IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LITERATURE REVIEW ON SPEECH EMOTION RECOGNITION

[1]Tiya Maria Joshy, [2]Dr. Anjana S Chandran

[1]Scholar, [2]Assistant Professor

[1]Master of Coumputer Applications,

[1]SCMS School of Technology and Management, Cochin, India

Abstract: Emotion recognition is a part of speech recognition. There are methods to recognize emotions using machine learning. Emotion recognition is the process of identifying human emotion whereas speech recognition enables the recognition and translation of spoken language into text by computers. This paper explains about machine learning, trends in machine learning, deep learning, and its trends and applications, emotion recognition and speech emotion recognition.

Index Terms **- Machine learning, deep learning, emotion recognition, speech emotion recognition.**

I. INTRODUCTION

Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can achieve input data and use statistical analysis to predict an output while updating outputs as new data becomes available. Machine learning is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment. Machine learning systems are used to identify jobs in images, transcribe speech into text, match news items, posts or products with user's interests and select relevant results of search.

These applications make use of a class of technique called deep learning. Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. Deep learning discovers intricate structure in large data sets by using the back propagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer.

Emotion recognition is the process of identifying human emotions. Whereas speech emotion recognition enables the recognition and translation of spoken language into text by computers which is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT). Speech Emotion Recognition (SER) recognize the human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. SER is tough because emotions are subjective and annotating audio is challenging. The aim of SER system is to extract the emotion from the unknown input speech. SER is more or less a pattern recognition system.

II. MACHINE LEARNING

The main four directions that machine learning gives greater progress is – by learning ensembles of classifiers improvement of classification accuracy, methods for scaling up supervised learning algorithms, reinforcement learning and learning of complex stochastic models. Ensemble classifiers are set of classifiers whose decisions are combined to classify new examples. Ensemble classifiers are more accurate than component classifiers. Reinforcement learning addresses the problems of sequential decision making and control. Reinforcement learning is analyzed as a form of online, approximate dynamic programming. The online approximation of value iteration is Q-learning. The stochastic model describes the real-world process by which observed data is generated. The probabilistic network or stochastic model is a graph structure that captures the probabilistic dependencies among a set of random variables [1].

Machine learning offers different techniques to extract information from data that are translated into knowledge. Machine learning algorithms are categorized into supervised, unsupervised and semi-supervised learning. Supervised learning is learning from labeled data that provides corrective information to the algorithm. Unsupervised learning is the learning of patterns without labeled training data. Whereas semi-supervised learning is the learning with partially labeled data or by receiving a reward from the environment. Fluid mechanisms were traditionally concerned with big data and currently, fluid mechanisms are beginning to tap into the full potential of the powerful methods [3].

Supervised learning is a widely used machine learning method. It includes spam classifiers of email, face-recognizers over images and medical diagnosis systems for patients. Gradient-based optimization algorithms are being used by the deep learning systems. Major trend is the growing concern with the environment in which a machine-learning algorithm operates. New machine-learning methods that are capable of working collaboratively with humans to jointly analyze complex data sets. Recent progress in machine learning had driven the development of new algorithms. Adoption of data-intensive methods can be found throughout science, technology, and commerce which leads to more evidence-based decision making [2].

III. DEEP LEARNING

Deep learning methods are methods with multiple levels of representations. The common form of machine learning is supervised learning. A multilayer neural network can distort the input space. The back propagation equation can be applied to propagate gradients through all modules, starting from the output at the top to the bottom. Feed forward neural network architectures are used by many applications of deep learning. CNN process the data which are in the form of multiple arrays. RNN process input sequence one element at a time. It's a very powerful dynamic system but training them is problematic as the back propagated gradients either grow or shrink at each step [6].

The deep learning framework enables deep learning systems into eLearning and its development, benefits and future trends of deep learning in eLearning. The four common algorithms of deep learning are supervised, unsupervised, semi-supervised and reinforcement learning algorithms. Deep learning applications in eLearning are personalized learning paths, Chabot's, performance indicators, and virtual teaching assistants. Deep learning supports the better allocation of online resources. Deep learning offers an overview of big data and uses it to predict the outcome [12].

Deep learning methods make use of multiple processing layers to learn the hierarchical representation of data. NLP helps computers to perform a wide range of tasks at all levels. Word embedding's can capture syntactic and semantic information. With distributed representation different deep models have become the new state-of-art method for NLP problems [7].

IV. APPLICATIONS

The issues in the design of deep learning models and its application are the definition of pattern classes, sensing environment, pattern representation, feature extraction and selection, cluster analysis, classifier design and learning, selection of training and test samples and performance evaluation. The main idea of the deep learning algorithm is to computerize the extraction of portrayals from the information [13].

Deep learning algorithms have been applied to different fields. The applications that applied deep learning are automatic speech recognition, image recognition, natural language processing, drug discovery and toxicology, customer relationship management, recommendations systems, and bioinformatics. To drive speech features, deep learning uses multiple layers of nonlinear transformation [9]. Survey on supervised convolutional neural network gives a broad view of different supervised Convolutional Neural Network applications with its features. The different fields that use CNN are computer vision for pattern and object detection, natural language processing, speech recognition, medical image analysis [10].

The disadvantage of traffic identification is to find the features in the flow data. The process of traffic identification is time-consuming. Artificial Neural Network (ANN) is mainly used in pattern recognition. Deep learning is a branch of machine learning. Deep learning methods are Deep Neural Network (DNN), Convolutional Neural Network (CNN), Deep Belief Network (DBN) and Stacked Auto-Encoder (SAE). Deep learning replaced handcrafted features with efficient algorithms for unsupervised and semi-supervised learning. Applications are data set, automatic feature learning, protocol classification, and anomalous protocol detection, unknown protocol identification [8].

Robotic medical surgical procedure is training in medical procedures. The mechanical arms are controlled by a specialist using the computer. The mechanical robotic arm developments are constrained by the specialist hand development. The outcomes show the amazingness of machine learning adapting automatic speech recognition in medical-surgical procedures. Machine learning and ASR techniques have better exactness [11].

V. EMOTION RECOGNITION

Emotion recognition is the process of identifying human emotions. A new method was introduced to identify human emotions using heart rate. Using smart bracelet data was collected. Experiments proved that the method is effective. It helps to promote the application and also the development of wearable devices for monitoring human emotional moods in both static and quasi-static states [4]. Emotions are the affective state which influences the behavior and cognitive processes. As a result of external and internal stimuli they appear. Different types of stimuli used for emotion elicitation. Using PPG and GSR signals emotion recognition is possible. Feature selection techniques can maximize the performance with a less number of predictors. Different classification models are trained to select the one that maximizes accuracy and ROC [5].

Using existing data results of the analysis were 31 to 81 percent correct and by using Fuzzy logic 72 to 81 percent for two classes of emotions. The proposed system depends on human brain activities or emotions. In emotion recognition using brain activity, brain activities using EEG signals are used. It's a tough task as it becomes expensive, complex and time-consuming while measuring the human brain with Electroencephalography (EEG). By using a system that is trained by neural networks has achieved 97 percent accurate results [17]. Research challenges are categorization of emotions, the basis of emotion annotations, conversational context modeling, speaker-specific modeling, listener specific modeling, presence of emotion shift, fine-grained emotion recognition, multiparty conversation, presence of sarcasm, and emotional reasoning. An effective emotion-shift recognition model and context encoder can yield significant performance improvement [18].

VI. SPEECH EMOTION RECOGNITION

In deep representation learning in speech processing : challenges, recent advances and future trends present a survey on different techniques of speech representation learning across three research areas including Automatic Speech Recognition (ASR), Speaker Recognition (SR) and Speech Emotion Recognition (SER). Drawbacks of speech processing are feature engineering being manual and require human knowledge and also the designed features may not be the best for the objective. The significance of representation learning has increased with deep learning. Representations are less dependent on human knowledge and are more useful. LSTM/GRU-RNNs in combination with CNN is suitable for capturing speech attributes [15].

HER provides information in applications like human-centered computing, behavior analysis, and cognitive science. The proposed method incorporates both bottom-up and top-down components in the group level emotion recognition. The three-level feature employed was attention based scene level feature, CNN-LSTM face level feature, and RtPose skeleton level feature. This method achieved 62.90% test accuracy [19].

To develop a working model two datasets were used: extended Cohn-Kanade dataset and Japanese Female Facial Expression database in facial emotion recognition in real time. JAFFE provides additional images with more subtle facial expressions. Using the VGGS network along with a face detector an application where an emoji indicates one of the six expressions. The applications of emotion recognition examine the static images of facial expressions. Researchers have developed a system for detecting human emotions in different scenes, angles and lighting conditions in real-time. The result was an application where an emotion - indicating emoji is superimposed over the face [23].

Facial expression recognition faces challenges due to high intra-class variation. Traditional approaches to facial expression recognition were SIFT, HOG, and LBP. The proposed system can focus on important parts of the face and achieve improvements over previous models. From the experiments, it was proved that different emotions seem to be sensitive to different parts of the face. A visualization method to highlight the salient regions of the face images were deployed which highlights the salient regions of face images which are considered the crucial parts in detecting different facial expressions [21].

DNN based ASR system is used to model the speech recognition threshold. DNN based model produces higher prediction accuracy than baseline models. With multi-condition training accurate predictions are obtained. DNN serves as a valuable tool to uncover signal recognition strategies. ASR based prediction of human speech recognition performs the baseline models [14].

Speech recognition involves capturing and digitizing sound waves, converting to basic language units, constructing words and analyzing words to ensure correct spelling for words that sound alike. Neural networks are naturally discriminative. Attributes of the neural network are set of the processing units, set of connections, computing procedure, and training procedure [16].

Big data includes speech and video. In the proposed system, the speech signal is processed in the frequency domain to obtain Mel-spectrogram. Mel-spectrogram is fed to the convolutional neural network (CNN). Representative frames from video segments are extracted and fed to CNN for video signals. The output gives a support vector machine (SVM) for the classification of emotions. 2D CNN for speech and 3D CNN for video signals were used in the proposed system. Compared to the classifier's combination, ELM based fusion performance was better [20].

An EMOTIC database is presented which contains a dataset of images in non-controlled environments. According to the people's apparent emotional state, images are annotated combining 2 types of annotations. Instead of recognizing emotions some recent works on facial expression use the continuous dimensions of the VAD Emotional State Model to represent emotions. The VAD describes emotions using 3 dimensions: Valence (V), Arousal (A), and Dominance (D). EMOTIC database is composed of images. The database contains a total number of 18,316 images having 23,788 annotated people. The database contains two different emotion representation formats - discrete categories and continuous dimensions [22].

VII. CONCLUSION

Speech is the most natural and environment friendly way of verbal exchange between humans. Lots of efforts have been made to boost a human computer interface so that one can easily have interaction and talk in an unskilled way. With the dawn of AI human like Automatic Speech Recognition (ASR) accuracy is doable in voice primarily based shrewd agents. Emotion recognition from speech statistics can make the machine extra human like and it can resource the robot to respond according to the tone and sentiments of the user’s voice.

REFERENCES

[1] Thomas G. Dietterich. 1997. Machine-Learning Research Four Current Directions. AI Magazine Volume 18 Number 4.

[2] M. I. Jordan, and T. M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects.

[3] Steven L. Brunton. 2020. Machine learning for fluid mechanisms.

[4] Lin Shu, Yang Yu, Wenzhuo Chen, Xiangmin Xu. 2020. Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet.

[5] J.A. Dominguez-Jimenez, K.C. Campo-Landines, J.C. Martinez-Santos, E.J. Delahoz, S.H. Contreras-Ortiz. 2020. A machine learning model for emotion recognition from physiological signals. – Biomedical Signal Processing and Control

[6] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. 2015. Deep learning

[7] Tom Young, Devamanyu Hazarika, Soujanya Poria and Erik Cambria. 2018. Recent Trends in Deep Learning Based Natural Language Processing.

[8] Zhanyi Wang. 2015. The Applications of Deep Learning on Traffic Identification.

[9] Nur Farhana Hordri, Siti Sophiayati Yuhaniz and Siti Mariyam Sha. 2016. Deep Learning and Its Applications: A Review.

[10] D. T. Mane and U. V. Kulkarni. 2017. A Survey on Supervised Convolutional Neural Network and Its Major Applications. International Journal of Rough Sets and Data Analysis Volume 4 Issue 3.

[11] J.Ruby, Susan Daenke, Yanmin Yuan, William Harry, J.Tisa, J.Nedumaan, Yang Yung, J.Lepika, Thomas Binford and P.S.Jagadeesh Kumar, Wenli Hu. 2020. Automatic Speech Recognition and Machine Learning for Robotic Arm in Surgery. American Journal of Clinical Surgery.

[12] Anandhavalli Muniasamy, Areej Alasiry. 2020. Deep Learning: The Impact on Future eLearning. iJET Vol. 15 No. 1.

[13] Redouane Lhiadi. 2020. Deep Learning Algorithm and Their Applications in the Perception Problem.

[14] Birger Kollmeier, Constantin Spille, Angel Mario Castro Martinez, Stephan D. Ewert and Bernd T. Meyer. 2020. Modelling human speech recognition in challenging noise maskers using machine learning. The Acoustical Society of Japan Acoust. Sci. & Tech. 41, 1.

[15] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Bjorn W. Schuller. 2020. Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends.

[16] Akhilesh Halageri, Amrita Bidappa, Arjun C, Madan Mukund Sarathy and Shabana Sultana. 2015. Speech Recognition using Deep Learning. IJCSIT, Vol. 6 (3).

[17] Dilbag Singh. 2012. Human Emotion Recognition System. I.J. Image, Graphics and Signal Processing.

[18] Soujanya Poria, Navonil Majumder, Rada Mihalcea and Eduard Hovy. 2019. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. IEEE Access.

[19] Li, Dejian, Luo, Ruiming, Sun and Shouqian. 2020. Group-Level Emotion Recognition Based on Faces, Scenes, Skeletons Features. Graphics and Image Processing (ICGIP).

[20] M. Shamim Hossain and Ghulam Muhammad. 2019. Emotion Recognition Using Deep Learning Approach from Audio-Visual Emotional Big Data. Information Fusion, vol. 49.

[21] Shervin Minaee and Amirali Abdolrashidi. 2019. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network.

[22] Ronak Kosti, Agata Lapedriza, Jose M. Alvarez and Adria Recasens. 2017. Emotion Recognition in Context - Pattern Recognition.

[23] Dan Duncan, Gautam Shine, Chris English. 2016. Facial Emotion Recognition in Real Time.