Ubiquitous Emotion Recognition using Audio and Video Data

Rahatul Jannat
University of South Florida
Tampa, FL, USA
scanavan@usf.edu

Iyonna Tynes
University of South Florida
Tampa, FL, USA
iyonnatynes@usf.edu

Lott LaLime
University of South Florida
Tampa, FL, USA
lalime@usf.edu

Juan Adorno
University of Puerto Rico Bayamon
Bayamon, Puerto Rico
Juan_ardorno@me.com

Shaun Canavan
University of South Florida
Tampa, FL, USA
scanavan@usf.edu

Abstract
In this paper we present a method for recognizing emotions using video and audio data captured from a mobile phone. A mobile application is also presented that captures audio and video data, which were used to predict emotion with a convolutional neural network. We show results of our deep network on images taken from the BP4D+ [11] database, and audio signals taken from the RAVDESS dataset [4], which were also used to train the CNN used in the presented mobile application.

Author Keywords
Ubiquitous computing, emotion recognition, machine learning, audio, video

ACM Classification Keywords
H.m[Information Systems]: Miscellaneous

Introduction and Background
An important part of human intelligence is recognizing emotion [7], as it has applications in fields such as entertainment, transportation, medicine, and psychology. The use of video (images) and audio data have shown promise in emotion recognition. Liu et al [5] used a boosted deep belief network for this task. They developed a network that performed feature learning, selection, and classifier construction iteratively in a unified loopy framework. Sinith et al [8] conducted emotion recognition experiments using audio signals with a support vector machine. Features were extracted from the signals showing

promising results on 4 emotions. Recently, with devices becoming more ubiquitous, research has turned towards ubiquitous emotion recognition. Suk et al [9] classified 6 emotions by training a support vector machine by creating dynamic features from landmarks fit with an Active Shape Model [1]. Considering the success of these works, in this paper we propose the fusion of video (images) and audio signals from a mobile phone for emotion recognition. The rest of the paper details our proposed method, including the presentation of a new mobile application, results, and a discussion of the results along with ideas on future methods for ubiquitous emotion recognition using video and audio data.

Methodology

Our ultimate goal is to fuse audio and video data for the task of emotion recognition. To do this we investigate the use of 2 publicly available datasets. For video (image) data, we use the BP4D+ multimodal emotion corpus [11]. This database is a large (~14 TB in size) collection of multimodal data consisting of 140 subjects expressing 10 targeted emotions (e.g. happy, said, pain, fear). The data includes RGB images, 3D face models, 2D and 3D facial landmarks, thermal data, physiological data, and action units. For our experiments we use approximately 13,000 RGB images from this dataset.

The next dataset we use is the Ryerson Audio-Visual Database of Emotional Speech and Song [4]. This database consists 24 actors vocalizing statements in a North American accent. There is a total of 7356 recordings available with 7 emotions (e.g. calm, happy, sad, disgust). For our experiments, we use approximately 700 recordings from this dataset.

Audio and Video Pre-processing

Before we can efficiently fuse this multimodal data, we must first pre-process the data. For the image data, we first detect the face in each image using Haar features [10]. Once the face is detected, we crop the image to include the face region and scale it to 256x256 (see figure 1). To pre-process the audio data, we plot the raw audio signal onto the 2D image plane. The final waveform image is also scaled 256x256 (see figure 1), to be consistent with the face data.

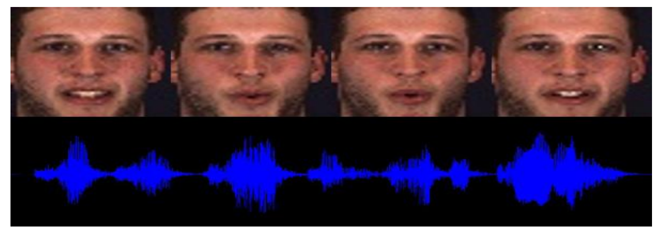

Figure 1. Pre-processed data. Top row: Cropped faces from BP4D+; bottom row: plot of raw audio signal from RAVDESS.

Convolutional Neural Networks

Convolutional Neural Networks (CNN) are a popular deep network for images, and audio data. They have shown success in areas such as emotion recognition [12], face recognition [6], and speech recognition [3]. Due to their success in these areas, we have chosen to employ them for our multimodal emotion recognition task. We used an Inception V3 CNN with 3 convolutional layers of size 32, 64, and 128, each followed by max-pooling, with a final fully connected layer used in the output. The Adam optimizer was used with a learning rate of 0.0003.

Experimental Design and Results

Given pre-processed image data that includes faces and audio waveforms, we then train our deep network to recognize emotion. We conducted 3 experiments, with 3 separately trained networks, to do this. We trained one network only on image data, another only on the plotted audio waveforms, and the third on both image and waveform data. This was done to test the accuracy of our method using a single modality compared to a multimodal approach.

To test our networks, we trained on 2 emotions (happy and sad) for both audio and video data. We employed a 90/10 split of training and testing data (trained on 90% of the data and tested on 10%). To train our network that contains both audio and image data, each of the signals were used as a single instance of emotion (i.e. one face had a class of happy, one separate waveform had a class of happy). Results of each of the experiments (audio, video (images), audio and video) are detailed in table 1.

Table 1. Experimental results.

Experiment	Loss	Accuracy
Image	17.29%	99.22%
Audio	58.28%	66.41%
Image/Audio	19.70%	96.09%

As can be seen from table 1, the image data resulted in the highest accuracy with 99.22% accuracy and the waveform format of the audio signals had the lowest recognition with 66.41%. Interesting to note is that when our network was trained on both audio and video data, the results are only slightly lower compared to video alone (<3%). This is compared to audio alone which has a recognition rate that is >30% less than images alone. In our final discussion we will discuss fusion techniques, and other formats for the audio signals that may increase accuracy.

Ubiquitous Emotion Recognition

The presented ubiquitous application was developed to run on PCs (e.g. laptops), and Android devices. The application captures face and audio data. Like the offline pre-processing, the face was detected using Haar features, and cropped to a size of 256x256 (see figure 3). To be able convert the captured raw audio signal to a plotted waveform, we created a lightweight Node Js server. The audio file is captured from the application, sent to the server, which is then transformed to the 2D image plane (waveform plot), and finally sent back to the phone. See figure 2 for an overview of this process.

Figure 2. Overview of Node Js server to plot waveform.

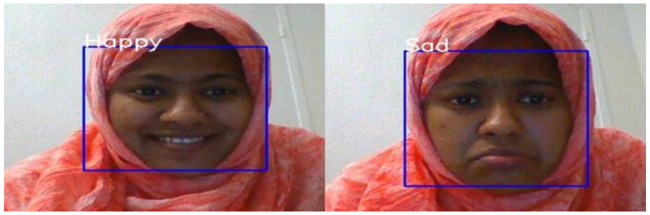

Figure 3. Ubiquitous application running in real-time with detected face and recognized emotion shown.

Discussion

We have presented a method for recognizing emotion using audio and video data, including a method for representing raw audio signals as a plotted waveform. We have also presented an application that uses these modalities in real-time. Although the results are encouraging with image data, the audio data results in a low accuracy when used independently and lowers the image accuracy results when used together. To address this, we propose 2 possible solutions to this problem. First, is to use the raw audio signals by splitting them into blocks of time and using this raw data to train our deep networks. Second, is the fusion of the modalities. This can be done by creating a new image from the face and audio images. This approach to image fusion has shown success in face recognition [1]. It is important to note that for this paper, we only trained and tested our networks on 2 emotions. Intuitively, as the number of emotion classes increases the recognition accuracy of a single modality will decrease. Our approach to multimodal fusion of audio and video data can address this by providing a strong multimodal representation of emotion.

References

1. S. Canavan, et al., "Face rec. by multi-frame fusion of rotating heads in videos," BTAS 2007.
2. T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models-their training and application," Computer Vision and Image Understanding, 61(1): 38-59, 1995.
3. G. Hinton, et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," IEEE Signal Processing Magazine, 29(6): 82-97, 2012.
4. S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expression in North American English," PloS one, 13.5: e0196391, 2018.
5. P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," CVIU, 2014.
6. O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," BMVC, 1(3): 6, 2015.
7. R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," IEEE Trans. on PAMI, 23(10): 1175-1191, 2001.
8. M. Sinith, et al., "Emotion recognition from audio signals using support vector rec.," Recent Advances in Intelligent Computational Systems, 2015.
9. M. Suk and B. Prabhakaran, "Real-time model expression recognition system – a case study," CVPR Workshops, 2014.
10. P. Viola and M. Jones, "Robust real-time face detection," Intl. Journal of Computer Vision, 57(2): 137-154, 2004.
11. Z. Zhang, et al., "Multimodal spontaneous emotion corpus for human behavior analysis," CVPR, 2016.
12. Y. Zhiding and C. Zhang, "Image based static facial expression recognition with multiple deep network learning", ICMI, 2015