

Deep Convolutional and Recurrent Neural Networks for Emotion Recognition from Human Behaviors

James J. Deng[1(✉)] and Clement H. C. Leung[2]

[1] MindSense Technologies, Hong Kong, People's Republic of China
james@mindsense.ai
[2] The Chinese University of Hong Kong, Shenzhen, People's Republic of China
clementleung@cuhk.edu.cn

Abstract. Human behaviors and the emotional states that they convey have been studied by psychologist and sociologists. The tracking of behaviors and emotions is becoming more pervasive with the advent of the Internet of Things (IoT), where small and always connected sensors can continuously capture information about human gestures, movements and postures. The captured information about readable behaviors conveys significant information that can be represented as time series. Few studies in emotion recognition and affective computing have explored the connection between the time series sensors data and the emotional behavior they conveys. In this paper, an innovative approach is proposed to study the emotions and behaviors connected to the time series data. A convolutional network augmented with attention-based bidirectional LSTM is introduced to represent the correlations between behaviors and emotions. The advantage of this model is that it can well recognized emotions by exploiting the data captured by sensors. The experimental results show that the proposed deep learning method outperforms separate schemes and achieves a high degree of accuracy for modelling human behaviors and emotions.

Keywords: Emotion recognition · Convolutional neutral network · Bi-directional LSTM · Attention model · Human behaviors

1 Introduction

Emotion and cognition are the advanced capabilities of artificial intelligence. The theories to model human emotion and cognition are at the base of affective computing [20] and emotion recognition. Many studies [11,13,17] on these areas, conducted by different researchers, brought to different approaches and models.

In psychology and sociology different studies [6,14] elicited the connection between human behaviors like gestures, movement and posture and the emotion they conveyed, for example the frequency of movements, the strength of

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12250, pp. 550–561, 2020.
https://doi.org/10.1007/978-3-030-58802-1_39

gestures. The challenge is how to represent the human behaviors and the connected emotions in an accurate and effective way. On the other hand, capturing the human behaviors has become more pervasive and efficient with the advent and development of the Internet of things (IoTs). In fact, smaller and smarter connected mobile devices and sensors, paired with cloud computing for big data storage and analysis, rendered feasible the near real-time behaviors detection and emotions recognition.

The analysis and modeling of human behaviors and emotions using deep learning techniques is motivated by the fact that these human activities are characterized by long and short term sequence features. Recurrent neural network (RNN) and Long Short Term Memory (LSTM) have been widely used in sequence modeling. Furthermore, bidirectional LSTM can use both past and future information with two separate hidden layers, which can represent different states and grades of human behaviors and emotions. An attention-based mechanism [1,3] is then used to focus on the most important information using a separate hidden layer. The deep architecture models can significantly outperform those with shallow architecture, and greatly improves training speed and effectiveness. The convolutioanl neural networks can well extract viiusal information. Therefore, a convolutional neural network augmented with Attention-based deep bi-directional Long Short Term Memory (CNN-ADBLSTM) network architecture is proposed to model human behaviors and emotions, which can well perform prediction tasks such as emotion recognition.

2 Literature Review

Different models have been proposed for emotion representations by different researchers. Usually based on two emotion theories: discrete emotion theory and dimensional emotion theory. Discrete emotion theory employs a finite number of emotional descriptors or adjectives [18] to express basic human emotions (e.g., joy, sadness, anger, contempt, happiness). Ortony et al. [19] proposed an emotion cognition model commonly known as OCC model to hierarchically describe 22 emotion descriptors. More coarser-grained partition (e.g. happy, neutral, and sad) as well as abstraction and similarity of emotions [2,12] have been also been used in various works. The discrete emotion theory main advantage is its ease to explain and to use in practical applications. Dimensional emotion theory states that emotion should be depicted in a psychological dimensional space, which can overcome the disadvantages of discrete theory such as the difficulty to represent continuous changes and to distinguish between similar emotion. Dimensional emotion models such as arousal-valence [22], resonance-arousal-valence [9], and arousal-valence-dominance are widely used in different application domains. Dimensional emotion theory is more likely to be used in computational emotion systems. We shall adopt the discrete emotion theory in this study.

Emotion recognition tasks often adopt human facial expressions, while some studies include body postures [4] and gestures [16,21]. Other studies focus on

movements where angry movements tend to be large, fast and relatively jerky, while fearful and sad movements are less energetic, smaller, and slower. Some studies use multiple resources, such as images or sensors, for human behavior recognition [5]. We shall focus on simple human behaviors data captured by accelerometer. The emotions and behaviors analysis and modeling can use traditional machine learning methods for time series analysis [8]. However, deep learning techniques have been successfully applied to the image, video, and audio, with many works carrying out emotion recognition, while fewer studies use sensors data [7,24]. We make some explorations on this area by using sequence models in deep neural network. In our previous work [10], an attention-based bidirectional LSTM model have been applied to recognize emotion from human behaviors. In this paper, we explore the convolutional network augmented attention-based LSTM to evaluate the performance.

3 Human Behavior Representation

Human beings are complex organisms that consist of various subsystems such as neural system, circulatory system, muscular and skeletal systems. All These subsystems work together to activate different physical and physiological changes. Human behaviors are the response of individuals internal and external stimuli, and usually is very strongly influenced by emotions. During the past decades, emotions have been well studied using facial expressions or voice, while there is less research on our bodies and behaviors that also convey rich emotional information or messages, such as a greater amount of muscle tension implying stress, a frequently walk up and down indicating anxiety, and a thumb upward representing an approval and admiring. Considering convolutional and recurrent network have already achieved success in image and time series signals, here we attempt to build time series expressions of behaviors such as gesture, posture and movement, and construct spatial-temporal network to model behaviors and emotions associated with human.

As IoTs and various sensors have been widely used in mobile and wearable devices, human behaviors like gesture, posture or movement can be represented by a set of time series signals obtained from specific sensors. Considering that there are many different sensors like Electrocardiography (ECG), Electromyography (EMG), gyroscope and accelerometers for specific purpose such as health, sports, gesture control and motion in data collection, to simplify, in this paper we only use the accelerometers for human behavior measurement. An accelerometer is a sensor that can measure proper acceleration. Here, we directly use three-dimensional accelerometer data to represent time series human behaviors. For notation, given a recognized human behavior B, for example hitting or shaking, corresponding to a sequence of three-dimensional accelerometer signals, behavior B can be represented by $B = \{x_1, x_2, \cdots, x_N\}$ where x_i denotes for the three-dimensional accelerometer data at timestamp i. Our goal is to predict emotion E conveyed by given behavior B. The following sections will describe the spatial-temporal neural network for modeling behavior and emotion.

4 Modeling Human Behaviors and Emotions

Human behaviors have temporal structure and associated with specific emotions, time sequence provides a natural and intuitive representation. The ultimate goal of this paper is to build a model that can recognize emotion conveyed by human behaviors that may be measured by of multiple temporal physical and physiological signals. As mentioned previously, to simplify, here we only use time series movement to represent human behaviors.

Long Short Term Memory (LSTM) are effective and have shown satisfied results in many sequential information processing tasks (e.g., natural language processing, speech recognition, and human activity recognition). Furthermore, Our previous work has demonstrated that the bidirectional LSTM assisted by attention mechanism achieved the state-of-the-art performance. However, human behaviors like posture or gesture also contains rich spatial information for decoding associated emotion. In order to take account both temporal and spatial information, we shall use the attention-based bidirectional LSTM and convolutional neural network to model human behaviors and emotions in this paper.

4.1 Deep Attention-Based Bidirectional LSTM

There are several methods to construct deep bidirectional LSTM structures. One method is to simply stack bidirectional LSTM in each layer, while another efficient method is that the bidirectional LSTM is located in the bottom layer and the other layers are unidirectional as shown in Fig. 1. The former method works only to a certain number of layers, beyond which the network becomes too slow and difficult to train, likely due to the gradient exploding and vanishing problems, while the latter can resolve these problems. Consequently, in this paper, we apply the latter method to construct a deep structure to modeling human behaviors and emotions in temporal as shown in Fig. 1. The attention model based on the top layer is defined as follows,

$$\alpha_t(s) = \frac{\exp(score(h_t, \bar{h}_{\hat{s}}))}{\sum_{s'} \exp(score(h_t, \bar{h}_{\hat{s}}))} \tag{1}$$

where $socre()$ is an activation function, $f()$ is tahn or relu function, with weight and bias W and b for the network.

$$score(h_t, \bar{h}) = f(h_t^T W \hat{h}_s + b) \tag{2}$$

4.2 Temporal Convolutional Network

The temporal convolutional network plays an important role in processing time series signals. The time series accelerometer signals is feeded to a convolutional network. The 3D filters are used in each convolutional layer, followed by batch normalization and activation steps, where activation function is the Rectified Linear Unit (ReLU). Global average pooling is applied to decrease the dimensionality following the final convolution block. An illustration of the structure of convolutional network is given in Fig. 2.

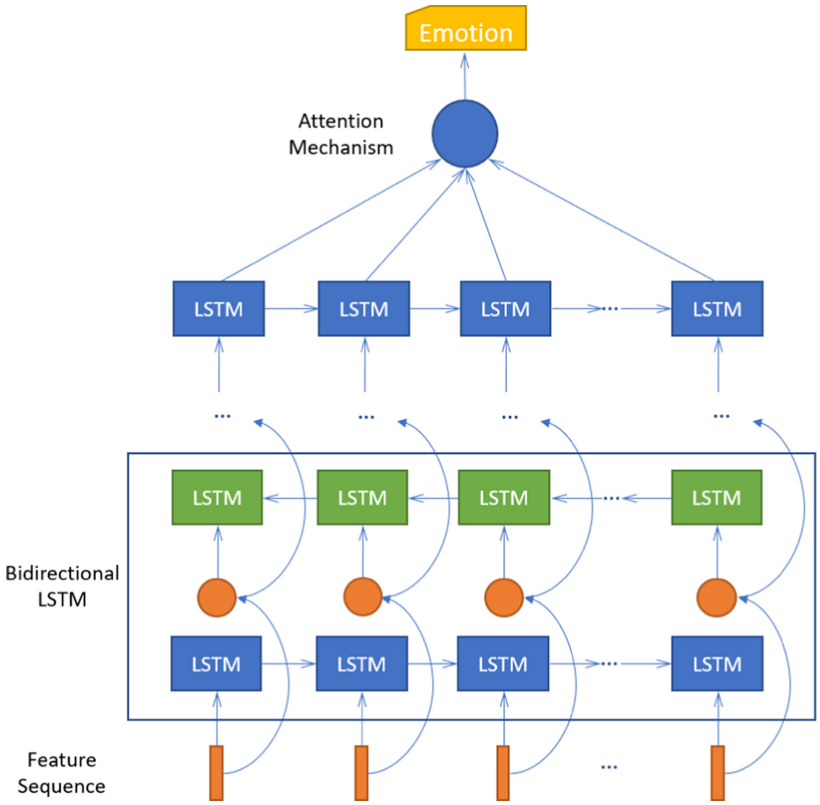

Fig. 1. Deep Attention-based bidirectional LSTM network. Blue rectangles represent forward LSTM units, and green rectangles represent backward LSTM units. Attention mechanism is in place.

4.3 Convolutional and Attention-Based Bidirectional LSTM Network

In order to take advantages of both convolutional and recurrent neural networks, we combine both convolutional and attention-based bidirectional LSTM to construct a hybrid architecture. The convolutional block and LSTM block perceive the same time series input in two different paths. The convolutional block takes 3D accelerometer signals with multiple time steps. Meanwhile, the LSTM block receives 3D accelerometer signals. Through bidirectional and residual connections, the upper LSTM outputs are finally input to attention mechanism.

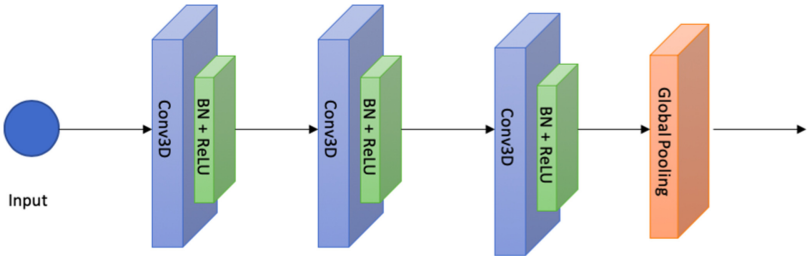

Fig. 2. Convolutional Neural Network. BN denotes for batch normalization.

The residual connections between adjacent layers are given in Eq. 3, and attention layer is calculated following previous Eq. 1.

$$\begin{aligned} h_t^i &= \Lambda(x_t^i, h_{t-1}^i; W^i) \\ x_t^{i+1} &= h_t^i + x_t^i \\ h_t^{i+1} &= \Lambda_{i+1}(x_t^{i+1}, h_{t-1}^{i+1}; W^{i+1}) \end{aligned} \tag{3}$$

The output of global pooling layer and attention layer is concatenated and passed onto a softmax function for emotion recognition. Figure 3 shows the whole network architecture of Convolutional and Attention-based Bidirectional LSTM.

5 Experiments

5.1 Experimental Setup

Ten people (7 males and 3 females) have participated in the experiments. We used wearable brands to collect 3-dimensional accelerometer data. The sampling rate of the accelerometer of given brand is set to 200 HZ. We set five predefined behaviors or movements corresponding to emotions as shown in Table 1. Each participant carried out given behaviors of movement and emotions. Each specific behavior and corresponding emotion is performed by multiple participants. The total amount of dataset we collected is more than 100,000 s. After collecting the original dataset, we normalized each dimensional with zero-mean and unit-variance. Furthermore, we can set more features such as accelerometer shape or contrasts as model input in experiments. Shape features contain curve centroid, flux, flatness and roll-off, and contrast features contain kurtosis, skewness, zero-crossing rate.

5.2 Training and Evaluation Criteria

After preprocessing the dataset, training deep ALSTMmodel with suitable parameters is vitally important for model performance. We divided the dataset into three parts, where 60% of it were used for training, 20% of it were used for

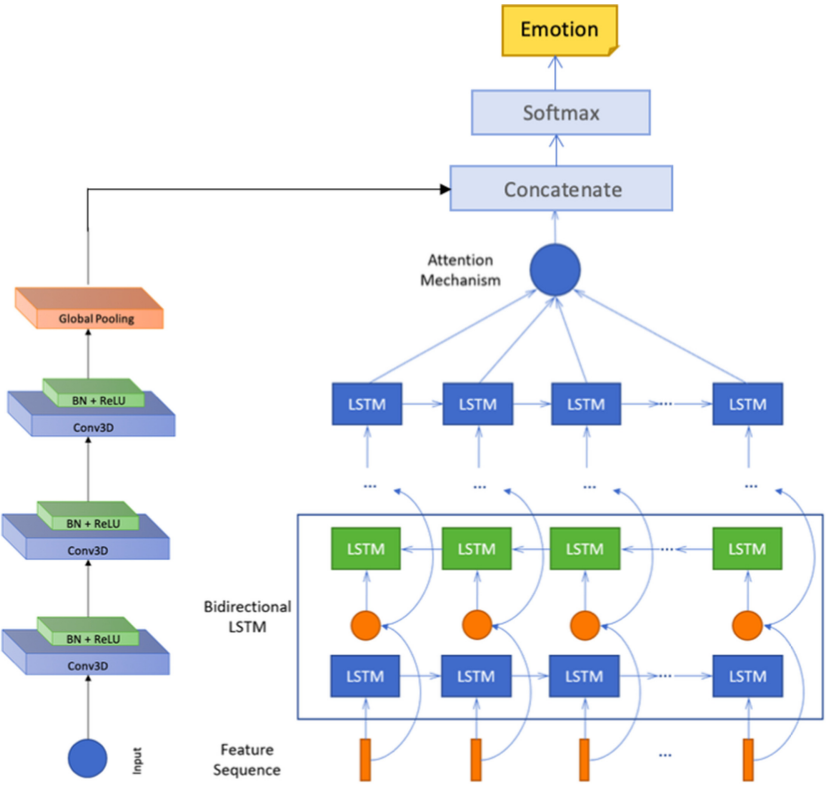

Fig. 3. Hybrid network architecture convolutional neural network augmented with attention-based dep bidirectional LSTM. Blue rectangles represent forward LSTM units, and green rectangles represent backward LSTM units. Attention mechanism is in place. BN denotes for batch normalization. The input and feature space are accelerometer signals.

validation, and the remaining 20% were used for testing. We set the mini-batches with size 128 for training and use Adam optimizer. We attempted different layers with different numbers of hidden states. The length of the training sequence is initially set at 24, which can be larger when more layered are set. As the input sequence dimension is not large, the maximum depth of layers is set as 4 to avoid model overfitting. After the deep convolutional augmented ABLSTM model has been trained, we can apply it to predict the emotion of testing accelerometer data. The emotion recognition is evaluated based on accuracy. The comparison of this hybrid network with separate network is carried out to evaluate its performance.

Table 1. Behavior types and corresponding emotions in data collection

Behaviors ID	Behaviors	Emotions
1	Walk fast and relatively jerky	Angry
2	Walk slowly and smaller	Sad
3	Short movement	Anxiety
4	Up and down	Exciting
5	Arms stretched out to the front	Joy

5.3 Experimental Results

We have trained different deep models using 3D accelerometer data, shape and contrast features. Table 2 shows the comparison of several different layers of Deep LSTM (DLSTM) models and the corresponding bidirectional model (DBLSTM) with and without attention mechanism. Attention-based LSTM (ALSTM) and its bidirectional variant (ABLSTM) models perform better than those without attention mechanism. We can see that the accuracy of ABLSTM (accuracy = 96.1%) is higher than that of DBLSTM (accuracy = 95.2%). This indicates that time series human's behavior data can be well decoded for emotion recognition, that's because different segments of human's time series data can expressing different weights for emotion decoding. Furthermore, as proved in our previous work, deeper models like ADBLSTM-3L, ADBLSTM-4L also exhibit the better performance than ALSTM and ABLSTM. However, it does not mean that the deeper the better for all in practice. Some works [23] have shown that the maximum number of layers is 8 in neural machine translation. Here, we find that when the layers are set to 4, the best performance is achieved, and the average accuracy in validation and testing can be obtained at 97.3%. As the attention is added to the top layer, there are more computation required on model training. However, we find that the loss of mini-batch of training dataset decrease faster yet with higher accuracy than those without attention operation at the same training iterations.

We also evaluated accuracy for five emotion categories using the ADBLSTM-4L, and the results are given in Fig. 4. We can see that the emotion "exciting" shows the highest accuracy, that is because the people's behaviors go up and down repeatedly, which is apparently different from other walking behaviors. Emotion "anxiety" has the lowest accuracy, which means that short movements do not always indicate anxious feelings. In addition, we also divide these five emotions into two large categories of positive and negative. Figure 5 shows the emotion recognition performance in this two coarse categories using the same model. Positive emotions (accuracy = 96.8%) outperform negative emotions (accuracy = 92.6%), suggesting that the two positive emotions such as joy and excitement are easier to recognize from behaviors, while the three negative emotions such as sad, anxiety, and angry are more difficult to recognize. The above results should also depend on the cultural milieu of the person and groups.

Table 2. Comparison of different deep recurrent neural network models

Models	Validation	Testing
LSTM	94.6%	94.9%
BLSTM	95.2%	95.7%
DBLSTM-3L	95.4%	97.0%
DBLSTM-4L	96.7%	97.1%
ALSTM	95.7%	95.2%
ABLSTM	96.1%	96.7%
ADBLSTM-3L	96.8%	97.6%
ADBLSTM-4L	97.2%	97.4%

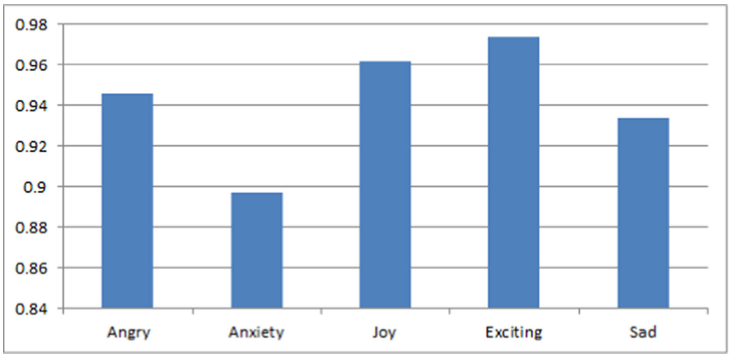

Fig. 4. Comparison of five emotion categories recognition performance

To evaluate the hybrid network performance, we compare it with separate network. Three stacked temporal convolutional blocks was used in the experiment. The convolution kernel was initialized following the initialization proposed by [15]. The learning rate was reduced by a specific factor of every 100 epochs of no improvement in the validation score, until the final learning rate was reached. The comparison results are given in Table 3. We can see that the convolutional network contributes the time series data, and convolutional model augmented with ADBLSTM outperforms the existing state-of-the-art models for emotion recognition. The model CNN-ADBLSTM-4L achived the highest accuracy in validation and testing dataset.

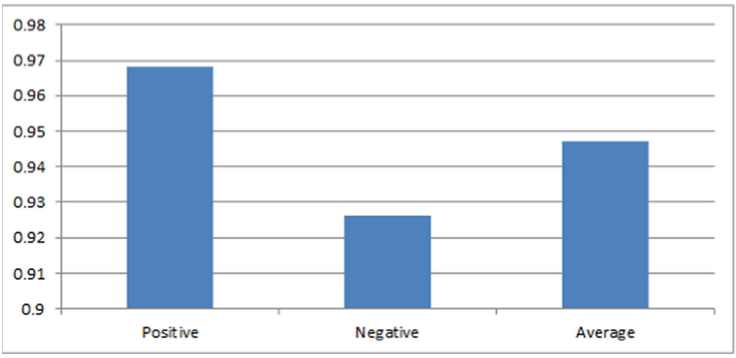

Fig. 5. Two coarse emotion categories recognition performance

Table 3. Comparison of convolutional augmented with attention-based bidirectional LSTM with separate network

Models	Validation	Testing
CNN	95.9%	94.7%
ABLSTM	96.1%	96.7%
CNN-ABLSTM	96.1%	96.4%
ADBLSTM-3L	96.8%	97.6%
CNN-ADBLSTM-3L	97.3%	97.4%
ADBLSTM-4L	97.2%	97.4%
CNN-ADBLSTM-4L	97.3%	97.6%

6 Conclusion

In this paper we introduced an innovative deep learning model for modeling human behaviors and emotions. Using data from IoT devices and recent advances in technology, human posture, gesture, and movement behaviors can be captured by sensors, and be analyzed as well as modeled by deep neural networks. Considering the interaction and correlation of human behaviors and emotions, we introduced a methodology that makes use of the deep convolutional neural network and the attention-based bidirectional LSTM networks to build hybrid network architecture. The bidirectional LSTM is deployed in the bottom layer and an attention-based mechanism is added to focus on important information. This sophisticated design is able to facilitate deep model training. The convoluational blocks can make usage of visual information. Both convoluational and recurrent network outputs are concatenate and used for emotion recognition from time series human behaviors. The experimental results show that the proposed method is able to obtain good emotion recognition performance. Furthermore, this method should scale well for modeling various behaviors and emotions through multiple sensors.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473 (2014)
2. Biondi, G., Franzoni, V., Li, Y., Milani, A.: Web-based similarity for emotion recognition in web objects. In: Proceedings of the 9th International Conference on Utility and Cloud Computing, UCC 2016, New York, NY, USA, pp. 327–332. ACM (2016)
3. Cho, K., Courville, A.C., Bengio, Y.: Describing multimedia content using attention-based encoder-decoder networks. IEEE Trans. Multimedia **17**(11), 1875–1886 (2015)
4. Coulson, M.: Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. J. Nonverbal Behav. **28**(2), 117–139 (2004). https://doi.org/10.1023/B:JONB.0000023655.25550.be
5. Crane, E., Gross, M.: Motion capture and emotion: affect detection in whole body movement. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 95–101. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74889-2_9
6. De Gelder, B.: Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. Philos. Trans. Roy. Soc. B Biol. Sci. **364**(1535), 3475–3484 (2009)
7. Deng, J., Leung, C., Li, Y.: Beyond big data of human behaviors: Modeling human behaviors and deep emotions. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 282–286, April 2018
8. Deng, J.J., Leung, C.H.: Dynamic time warping for music retrieval using time series modeling of musical emotions. IEEE Trans. Affect. Comput. **6**(2), 137–151 (2015)
9. Deng, J.J., Leung, C.H., Milani, A., Chen, L.: Emotional states associated with music: classification, prediction of changes, and consideration in recommendation. ACM Trans. Interact. Intell. Syst. (TiiS) **5**(1), 4 (2015)
10. Deng, J.J., Leung, C.H.C., Mengoni, P., Li, Y.: Emotion recognition from human behaviors using attention model. In: 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 249–253. IEEE (2018)
11. Dolan, R.J.: Emotion, cognition, and behavior. Science **298**(5596), 1191–1194 (2002)
12. Franzoni, V., Li, Y., Mengoni, P.: A path-based model for emotion abstraction on facebook using sentiment analysis and taxonomy knowledge. In: Proceedings of the International Conference on Web Intelligence - WI 2017, pp. 947–952 (2017)
13. Gratch, J., Marsella, S.: A domain-independent framework for modeling emotion. Cogn. Syst. Res. **5**(4), 269–306 (2004)
14. Harrigan, J.A.: Proxemics, kinesics, and gaze. In: Harrigan, J., Rosenthal, R., Scherer, K.R., Scherer, K. (eds.) The New Handbook of Methods in Nonverbal Behavior Research, pp. 137–198. Oxford University Press, Oxford (2005)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
16. Hicheur, H., Kadone, H., Grezes, J., Berthoz, A.: The combined role of motion-related cues and upper body posture for the expression of emotions during human walking. In: Mombau, K., Berns, K. (eds). Modeling, Simulation and Optimization of Bipedal Walking. Cognitive Systems Monographs, vol 18. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36368-9_6

17. Hudlicka, E.: Beyond cognition: Modeling emotion in cognitive architectures. In: ICCM, pp. 118–123 (2004)
18. zard, C.E., Malatesta, C.Z.: Perspectives on emotional development i: differential emotions theory of early emotional development. In: The first draft of this paper was based on an invited address to the Eastern Psychological Association, 1 April 1983. Wiley, Hoboken (1987)
19. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1990)
20. Picard, R.W., et al.: Affective Computing. MIT Press, Cambridge (1995)
21. Roether, C.L., Omlor, L., Christensen, A., Giese, M.A.: Critical features for the perception of emotion from gait. J. Vis. **9**(6), 15–15 (2009)
22. Thayer, R.E.: The Biopsychology of Mood and Arousal. Oxford University Press, Oxford (1990)
23. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
24. Zhang, Z., Song, Y., Cui, L., Liu, X., Zhu, T.: Emotion recognition based on customized smart bracelet with built-in accelerometer. PeerJ **4**, e2258 (2016)