2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)

Emotion Recognition from Human Behaviors using Attention Model

James J. Deng
Hong Kong Applied Science and Technology Research Institute
Hong Kong, China
jamesjdeng@gmail.com

Clement H. C. Leung
Centre for Applied Informatics Victoria University
Melbourne, Victoria, Australia
clement.leung@vu.edu.au

Paolo Mengoni
Dept. of Mathematics and Computer Science University of Florence
Florence, Italy
paolo.mengoni@unifi.it

Yuanxi Li
Dept. of Computer Science Hong Kong Baptist University
Hong Kong, China
csyxli@comp.hkbu.edu.hk

***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 signicant 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 sensors time series data. A deep learning attention-based bidirectional LSTM is introduced to represent the correlations between data and emotions. The attention-based mechanism is then used to focus on the most significant information. The advantage of this model is that it can represent various human emotions by exploiting the data captured by sensors. The experimental results show that the proposed deep learning method outperforms shallow models and achieves a high degree of accuracy for modelling human behaviors and emotions.**

***Index Terms*—emotion recognition, deep learning, attention model, LSTM, bi-LSTM**

I. Introduction

The theories to model human emotion and cognition are at the base of affective computing[19] and emotion recognition. Multiple studies[10], [12], [15], conducted by different authors, brought to various methods and different approaches. In psychology and sociology different studies [13], [6] elicited the connection between gestures, movement and postural behaviors and the emotion they convey e.g. the frequency of movements, the strength of gestures, etc. The challenge is now 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 (IoT). In fact, smaller and ever 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. Therefore, a deep Attention-based bi-directional Long Short Term Memory (ABLSTM) network architecture is proposed to model human behaviors and emotions, which can well perform prediction tasks such as emotion recognition.

II. 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 [16] to express basic human emotions (e.g., joy, sadness, anger, contempt, happiness). Ortony et al. [18] 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], [11] 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

978-1-5386-9555-5/18/$31.00 ©2018 IEEE
DOI 10.1109/AIKE.2018.00056

gestures [20], [14]. 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 mltiple 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 exploreations on this area by using sequence models in deep neural network.

III. Human Behavior Representation

In the human organism different individual systems such as neural system, circulatory system, muscular system and so on, work together to activate various human physical and physiological changes. Behaviors and emotions are formed and connected to various subsystems. Emotions have been well studied using facial expressions, while there is less research on our bodies and behaviors that 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. Therefore, here we want to build expressions of behaviors such as gesture, posture, and movement that model humans emotion well.

As IoTs and various sensors have been widely adopted to measure many aspects of human behaviors (e.g., running, shaking, and hitting) in mobile devices, we also apply sensors to collect data generated by human behaviors. Considering that many different types of sensors such as Electrocardiography (ECG), touch sensors, and accelerometers, we shall focus on accelerometer sensors in this paper. An accelerometer is a sensor that can measure proper acceleration which is suitable for movement measurement. Therefore, we directly use three-dimensional accelerometer data to represent human behaviors.

IV. Modeling Human Behavior and Emotion

As previously mentioned, human behaviors have temporal structure and relationship with emotions, time sequence analysis provides a powerful way to learn and model behaviors and emotions. The ultimate goal of this paper is to recognize emotion conveyed by human behaviors measured by of multiple temporal physical and physiological change in time sequence. Here, we consider human movement behaviors. Considering that Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) are effective and have shown considerable success in many sequential information processing tasks (e.g., natural language processing and speech recognition) [21], [17], we also apply RNN-LSTM to model time series human movement of behaviors. As the characteristic of interaction between behaviors and emotions, and attention mechanism have shown success in image, audio and NLP domains, we shall use bidirectional LSTM to model them assisted by attention-based mechanism.

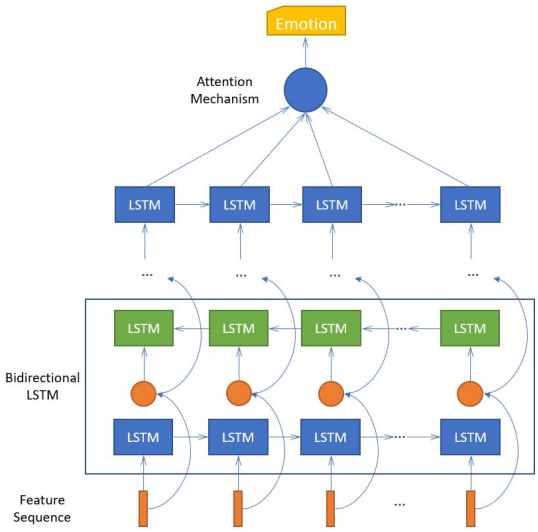

Fig. 1. Deep Attention-based bidirectional LSTM network. Blue rectangles represent forward LSTM units, and green rectangles represent backward LSTM units. Attention mechanism is in place.

A. Attention-based bidirectional LSTM for Modeling Behaviors and Emotions

We use LSTM to modeling human behavior and emotion. The attention-based LSTM (ALSTM) model use a mechanism to choose the relevant information from the input vectors encoded in the hidden layers. Using the attention weights computed using this mechanism, ALSTM can recognize and model the emotional conveying information. As this information is usually unevenly distributed through the events time sequence, the attention mechanism help the algorithm to differentiate the various sequences of events and focus on the most important ones. Conventional LSTM makes use only of the past sequence information. They cannot consider the future sequences as the information flow is unidirectional. On the other hand, the bidirectional LSTM can use both past and future sequence information using two separate hidden layers, which can completely represent human behaviors and emotions interactions and correlations.

B. Deep Attention-based bidirectional LSTM

Considering that many research works have shown that deep architecture models can outperform shallow architecture, we also apply deep network architecture for Attention-based bidirectional LSTM. There are several deep structures. One method is to simply stack ALSTM 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 Figure 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

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

TABLE I
BEHAVIOR TYPES AND CORRESPONDING EMOTIONS IN DATA COLLECTION

problems. Consequently, in this paper, we apply the latter method to construct a deep structure for modeling human behaviors and emotions. The attention model based on the top layer is defined as follows,

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

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

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

V. EXPERIMENTS

A. 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 200HZ. We set five predefined behaviors or movements corresponding to emotions as shown in Table I. 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 seconds. Figure 2 shows an example of three-dimensional distribution of accelerometer data. 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.

B. Training and Evaluation Criteria

After preprocessing the dataset, training deep ALSTM model 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 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 ALSTM model has been trained, we can apply

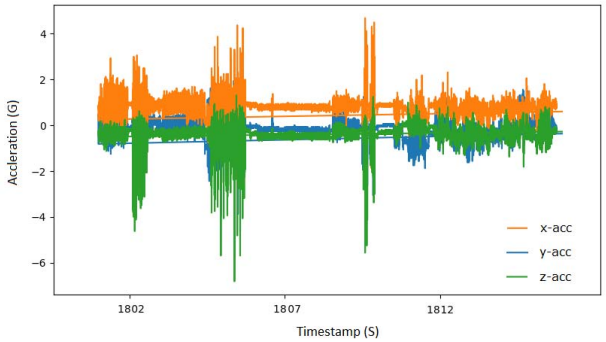

Fig. 2. An example of three-dimensional accelerometer data

it to predict the emotion of new accelerometer data. The emotion recognition is evaluated based on accuracy.

C. Experimental Results

We have trained different deep models using 3D accelerometer data, shape and contrast features. Table II 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 nd 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 nd 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 Figure 3. 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 4 shows the emotion recognition performance in this

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%

TABLE II
COMPARISON OF DIFFERENT DEEP MODELS

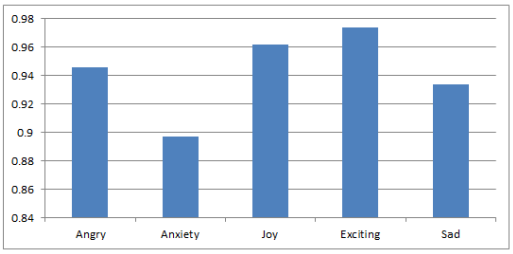

Fig. 3. Comparison of five emotion categories recognition performance

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.

VI. 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 an attention-based bidirectional LSTM networks 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, and the experimental results show that the proposed method is able to obtain good emotion recognition performance, enriched by the attention mechanism. Furthermore, this method should scale well for modeling various behaviors and emotions.

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

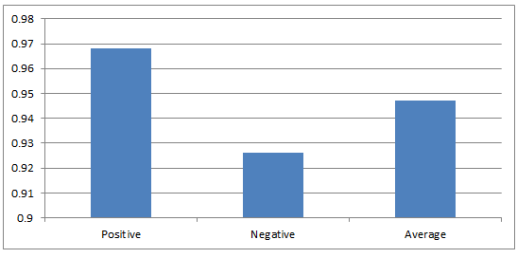

Fig. 4. Two coarse emotion categories recognition performance

[2] G. Biondi, V. Franzoni, Y. Li, and A. Milani. Web-based similarity for emotion recognition in web objects. In *Proceedings of the 9th International Conference on Utility and Cloud Computing*, UCC '16, pages 327–332, New York, NY, USA, 2016. ACM.

[3] K. Cho, A. C. Courville, and Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Trans. Multimedia*, 17(11):1875–1886, 2015.

[4] M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, 28(2):117–139, 2004.

[5] E. Crane and M. Gross. Motion capture and emotion: Affect detection in whole body movement. In *International Conference on Affective Computing and Intelligent Interaction*, pages 95–101. Springer, 2007.

[6] B. De Gelder. Why bodies? twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3475–3484, 2009.

[7] J. Deng, C. Leung, and Y. Li. Beyond big data of human behaviors: Modeling human behaviors and deep emotions. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 282–286, April 2018.

[8] J. J. Deng and C. H. Leung. Dynamic time warping for music retrieval using time series modeling of musical emotions. *IEEE Transactions on Affective Computing*, 6(2):137–151, 2015.

[9] J. J. Deng, C. H. Leung, A. Milani, and L. Chen. Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(1):4, 2015.

[10] R. J. Dolan. Emotion, cognition, and behavior. *science*, 298(5596):1191–1194, 2002.

[11] V. Franzoni, Y. Li, and P. Mengoni. A path-based model for emotion abstraction on facebook using sentiment analysis and taxonomy knowledge. *Proceedings of the International Conference on Web Intelligence - WI '17*, pages 947–952, 2017.

[12] J. Gratch and S. Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.

[13] J. A. Harrigan. Proxemics, kinesics, and gaze. *The new handbook of methods in nonverbal behavior research*, pages 137–198, 2005.

[14] H. Hicheur, H. Kadone, J. Grezes, and A. Berthoz. The combined role of motion-related cues and upper body posture for the expression of emotions during human walking. In *Modeling, simulation and optimization of bipedal walking*, pages 71–85. Springer, 2013.

[15] E. Hudlicka. Beyond cognition: Modeling emotion in cognitive architectures. In *ICCM*, pages 118–123, 2004.

[16] C. E. Izard and C. Z. Malatesta. 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, Apr 1, 1983*. John Wiley & Sons, 1987.

[17] Y. Miao, M. Gowayyed, and F. Metze. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 167–174. IEEE, 2015.

[18] A. Ortony, G. L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge university press, 1990.

[19] R. W. Picard et al. Affective computing. 1995.

[20] C. L. Roether, L. Omlor, A. Christensen, and M. A. Giese. Critical features for the perception of emotion from gait. *Journal of vision*, 9(6):15–15, 2009.

[21] H. Sak, A. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*, 2014.

[22] R. E. Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1990.

[23] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[24] Z. Zhang, Y. Song, L. Cui, X. Liu, and T. Zhu. Emotion recognition based on customized smart bracelet with built-in accelerometer. *PeerJ*, 4:e2258, 2016.