

EMOTION ANALYSIS USING SIGNAL AND IMAGE PROCESSING APPROACH BY IMPLEMENTING DEEP NEURAL NETWORK

S. Sushma [1], T. Christy Bobby [2] and S. Malathi [2]

[1] Department of Computer Science and Engineering,
[2] Department of Electronics and Communication Engineering,
Ramaiah University of Applied Sciences, Bengaluru, Karnataka, India

Corresponding Author: S. Sushma
Email: sushma.snk@gmail.com
https://doi.org/10.34107/BiomedSciInstrum.57.04313

ABSTRACT

Emotion recognition is important in human communication and to achieve a complete interaction between humans and machines. In medical applications, emotion recognition is used to assist the children with Autism Spectrum Disorder (ASD to improve their socio-emotional communication, helps doctors with diagnosis of diseases such as depression and dementia and also helps the caretakers of older patients to monitor their well-being. This paper discusses the application of feature level fusion of speech and facial expressions of different emotions such as neutral, happy, sad, angry, surprise, fearful and disgust. Also, to explore how best to build the deep learning networks to classify the emotions independently and jointly from these two modalities. VGG-model is utilized to extract features from facial images, and spectral features are extracted from speech signals. Further, feature level fusion technique is adopted to fuse the features extracted from the two modalities. Principal Component Analysis (PCA is implemented to choose the significant features. The proposed method achieved a maximum score of 90% on training set and 82% on validation set. The recognition rate in case of multimodal data improved greatly when compared to unimodal system. The multimodal system gave an improvement of 9% compared to the performance of the system based on speech. Thus, result shows that the proposed Multimodal Emotion Recognition (MER outperform the unimodal emotion recognition system.

Keywords: Multimodal emotion recognition, deep learning network, feature-level fusion, principal component analysis.

INTRODUCTION

Human emotion recognition is multimodal in nature as people not only listen to what others say, but also observe voice modulations, facial expressions, and body gestures of the speaker during any conversation. Audio-visual data plays a vital role in improving human-machine interaction.

The purpose of speech emotion recognition is to automatically detect emotional state based on his/her voice. The emotional state of a person hidden in speech is an important factor of human interaction as it gives feedback in communication without altering linguistic contents. The analysis of semantic information has been the focus of research in the field of speech emotion recognition. Generally, features like pitch and energy with their means, medians and standard deviations are concatenated with higher level features like word duration and speaking rate. Gaussian Mixture Model – Hidden Markov Model (GMM-HMM approach has been the dominant technique for a long time in speech recognition, with an underlying shallow generative model of context dependent GMMs and HMMs [1]. Inspired by the success of deep learning in multiple fields, many deep learning approaches have been investigated for the task of Speech Emotion Recognition (SER. Deep learning and Deep Neural Network (DNN began making their influence on speech recognition in 2010, after close collaboration between academic and industrial researchers. According to the study, most important audio features for affect recognition are pitch and energy related. Recently, some spectrum features such as, Linear Prediction Coefficients [2], Linear Prediction Cepstrum Coefficients [3], Mel-frequency Cepstrum Coefficients (MFCCs [4] and its first derivative, are used for emotion recognition. In case of SER, signal disturbed by noise and differences between voices of different people are the factors affecting the performance of emotion recognition method.