Emotion Features Selection and Detection using Two Modality Neural Network
Ganesh Kumaarr. S¹, Kamakshi Kumar. A. V², Mr. S. Balasubramanian³
UG Student¹,², Associate Professor³
Department of Electronics and Communication Engineering
Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India

Abstract:
Human-Computer-Interactions are increasing day by day. Adding emotions to the interactions make the computers feel more human. Emotion addition involves emotion detection in the first hand. To identify emotions we will study them using two modalities i.e., a particular form of sensory perception and here it is the visual (image) and auditory (speech) perceptions. Then analyze them which involves steps such as wavelet transformations, feature extractions. Here we focus on textural and MFCC analysis for image and speech respectively. In this system different step for emotion detection i.e., first the Input Acquisition, segmentation, pre-processing, extracting features that define different emotions, training the neural networks using them and finally classification methods are used. The Feature extraction by digital image processing method includes detection of edges and textures followed by detection of low level features. The Feature extraction by digital signal processing includes determination of mathematical coefficients. Here we proposed the Convolutional Neural Network for Image and Recurrent Neural Network for signal to classify different emotions.

Keywords: Emotion Classification, RGB2GRAY, Median filter, DRLBP, GLCM, Discrete Wavelet Transform, Mel Frequency Cepstral Coefficients (MFCC).

1. INTRODUCTION

Emotion is that one thing that differentiates humans from machines. But the increased interactions with machines made humans want the machines to feel like them. In order to exhibit emotions, it is necessary to learn them before. We have taken data for two sensory perceptions, visual and audio, that contain wide range of emotions from different persons and used them as the datasets to train the neural networks that forms a machine’s brain and enable it to classify emotions with accuracy similar to how humans do.

1.1 Research Motivation
Though the technology has seen a fast growth in the last decade, sophisticated machines are being developed, they are still being treated as machines and that’s because they lack emotions. So in order to make machines more human and make them to offer proactive suggestions we first need to teach them emotions. This is done through machine learning where the enormous emotion data are fetched to train the system containing neural networks that mimics the human brain, which helps the system to exhibit artificial intelligence that cleverly classifies emotions showed by users and offer them similar emotional responses or proactive responses or suggestions in turn.

1.2 Project Objective
Our project intends to develop a system that detects various emotions. Here we use facial images and speech cues containing different emotions to train the system and teach them about that. Then we use neural networks to classify emotions based on training and enable it to offer proactive responses. Here we use Convolutional and Recurrent neural networks to train the system learn emotions and perceive them in two different ways, visual and auditory perceptions and later classify them through artificial intelligence.

2. PROPOSED SYSTEM
This system contains two subsystems which are:

i. Facial emotion classification system
ii. Speech emotion classification system

2.1 Facial Emotion Classification System
It involves

- Wavelet transformation. Edge detectors, texture descriptors and extraction of low level features (Textural features)
- Convolutional Neural Network classifier

![Figure 1. Facial Emotion classification system](image-url)
2.2 Speech Emotion Classification System

It involves

- Wavelet transformation, extraction of acoustical coefficients (Mathematical features)
- Recurrent Neural Network classifier

Figure.2. Speech Emotion classification system

3. TEXTURAL FEATURES EXTRACTION

3.1 Input Image

Some of the input images are taken for feature extraction.

3.2 Face Detection

Face detection is a process to extract face regions from input image which has normalized intensity and uniform in size. The appearance features are extracted detected face part which describes changes of face such as furrows and wrinkles (skin texture). In this system model, the face detection process is based on Haar like features along with adaptive boosting method.

3.3 RGB2GRAY Conversion

The rgb2gray function converts RGB images to grayscale by eliminating the hue and saturation information while retaining the luminance.

3.4 Median Filter Process

The median of a distribution is the value for which larger and smaller values are equally probable. To calculate the median of a list of sample values, sort them in any order, and then peek the central value, or the mean between the two central values if the list is even-sized. If your list of values has a strong central tendency, which manifests itself as a single, well defined peak on the histogram, then the median is a good estimator of the peak position.

3.5 Discrete Wavelet Transform (2D)

The DWT provides a sparse representation for many natural signals. In other words, the important features of many natural signals are captured by a subset of DWT coefficients that is typically much smaller than the original signal. This "compresses" the signal. With the DWT, you always end up with the same number of coefficients as the original signal, but many of the coefficients may be close to zero in value. As a result, you can often throw away those coefficients and still maintain a high-quality signal approximation.

Figure.3. Image Decomposition after 2-D Discrete Wavelet Transform

3.6 DRLBP Descriptor

The descriptor local binary pattern is used to compare all the pixels including the centre pixel with the neighbouring pixels in the kernel to improve the robustness against the illumination variation. DRLBP is represented in terms of set of normalized histogram bins as local texture features. It is used to discriminate the local edge texture of face invariant to changes of contrast and shape.

Figure.4. DRLBP Image

3.7 GLCM Features

It is a texture mapper. The graycomatrix function creates a gray-level co-occurrence matrix (GLCM) by calculating how often a pixel with the intensity (gray-level) value \( i \) occurs in a specific spatial relationship to a pixel with the value \( j \). Each element \((i,j)\) in the resultant GLCM is simply the sum of the number of times that the pixel with value \( i \) occurred in the specified spatial relationship to a pixel with value \( j \) in the input image.

Figure. 5. GLCM Matrix Creation

When you calculate statistics from these GLCMs, you can take the average. These statistics provide information about the texture of an image. Statistic such as Contrast, Correlation, Energy, Homogeneity, Entropy gives information about image.
4. MATHEMATICAL FEATURES EXTRACTION

4.1 Input Signal
Some of the speech signals are taken for mathematical features extraction.

4.2 Discrete Wavelet Transform (1D)
The DWT provides a sparse representation for many natural signals. In other words, the important features of many natural signals are captured by a subset of DWT coefficients that is typically much smaller than the original signal. This “compresses” the signal. With the DWT, you always end up with the same number of coefficients as the original signal, but many of the coefficients may be close to zero in value. As a result, you can often throw away those coefficients and still maintain a high-quality signal approximation.

4.3 Mel Frequency Cepstral Coefficients (MFCC)
The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.

5. NEURAL NETWORK

5.1 Convolutional Neural Network
In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery. CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.

5.2 Recurrent Neural Network
A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Unlike feedforward neural network, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition or image classification.

6. RESULTS

Figure 12. Classified Emotion using CNN
they have fast speed on training and simple structure. More images and signals were used to train the classifiers and tests were run on different sets of data to examine the classifiers accuracy. At last, the output from both the classifiers were compared using a relational operator and the final output(s) were generated which were left to the kind of application to make a conclusion.

7.1 Future Scope
The future scope of the emotion detection system is that it can be more accurate and efficient. The system can be trained using dynamic data from the static data used in this system. The system can be improved to an always vigilant system. The system can be more reliable and robust if it is further trained using enormous sizes of datasets. In future it can be made more interactive and user friendly for detecting the right emotion in an application.

ACKNOWLEDGEMENTS
We feel elated in manifesting a deep sense of gratitude and thankfulness to the management of Meenakshi Sundararajan Engineering College, our beloved Head of the Department and all our staffs for their constant motivation.

8. REFERENCES


[7]. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)” by Livingstone & Russo is licensed under CC BY-NA-SC 4.0.