A Literature Review on Emotion Recognition using Various Methods

Omar Sharif

Received: 9 December 2016 Accepted: 5 January 2017 Published: 15 January 2017

Abstract

Emotion Recognition is an important area of work to improve the interaction between human and machine. Complexity of emotion makes the acquisition task more difficult. Quondam works are proposed to capture emotion through unimodal mechanism such as only facial expressions or only vocal input. More recently, inception to the idea of multimodal emotion recognition has increased the accuracy rate of the detection of the machine. Moreover, deep learning technique with neural network extended the success ratio of machine in respect of emotion recognition. Recent works with deep learning technique has been performed with different kinds of input of human behavior such as audio-visual inputs, facial expressions, body gestures, EEG signal and related brainwaves. Still many aspects in this area to work on to improve and make a robust system will detect and classify emotions more accurately. In this paper, we tried to explore the relevant significant works, their techniques, and the effectiveness of the methods and the scope of the improvement of the results.

Index terms—

A Literature Review on Emotion Recognition using Various Methods Reeshad Khan ? & Omar Sharif ?

1 I. Introduction

ost common exposition of an idea of emotion could be found as "a natural instinctive state of mind deriving from one's circumstances, mood, or relationships with others". Which misses depicting the driving force behind all motivation which may positive, negative or neutral. This is very important information to understand emotion as an intelligent agent. It is very complicated to detect the emotions and distinguish among them. Before a decades or two emotion started to become a concern as an important addition towards the modern technology world. Rises the hope of new dawn for intelligence apparatus. Imagine a world where machines do feel what humans need or want. With the special kind of calculation then that machine could predict the further consequences and by which mankind could avoid serious circumstances and lot more. Humans are far more strong and intelligent due to the addition of the emotion but less effective than machines. But what if machines get this special features of human? It will be the strongest addition to the technology ever. And to make the dreams come true this is the first step; train a system to spot and recognize emotions. This is the start of an intelligent system. Intelligent Systems are becoming more efficient by predicting and classifying decision in various aspects of practical life. Particularly, emotion recognition through deep learning has become intriguing research area for its innovative nature and practical implication. This technique mainly consists of detecting emotion through various kinds of input taken from different human behavior and condition. A technology namely neural network detects emotion through deep learning. For its complication mentioned earlier, an emotion recognition system with stellar efficiency and accuracy is needed.

2 II. Recent Related Work in the Relevant Field

Previous works are focused on eliciting results from unimodal systems. Machines used to predict emotion by only facial expressions [1] or only vocal sounds [2]. After a while, multimodal systems that use more than one features to predict emotion has more effective and gives more accurate results. So that, the combination of features such as audio-visual expressions, EEG, body gestures have been used since. More than one intelligent machine and neural networks are used to implement the emotion recognition system. Multimodal recognition method has proven more effective than unimodal systems by Shiqing et al. [3]. Research has demonstrated that deep neural networks can effectively generate discriminative features that approximate the complex non-linear dependencies between features in the original set. These deep generative models have been applied to speech and language processing, as well as emotion recognition tasks [4][5][6]. Martin et al. [7] showed that bidirectional Long Short Term Memory(BLSTM) network is more effective that conventional SVM approach.; In speech processing, Ngiam et al. [8] proposed and evaluated deep networks to learn audio-visual features from spoken letters. In emotion recognition, Brueckner et al. [9] found that the use of a Restricted Boltzmann Machine (RBM) prior to a two-layer neural network with fine-tuning could significantly improve classification accuracy in the Interspeech automatic likability classification challenge [10]. The work by Stuhlsatz et al. [11] took a different approach for learning acoustic features in speech emotion recognition using Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs). Yelin et al. [12] showed three layered Deep Belief Networks(DBNs') give better performance than two layered DBNs' by using audio-19 Year 2017 ( ) F

Abstract-Emotion Recognition is an important area of work to improve the interaction between human and machine. Complexity of emotion makes the acquisition task more difficult. Quondam works are proposed to capture emotion through unimodal mechanism such as only facial expressions or only vocal input. More recently, inception to the idea of multimodal emotion recognition has increased the accuracy rate of the detection of the machine. Moreover, deep learning technique with neural network extended the success ratio of machine in respect of emotion recognition. Recent works with deep learning technique has been performed with different kinds of input of human behavior such as audio-visual inputs, facial expressions, body gestures, EEG signal and related brainwaves. Still many aspects in this area to work on to improve and make a robust system will detect and classify emotions more accurately. In this paper, we tried to explore the relevant significant works, their techniques, and the effectiveness of the methods and the scope of the improvement of the results.

visual emotion recognition process. Samira et al [13] used Recurrent neural network combined with Convoluted Neural Network(CNN) in an underlying CNN-RNN architecture to predict emotion in the video. Some noble methods and techniques also enriched this particular research. They are more accurate, stable and realistic. In terms of performance, accuracy, reasonability and precision these methods are the dominating solutions. Some of them are more accurate but some are more realistic. Some take much time and require greater computation power to produce the more accurate result but some compromises accuracy over performance. The idea of being successful might differ but these solutions are the best possible till now.

Yelin Kim and Emily Mower Provos explore whether a subset of an utterance can be used for emotion inference and how the subset varies by classes of emotion and modalities. They propose a windowing method that identifies window configurations, window duration, and timing, for aggregating segment-level information for utterance-level emotion inference. The experimental results using the IEMOCAP and MSP-IMPROV datasets show that the identified temporal window configurations demonstrate consistent patterns across speakers, specific to different classes of emotion and modalities. They compare their proposed windowing method to a baseline method that randomly selects window configurations and a traditional all-mean method that uses the full information within an utterance. This method shows a significantly higher performance in emotion recognition while the method only uses 40-80% of information within each utterance. The identified windows also show consistency across speakers, demonstrating how multimodal cues reveal emotion over time. These patterns also align with psychological findings. But after all achievement, the result is not consistent with this method [15].

A. Yao, D. Cai, P. Hu, S. Wang, L. Shan, and Y. Chen used a well-designed Convolutional Neural Network (CNN) architecture regarding the video based emotion recognition [14]. They proposed the method named as HOLONET has three critical considerations in network design. (1) To reduce redundant filters and enhance the non-saturated non-linearity in the lower convolutional layers, they used modified Concatenated Rectified Linear Unit (CReLU) instead of ReLU. (2) To enjoy the accuracy gain from considerably increased network depth and maintain efficiency, they combine residual structure and CReLU to construct the middle layers. (3) To broaden network width and introduce multi-scale feature extraction property, the topper layers are designed as a variant of the inception-residual structure. This method more realistic than other methods here. It's focused on adaptability in real-time scenario than accuracy and theoretical performance. Though its accuracy is also impressive but only this method is applicable only in the video based emotion recognition. Other types of data rather than video, this method can't produce results [14].

Y. Fan, X. Lu, D. Li, and Y. Liu. proposed a method for video-based emotion recognition in the wild. They used CNN-LSTM and C3D networks to simultaneously model video appearances and motions [16]. They found that the combination of the two kinds of networks can give impressive results, which demonstrated the effectiveness of the method. In their proposed method they used LSTM (Long Short Term Memory) -a special kind of RNN, C3D -A Direct Spatio-Temporal Model and Hybrid CNN-RNN and C3D Networks. This method gives a great accuracy and performance is remarkable. But this method is much convoluted, time-consuming and less realistic. For this reason, efficiency is not that impressive [16].

Zixing Zhang, Fabien Ringeval, Eduardo Coutinho, Erik Marchi and Björn Schüller proposed some improvement in SSL technique to improve the low performance of a classifier that can deliver on challenging recognition tasks reduces the trust ability of the automatically labeled data and gave solutions regarding the noise accumulation problem -instances that are misclassified by the system are still used to train it in future iterations [17]. they exploited the complementarity between audio-visual features to improve the performance of the classifier during the supervised phase. Then, they iteratively re-evaluated the automatically labeled instances to correct possibly mislabeled data and this enhances the overall confidence of the system's predictions. This technique gives a best possible performance using SSL technique where labeled data is scarce and/or expensive to obtain but still, there are various inherent limitations that limit its performance in practical applications. This technique has been tested on a specific database with a limited type and number of data. The algorithm which has been used is not capable of processing physiological data alongside other types of data [17].

Wei-Long Zheng and Bao-Liang Lu proposed EEG-based effective models without labeled target data using transfer learning techniques (TCA-based Subject Transfer) [18] which is very accurate in terms of positive emotion recognition than other techniques used before. Their method achieved 85.01% accuracy. They used to transfer learning and their method includes three pillars, TCA-based Subject Transfer, KPCA-based Subject Transfer and Transductive Parameter Transfer. For data preprocessing they used raw EEG signals processed with a bandpass filter between 1 Hz and 75 Hz and for feature extraction, they employed differential entropy (DE) features. For evaluation, they adopted a leave-onesubject-out cross-validation method. Their experimental results demonstrated that the transductive parameter transfer approach significantly outperforms the other approaches in terms of the accuracies, and a 19.58% increase in recognition accuracy has been achieved.

Though this achievement is limited to the positive emotion recognition only. This method is limited in terms of negative and neutral emotion recognition. Yet a lot improvement needed to recognize negative and neutral emotion more accurately [18].

3 Proposed Method

In terms of emotion recognition, there is no indefinite way or method which is the univocal solution. A lot of solution have come and many to comes in near future with significant improvement in terms of efficiency, accuracy, and usability. In past and the current research shows that multimodalities dominated the area of emotion recognition than unimodality. Using EEG and audio-visual signal yields the best possible results according to the newest researches. We assume LSTM-RNN is the best way to handle multimodalities. So our proposal is focused on emotion recognition by EEG and audio-visual signal using LSTM-RNN. This type of research has been done before. But our challenge is to improve the model where it will be trained by EEG and audiovisual data at the same time and will make a relation between this data wherein, if one type of data is not available in a situation, the model could still produce the result; finding the relation within the data. So, the training will have two part; training for the data and training to understand the relations between the data.

4 V. Conclusions

In this Paper we discussed about the work done on emotion recognition and for achieving that all

5 Future Work Scope

We are working towards a machine with emotions. A machine or a system, which can think like humans, can feel warmness of heart; can judge on events, prioritized between choices and with many more emotional epithets. To make the dream reality first we need the machine or system to understand human emotions, ape the emotion and master it. We just started to do that. Though there is some real example exists this days. Some features and services are getting popularity like Microsoft Cognitive Services but still there is a lot works required in the terms of efficiency, accuracy and usability. Therefore, in future Emotion Recognition is an area requires a great intentness.

6 III.

superior and novel approaches and methods. We have proposed a glimpse of a probable solution and method towards recognition the emotion. Work so far substantiate that emotion recognition using users EEG signal and audiovisual signal has the highest recognition rate and has highest performance. [1]

6 III.

1

A Literature Review on Emotion Recognition using Various Methods

Reference and year	Approach and Method	Performance	
Wei-Long Zheng and Bao-Liang Lu (2016)	EEG-based affective models without labeled target data using transfer learning techniques (TCA-based Subject Transfer)	Positive (85.01%) emotion recognition rate is higher than other approaches but neutral (25.76%) and negative (10.24%) emotions are often confused with each other.	Year 2017
Zixing Zhang, Fabien Ringeval, Fabien Ringeval, Eduardo Coutinho, Erik Marchi and Björn Schüller (2016)	Semi-Supervised Learning (SSL) technique	Delivers a strong performance in the classification of high/low emotional arousal (UAR = 76.5%), and significantly outperforms traditional SSL methods by at least 5.0% (absolute gain).	
Y. Fan, X. Lu, D. Li, and Y. Liu. (2016)	Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks	Achieved accuracy 59.02% (without using any additional Emotion labeled video clips in training set) which is the best till now.	
A. Yao, D. Cai, P. Hu, S. Wang, L. Shan and Y. Chen (2016)	HoloNet: towards robust emotion recognition in the wild	Achieved mean recognition rate of 57.84%.	( ) F
Yelin Kim and Emily Mower Provos (2016)	Data driven framework to explore patterns (timings and durations) of emotion evidence, specific to individual emotion classes	Achieved 65.60% UW accuracy, 1.90% higher than the baseline.	
		IV.	

Figure 1: Table 1 :

[Schuller et al.] , B Schuller , S Steidl , A Batliner , E N¨oth , A Vinciarelli , F Burkhardt , R Van Son , F Weninger , F .

[Ebrahimi et al. ()] *Christopher Pal-Recurrent Neural Networks for Emotion Recognition in Video?*, Samira Ebrahimi , Vincent Michalski , Kishore Konda , Goethe Roland Memisevic . 2015. Montréal, Canada. Kahou École Polytechnique de Montréal, Canada ; Universität Frankfurt, Germany; Université de Montréal

[Kim and Provos ()] *Data driven framework to explore patterns (timings and durations) of emotion evidence, specific to individual emotion classes*, Yelin Kim , Emily Mower Provos . 2016. Ann Arbor, Michigan, USA. University of Michigan Electrical Engineering and Computer Science

[Morgan ()] 'Deep and wide: Multiple layers in automatic speech recognition,? Audio, Speech, and Language Processing'. N Morgan . *IEEE Transactions on* 2012. 20 (1) p. .

[Deep learning for robust feature generation inaudiovisual emotion recognition Yelin Kim ()] *Deep learning for robust feature generation inaudiovisual emotion recognition Yelin Kim*, 2013. Ann Arbor, Michigan, US. Honglak Lee, and Emily Mower Provost * University of Michigan Electrical Engineering and Computer Science

[Levi and Hassner ()] *Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns*, Gil Levi , Tal Hassner . 2014. SC / Information Sciences Institute, the Open University of Israel

[Eyben and Bocklet ()] T Eyben , Bocklet . *The interspeech 2012 speaker trait challenge,? Interspeech*, (Portland, Oregon) 2012.

[Brueckner and Schuller ()] 'Likability classification -a not so deep neural network approach'. R Brueckner , Schuller . *? in Proceedings of INTERSPEECH*, 2012.

[Mohamed et al. ()] A Mohamed , G E Dahl , G Hinton . *Acoustic modeling using deep belief networks,? Audio, Speech, and Language Processing*, 2012. 20 p. .

[Ngiam et al. ()] J Ngiam , A Khosla , M Kim , J Nam , H Lee , A Y Ng . *Multimodal deep learning,? in Proceeding soft he 28 th International Conference on Machine Learning (ICML)*, 2011. p. .

[Wei-Long Zheng1 and Lu ()] *Personalizing EEG-Based Affective Models with Transfer Learning, Center for Brain-like Computing and Machine Intelligence*, Bao-Liang Wei-Long Zheng1 , Lu . 2016. Shanghai, China. Department of Computer Science and Engineering, Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Brain Science and Technology Research Center, Shanghai Jiao Tong University

[W¨ollmer et al. ()] *Shrikanth Narayanan; Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling; Institute for Human-Machine Communication*, Martin W¨ollmer , Angeliki Metallinou , Florian Eyben , Bj¨orn Schuller . 2010. Los Angeles, CA. Technische Universit¨at M¨unchen, Germany Signal Analysis and Interpretation Lab (SAIL), University of Southern California

[Sivaram and Hermansky ()] 'Sparse multilayer perceptron for phoneme recognition,? Audio, Speech, and Language Processing'. G Sivaram , H Hermansky . *IEEE Transactions on* 2012. 20 (1) p. .

[Kunhan et al. ()] 'Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine'. Dong Kunhan , Ivan Yu , Tashev . *One Microsoft Way* 2014. Microsoft Research. Department of Computer Science and Engineering, The Ohio State University

[Stuhlsatz et al. ()] A Stuhlsatz , C Meyer , F Eyben , T Zieike , G Meier , B Schuller . *Deep neural networks for acoustic emotion recognition: raising the benchmarks,? in Acoustics, Speech and Signal Processing*, 2011. IEEE. p. . (2011 IEEE International Conference on)

[Fan et al. ()] 'Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks'. Y Fan , X Lu , D Li , Y Liu . *Proceeding ICMI 2016 Proceedings of the 18th ACM International Conference on Multimodal Interaction*, (eeding ICMI 2016 eedings of the 18th ACM International Conference on Multimodal InteractionTokyo, Japan) November 12 -16, 2016. p. .

[Zhang et al. ()] 'Xiaoming Zhao; Multimodal Emotion Recognition Integrating Affective Speech with Facial Expression ; Institute of Image Processing and Pattern Recognition Taizhou University Taizhou 318000 CHINA, Hunan Institute of Technology Hengyang 421002 CHINA,Bay Area Compliance Labs'. Shiqing Zhang , Xiaohu Wang , Gang Zhang . *Corp. Shenzhen* 2014. p. 518000.

[Yao et al. ()] A Yao , D Cai , P Hu , S Wang , L Shan , Y Chen . *HoloNet: towards robust emotion recognition in the wild*, 2016.

[Zhang et al.] Zixing Zhang , Fabien Ringeval , Fabien Ringeval , Eduardo Coutinho , Erik Marchi , Björn Schüller . *Semi-Supervised Learning (SSL) technique*,