Music-Video Emotion Analysis Using Late Fusion of Multimodal

Yagya Raj PANDEYA and Joonwhoan LEE∗
Chonbuk National University, Jeonju, South Korea
∗Corresponding author

Keywords: Music-video dataset, Multimedia emotion, Audio-video multimodal, Late fusion.

Abstract. Music-video emotion is a high-level semantics of human internal feeling through singing music lyrics, musical instrument performance and visual expression. Any online and off-line music video are rich source to analysis the emotion using modern machine learning technologies. In this research we make music-video emotion dataset and extract music and video features from pre-trained neural networks. Two pre-trained audio and video networks are first fine-tuned and then extract the low level and high-level features. The music network use 2D convolutional network and video network use 3D convolution (C3D). Finally, we concatenate each music and video feature by preserving the time varying features. The long short-term memory (LSTM) network is used for long-term dynamic feature characterization and then various machine learning algorithms evaluates emotions. We also use late fusion to fuse the learned features of audio and video network. The proposed network performs better for music-video emotion analysis.

Introduction

Music is a language that communicates some emotion to anyone, even to plants or animals. The music composers try to speak through musical instruments so exact music emotion measurement is not possible even for music experts. The music makers generally use dynamics (tempo, meters) or articulation to evoke emotions in music but everybody have different strategies to includes certain feel. The human emotion is a very vague and person dependent situation that may have influence of visual or acoustic information, human though and environmental changes. In music-video composition, the composer always tries to make some feel through musical instrument, singing methods and visual expression. The various music-video emotion expresses various high-level semantics of human moods. This makes the automatic music-video emotion content analysis highly challenging for computer.

Human emotion representation is still an area of research and music video emotion is a specific to various human emotion specified in [1]. Individual audio or video emotion bases studies [6, 7] are done in past decades and in recently many [4, 8] research primarily focus on both audio and video emotion with early or late fusion. The primary limitation in this area of research is lack of labeled data and the methods based on data driven are not properly conducted. In this research our primary task is introduction of a small music-video dataset that can attract new researcher in this area and solve various problems related to music-video in television data center and several online video banks.

Organization of the Text

The DEAP [2] is a first music-video emotion annotated dataset with only 120 data sample, which is not enough for data driven emotion analysis. In this research we make six classes music-video emotion dataset including some DEAP dataset data samples. The number of class for our music-video emotion dataset are chosen according to [3, 4, 5] dataset representation but proper annotation of each data sample was great challenge to us. Finally, we make a dataset including various human emotion adjectives as in Table 1 and the V-A space representation is shown in Fig. 1.
Table 1. Description of music-video dataset with various adjectives and number of data samples in each emotion classes.

<table>
<thead>
<tr>
<th>Emotion Class</th>
<th>Emotion Adjectives</th>
<th>No. of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tension</td>
<td>Anger, Hate, Rage,</td>
<td>362</td>
</tr>
<tr>
<td>Fear</td>
<td>Horror, Fear, Scary, Disgust, Terror</td>
<td>362</td>
</tr>
<tr>
<td>Excitation</td>
<td>Happy, Fun, Love, Sexy, Joy, Pleasure, Exciting, Adorable, Cheerful, Surprising, Interest</td>
<td>359</td>
</tr>
<tr>
<td>Sad</td>
<td>Hate, Depressing, Melancholic, Sentimental, Shameful, Distress, Anguish</td>
<td>349</td>
</tr>
<tr>
<td>Neutral</td>
<td>Little (sad, Fearful, Exciting, Relax) Ecstasy, Mellow,</td>
<td>350</td>
</tr>
<tr>
<td>Relaxation</td>
<td>Calm, Chill, Relaxing</td>
<td>353</td>
</tr>
</tbody>
</table>

In this emotion dataset, the six emotions are differing by frequency, pitch, energy, zero crossing rate, motion intensity, color energy, lighting, rhythm regularity, etc. There is both inter class and intra class correlation and differentiation in data samples. The ‘Neutral’ class includes common characteristics of other emotion classes so the emotion classification using deep neural network is challenging task.

![Valence-Arousal emotion space](image)

Figure 1. A Valence-Arousal emotion space for music-video content analysis.

**Methodology**

This section covers proposed neural network architecture, music-video dataset feature extraction and classification of these features using various classifiers.

**Input Preprocessing and Transfer Learning**

The music data preprocessing for pre-trained convolutional neural networks (CNN)[9], first need zero padding to make the full length audio that can generate the fixed size mel spectrogram. The role of zero padding in the time domain signal is to increase the frequency resolution and make full periods in signal that remove the spectral leakage. The sampled music signals with 16 kHz rate is represented by mel spectrogram (96-mel bins x 1876 temporal frames) using fourier transform. To process the visual information, we first subsample the video by 300 frames (interval and colored) and then process
images to fit the 3D convolutional network for video (C3D)[10] pre-trained network input shape i.e. (112, 112, 32, 3).

After input processing, transfer learning and fine tuning is made of pre-trained network weight for music-video data. We separately make transfer learning for acoustic and visual feature from the respective pre-trained CNN audio network previously trained on million song dataset[11] and C3D video network trained on sport-1M dataset[12].

**Multimodal Architecture**

Transfer learning has been proved effective for deep audio-visual emotion recognition based on small scale training data. The features are extracted from pre-trained C3D and CNN using our emotion dataset. The features from pre-trained C3D network are first passed to recurrent neural network (RNN) to capture the varying time sequence behavior of video tube. The extracted features from audio and video networks are then concatenated and classify using various classifiers. The detailed architecture of our proposed neural network is shown in Fig. 2. The features are extracted from these two networks with and without finetuning but the fine tuning feature perform better performance.

![Multimodal Architecture](image)

**Feature Classification**

The features of pre-trained audio and video network are concatenated and then classified with various machine learning algorithms. We use four classifiers namely, random forest, k-nearest neighbors, extra tree and C-support vector machine classifiers and then ensemble them. In this experiment ensemble classifier perform best as it includes the collective features of different type of classifiers. The features with and without fine tuning are classified using five classifiers.

**Results and Discussion**

In our study, the learned features from audio and video networks are classified using popular sklearn ([http://scikit-learn.org](http://scikit-learn.org)) library. The individual classifier performance and their ensemble results are illustrated in Table 2.
Table 2. Classification of music-video emotion dataset after late fusion.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Random Forest</th>
<th>K-Nearest Neighbors</th>
<th>Extra Trees</th>
<th>C-Support Vector</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Set Accuracy</td>
<td>71.96</td>
<td>53.27</td>
<td>68.69</td>
<td>72.11</td>
<td>73.2</td>
</tr>
<tr>
<td>F-score</td>
<td>0.71</td>
<td>0.53</td>
<td>0.67</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>ROC AUC Score</td>
<td>0.933</td>
<td>0.721</td>
<td>0.918</td>
<td>0.943</td>
<td>0.958</td>
</tr>
</tbody>
</table>

The evaluation metric to evaluate our system performance are accuracy, F1-score, and area under the receiver operating characteristic curve (ROC-AUC) scores have been employed. The accuracy refers to the percentage of correctly classified unknown data samples, and F1-score computes harmonic mean between precision and recall. The ROC-AUC score is measured from each classifier in the receiver operating characteristic (ROC) curve. The ROC curve is used to visualize and analyze the performance of each classifier according to the various decision thresholds associated with it. The Confusion matrix shows the precise performances for the different classes. Our best performing ensemble classifier confusion matrix and the ROC-AUC curve are shown in Fig. 3.

**Figure 3.** Best performing ensemble classifier (Left) Confusion Matrix and (Right) ROC-AUC curve.

The music-video contents communicate some emotions according to composer requirements, but everybody is different and there is no right or wrong answer to certain feel. Hence, it can be concluded that listing some music always make some kinds of feel in human in certain way. For example, the sound of quite stream makes peacefulness or heavy traffic make feeling jittery. As music, the video tube also includes some emotions for human. Hence, music-video is a rich source of human emotion. But it is hard to evaluate the exact feeling of emotion in music-video. According to our labeled dataset, pre-trained features and classifier we obtain good results. If we can make more data samples, it obviously increases the system performance in future.

**Summary**

The automatic human emotion classification is one of the hardest task for computer as it needs highly diversified evaluation operators. We tried to solve this problem by, first introducing a small music-video emotion dataset and then classify the learned features from pre-trained neural network using different classifiers. Our result show comparable result with state-of-art results with very simple neural network architecture and classifiers. We hope that this research will attract new researchers in future for more accurate music-video emotion computation.
Acknowledgement

The research leading to these results, authors would like to thank BK21 under Chonbuk National University.

References


