete human body. Gesture recognition has a wide scope of applications, such as sign language for communication among the disabled, lie detection, monitoring emotional states or stress levels of studies, and navigating and/or manipulating in virtual environments. Recognition of emotion from gestures is challenging as there is no generic notion to represent a subject's emotional state by his or her gestures. Further, the gestural pattern has a wider variation depending on the subject's geographical origin, acculturation, and the power and intensity of his or her looks. Motions can be static, seeing a single pose or dynamic with a pre-stroke, stroke, and post-stroke phases [31]. Automatic identification of continuous gestures requires temporal segmentation. The most common gestural pattern, frequently used in emotion identification, is the hand movements. Glowinski et al. [32] proposed an interesting technique for hand (and head) gesture analysis for emotion recognition. Camurri et al. [33] classified expressive gestures from the human full body movement during the carrying into action of the subject in a dance. They identified motion cues and measured overall duration, contraction index, quantity of motion, and motion smoothness. On the base of these motion cues, they designed an automated classifier to classify four emotions (anger, fear, sadness, and happiness). Castellano et al. [34] employed hand gestures for emotion recognition.

5   Integrated Model

Figure 1 shows the integrated model which has both EE and ER for emotion in robots and creating better HRI and RRI [27]. For emotion recognition part, the data will come to CNN and the fusion will be answered based on their weight on human robot interaction and then we can count on the accuracy. For instance, if it receives 75 %
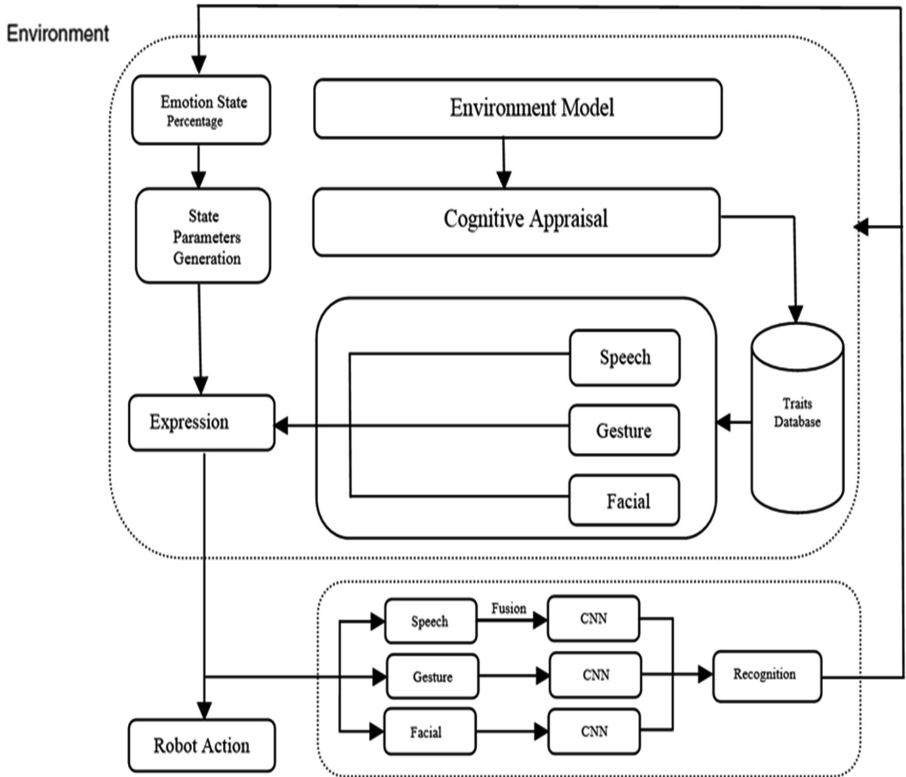
Fig. 1. Integrated model

from speech and 95 % by facial expression and 80 % of gesture, then grounded on their weights (for example here Mehrabian: 7 % by linguistic language, 38 % by paralanguage, and 55 % by facial expressions) they should multiply by these values and the average is the least accurate. On the other hand, for emotional reflection, established on the emotion recognition and cognitive appraisal, the scheme will force away the data from databases for words, gesture and facial expression which are more linked to the emotion recognition state that is recognized in the old state.

For ER and EE parts, we used the Decision Level Fusion of data in the ER part and Diffusion in the EE part. In decision level fusion each modality is first pre-classified independently, i.e., each biometric trait is captured, and features are then drawn out from that captured trait, based on that extracted features. The final classification is established on a merger of the yields of different modes. This is the highest stage of fusion with respect to human interface. In other words, the decision from each bio-metric system is concluded to construct the final determination [35].

6   Experimental Results

Table 2 indicates the confusion matrix of the emotion recognition system based on facial expressions. The overall functioning of this classifier was 80.4 %. Table 3 shows the performance of the emotion recognition system with respect to gesture analysis. The overall execution here is 86 %. Table 4 displays the confusion matrix of the emotion recognition system based on language. The overall execution of this classifier is 83 %. Table 5 shows the performance of the system with decision level integration using the best probability approach and 98.8 is overall accuracy.

Table 2. Confusion matrix for facial expressions

	Anger	Happy	Sad	Surprise	Disgust	Fear
Anger	**88.3**	1.2	0.9	4.4	1.5	0.8
Happy	3.7	**75**	4.7	2.5	0	4.1
Sadness	1.2	0.8	**82**	1.4	1.2	4.8
Surprise	3.2	2.3	1.8	**85.7**	2.6	0
Disgust	1.9	0.5	3.8	1.4	**72.4**	0
Fear	0.7	1.3	0	2.7	2.8	**79.1**

Table 3. Confusion matrix for gesture

	Anger	Happy	Sad	Surprise	Disgust	Fear
Anger	**98**	2	0	.8	0	0
Happy	.7	**74.2**	0	5.7	0	0
Sad	0	0	**73.8**	1.2	6.1	0
Surprise	0	2.1	0	**81.5**	5.3	0
Disgust	1.7	0	5.9	.6	**87.5**	0
Fear	0	0	0	0	3.6	**99.5**

Table 4. Confusion matrix for speech

	Anger	Happy	Sad	Surprise	Disgust	Fear
Anger	**98.2**	0	.6	0	.5	0
Happy	1.9	**64.2**	0	3.1	.7	2.1
Sad	1.9	0	**86**	1.3	0	.8
Surprise	2.3	1.3	0	**83**	1.6	0
Disgust	3.1	1.1	2.2	1.8	**71**	0
Fear	0	.7	0	.9	1.3	**93**

Table 5. Decision level fusion

	Anger	Happy	Sad	Surprise	Disgust	Fear
Anger	**99.7**	0	0	0	1.1	0
Happy	0	**99.3**	0	1.7	0	0
Sad	.9	0	**98.3**	0	.8	.7
Surprise	1.8	1.9	0	**97.4**	.6	1.2
Disgust	1.2	0	.9	.8	**97.8**	.9
Fear	0	0	0	1.8	1.7	**96.1**

7 Conclusion

We implemented the model for six basic emotion states and there are some other states of emotion, such as mix emotions, which are really laborious to be picked out by robots. We implemented our multi-modal system for automatic emotional state recognition. The proposed model achieves a more respectable performance when multimodal information is applied, in this case composed of facial expression, speech and gesture. The suggested model is able to learn from three different data streams: speech, facial expression and gesture. It deploys the CNN for better scholarship and identification. The results show more honest performance by comparing with old method. Our experiments show that a significant improvement of identification accuracy is accomplished when we use convolutional Neural Network (CNN) and multi-modal information system, from 91 % reported in the previous research [27] to 98.8 %. For future study, we plan to run along a mix emotion and test it on it and then enforce the model in a real-world scenario with a Telepresence Robot. We plan to move and test it on, double [36].

References

1. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124–129 (1971)
2. Gu, Y., Mai, X., Luo, Y.-J.: Do bodily expressions compete with facial expressions? Time course of integration of emotional signals from the face and the body. PLoS One **8**(7), 736–762 (2013)
3. Adolphs, R.: Neural systems for recognizing emotion. Current Opinion in Neurobiology **12**(2), 169–177 (2002)
4. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
5. Ghayoumi, M., Bansal, A.K.: Architecture of Emotion in Robots Using Convolutional Neural Networks. RSS, USA (2016)
6. Ghayoumi, M., Bansal, A.K.: Multimodal architecture for emotion in robots using deep learning. In: Future Technologies Conference, San Francisco, United States (2016)
7. Gunes, H., Piccardi, M.: A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: Proceeding of ICPR 2006 the 18th International Conference on Pattern Recognition, Hong Kong, China (2006)

8. Bänziger, T., Pirker, H., Scherer, K.: Gemep - Geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions. In: Deviller, L., et al. (eds.) Proceedings of LREC 2006 Workshop on Corpora for Research on Emotion and Affect, pp. 15–19, Genoa (2006)
9. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: towards a new generation of databases. Speech Commun. **40**(1), 33–60 (2003)
10. Gunes, H., Piccardi, M.: Bimodal emotion recognition from expressive face and body gestures. J. Network Computer Appl. **30**(4), 1334–1345 (2006)
11. el Kaliouby, R., Robinson, P.: Generalization of a vision-based computational model of mind-reading. In: Proceedings of First International Conference on Affective Computing and Intelligent Interfaces, pp. 582–589 (2005)
12. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Process. Magazine **18**(1), 32–80 (2001)
13. Pontiac, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1424–1445 (2000)
14. Mehrabian, A.: Silent Messages - A Wealth of Information about Nonverbal Communication (Body Language). Personality & Emotion Tests & Software: Psychological Books & Articles of Popular Interest (2009)
15. Ghayoumi, M., Bansal, A. K.; Real emotion recognition algorithm by detecting symmetry patterns with Dihedral group. In: MCSI (2016)
16. Schultz, W.: Neural coding of basic reward terms of animal learning theory, game theory microeconomics and behavioral ecology. Cur. Opin. Neurobiol. **14**(2), 139–147 (2004)
17. Panksepp, J.: Affective Neuroscience. Oxford University Press, New York (1998)
18. Laird, J.: The Soar Cognitive Architecture. MIT Press, Cambridge (2012)
19. Friesen, E., Ekman, P.: Facial action coding system: a technique for the measurement of facial movement, Palo Alto (1978)
20. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
21. Abrishami Moghaddam, H., Ghayoumi, M.: Facial image feature extraction using support vector machines. In: Proceeding VISAPP, Setubal, Portugal (2006)
22. Ghayoumi, M., Bansal, A.K.: An integrated approach for efficient analysis of facial expressions. In: SIGMAP, (2014)
23. Susskind, J.M., Hinton, G.E., Movellan, J.R., Anderson, A.K.: Generating facial expressions with deep belief nets. Affective Computing, Emotion Model. Synth. Recogn., 421-440 (2008)
24. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
25. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Computer Vision and Pattern Recognition (CVPR), pp. 1891–1898. IEEE (2014)
26. Ghayoumi, M., Bansal, A.: Unifying geometric features and facial action units for improved performance of facial expression analysis, CSSCC (2015)
27. Ghayoumi, M., Tafar, M., Bansal, A. K.: Towards formal multimodal analysis of emotions for affective computing. DMS (2016)
28. Huan, Y.: Wu, Ao., Zhang, G., Li, Y.: Extraction of adaptive wavelet packet filter-bank-based acoustic feature for emotion recognition. IET Signal Process. **9**(4), 341–348 (2015)
29. Kwon, O. W., Chan, K., Hao, J., Lee, T. W.: Emotion recognition by speech signals. In: 8th International Conference on Speech Communication and Technology (2003)

30. Lee, C.M., Narayanan, S.S.: Towards detecting emotions in spoken dialog. IEEE Trans. Speech Audio Process. **13**(2), 293–303 (2005)
31. Mitra, S., Acharya, T.: Gesture recognition: a survey. IEEE Trans. Syst. Man Cybern. **37**(3), 311–324 (2007)
32. Glowinski, D., Dael, N., Camurri, A., Volpe, G., Mortillaro, M., Scherer, K.: Toward a minimal representation of affective gestures. IEEE Trans. Affect. Comput. **2**(2), 106–118 (2011)
33. Camurri, A., Lagerlö, I., Volpe, G.: Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. Int. J. Hum. Comput. Stud. **59**(1), 213–225 (2003)
34. Castellano, G., Villalba, S.D., Camurri, A.: Recognising human emotions from body movement and gesture dynamics. In: Paiva, A.C., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 71–82. Springer, Heidelberg (2007)
35. Ghayoumi, M.: A Review of Multimodal Biometric Systems Fusion Methods and Its Applications. ICIS, USA (2015)
36. Ghayoumi, M., Khan, J., Pourebadi Khotbesara, M., Bauer, E., Hossain, A.: Follower Robot with an Optimized Gesture Recognition System. RSS, USA (2016)