Musically Classifying Emotion using Video Sourced Facial Action Units and 25-Keypoint Body Pose Positions to Train SVM, k-NN, and Neural Network Models

Orion Hsu - 301283481
Yoonhong Lee - 301267876
ohsu@sfu.ca
yla382@sfu.ca

ABSTRACT

Human emotions are expressed visually through modalities such as body language and facial movements, however many of the cues for emotion are subtle and subjective due to factors from culture, context, and secondary emotions. Classifying emotion with unimodal data from facial expressions can result in accurate identification of certain emotions largely expressed through facial movement, such as joy or sorrow, however it often lacks the contextual clues from body language, such as a slumped posture. In this paper, we seek to use continuous video sources to combine unimodal data generated from OpenPose's Body-25 model and OpenFace's Facial Action Unit identifier into a multimodal dataset. Using the combined OpenPose and OpenFace results frame by frame, the joined data is used to train the classification model. Using body language and facial expressions together provides a larger frame of context that can be used to classify emotions with more resolution and accuracy and allows usage of entire frames without cropping to the face while only requiring video footage. This approach provides insight into subtle social cues and cultural cues, albeit from a single cultural perspective.

KEYWORDS

OpenPose, OpenFace, Classifying Emotion, Facial-Bodily expression, Body-25 model, K-Nearest Neighbour, Neural Networks, Support Vector Machine

1 INTRODUCTION

Recognition of non-verbal emotional expressions is imperfect in computers and is also imperfect in people despite the ability being trained and honed throughout infancy and childhood as well as possessing the ability to generate these emotions themselves [1]. In facial-bodily emotions, there are many cultural nuances in social cues differentiating similar secondary emotions such as shame and sadness [2]. It is also important to note that while cultural differences in emotional expression and appraisal, there is some degree of universality in facial expressions [3], which is also evidenced by most emotions having similar accuracy of recognition across different cultures [2]. With these factors taken into consideration, a model can be constructed to classify emotions with potential accuracy higher than classifiers exclusively using facial emotion detection due to the integration of the modality of body pose, or body language into our model. Culture specific social cues can vary depending on the dataset used to train the model due to the social cues of the culture found in the videos comprising the dataset [2]. When annotating the data, the gender of the group or people making the annotations based off of the data would have an effect on the emotions identified. This is due to male and females identifying certain emotions differently [4]. The overall accuracy of emotion recognition depends on how broadly emotions are defined, modalities used, type of classifier used, and the dataset. Emotion recognition software often uses video data to capture facial expressions, as well as using physiological data such as electro-dermal activity and cardiac activity [5]. Our approach seeks to provide an accurate way of recognizing emotion by pairing it with a corresponding music clip using only a video camera.

2 METHODOLOGY

In this section, we explore the gathering and the preparation of the data, as well as the configurations of the libraries used. A brief overview will also be given outlining the models used to generate results. The process outlining preprocessing and model generation can be found at this Github repository.

2.1 System and Software Configuration

Generating the models and processing live data was done on consumer grade hardware (Intel Core i7 9750H and CUDA with Nvidia RTX 2060). Below is a list of software prerequisites to replicate the process done in this paper. Processing the data was done roughly at a rate of 25 frames per second, however, faster hardware will allow for faster processing of data.

- Windows 10
- Python 3.7
- CUDA Version 10
- Visual Studio 2017
- WSL (Ubuntu)

Below are the libraries used in the implementation of code.

- OpenFace
- OpenPose
- FFmpeg
- YouTube-dl
- scikit-learn
- tslearn
- UMAP
- Pandas
- NumPy

OpenPose was compiled using the Windows 10 CMake GUI instructions located on the Github repository for OpenPose, using the –BUILD_PYTHON flag in order to use the Python API. The precompiled Windows binaries for OpenFace were used. FFmpeg

and YouTube-dl can be obtained from their respective websites. The remainder are Python3 modules that can be installed using pip.

2.2 Collecting Data

The dataset used to train a model can affect the general effectiveness of it's performance. Datasets can be biased or have too much variance, either of which will have their own impact on the model [6].

Therefore the process of collecting data is important in generating a model that is accurate for it's intended purpose. In order to obtain data that has shows an adequate amount of facial-bodily expression, we decided to use music videos and acting samples from YouTube because they provide acted emotion combined with clear footage and a wider camera view. The types of videos chosen contained primarily one person, which allows for better recognition of facial-bodily expressions from OpenFace and OpenPose. An example of our specifically selected videos can be found in the appendix for the dataset as the first entry (A). In addition to using our own vetted videos, we selected a YouTube playlist of popular music videos, mostly comprising of a single person in the video.

To obtain the videos used in the dataset we used YouTube-dl to extract the videos from YouTube. Each video was downloaded as an mp4 file containing both the audio and the video frames. Each frame in the videos were processed by a custom written Python application that runs OpenPose on the GPU and OpenFace on the CPU in parallel on one video at a time. The result are csv files containing facial activation unit data and json files containing body pose coordinate data. This data is to be loaded into Pandas Dataframes to be used to generate the model. The data in each frame can be blended together to form a visualization as seen in Figure 1. Of the 366 music videos and acting samples, we selected 188 to be used for our dataset.

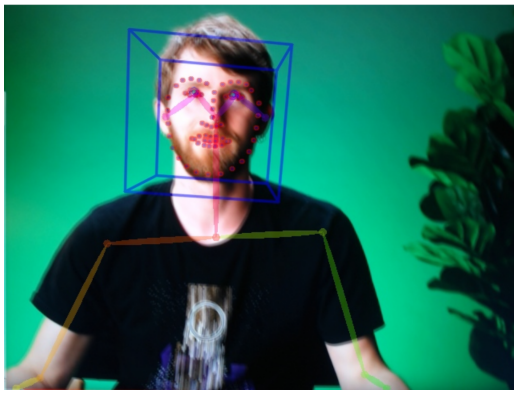

Figure 1: Example visualization from a trending YouTube video (Linus Tech Tips).

2.3 Preprocessing the Data

The OpenPose json files are appended together, then converted into csv files corresponding to each video instead of each frame. This will result on csv files for each video, for both OpenFace and OpenPose. Once the data is loaded into Dataframes, we begin filtering out data, starting with filtering out rows from the facial action unit data from OpenFace. This is done by first dropping unknown columns, then filtering out any csv files (videos) where the detected confidence was less than 85%, mean success value less than 1, and where there is only one face detected. Next, the confidence metrics were dropped from the dataframe. The OpenPose body coordinate data is filtered next. The dataframe is filtered by excluding any csv files where the mean confidence value is less than 0.002. Next, the confidence value column is dropped from the dataframe. The average pose coordinate values are then averaged as a metric of how much pose data was supplied by the video file. Average pose values that are greater than 200, and do not contain a negative coordinate are kept. Columns that are unsuitable for model inclusion such as filename are dropped, resulting in a final dataframe with 788 columns. Finally, each video file was split into 2 second fragments.

2.4 Feature Selection & Cluster Generation

After filtering the data, we used UMAP to reduce the dimensionality of the dataframe containing the dataset due to the large number of columns associated with each element in body pose and facial expressions. To determine how many clusters we need, we used silhouette scores and plotted it against the number of clusters.

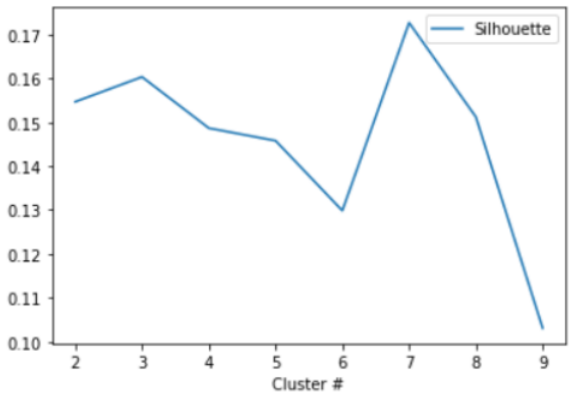

Figure 2: Plot of Silhouette scores.

We determined that 7 clusters would be the most optimal amount. A fixed seed of 9000 was used to ensure run to run consistency. We will explore the attributes of every cluster that was found with a visual representation of the most frame closest to the center.

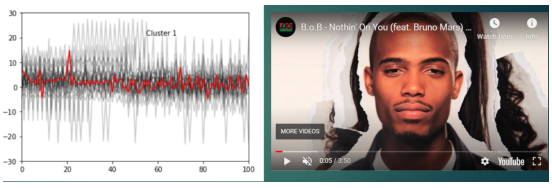

Figure 3: Cluster 1.

This cluster is a calm, slight smile that primarily uses the face. It appears to be a happy and content look.

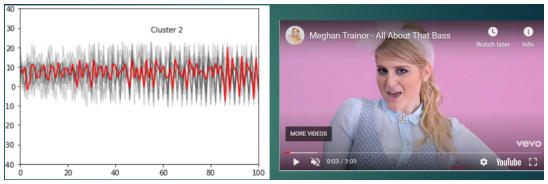

Figure 4: Cluster 2.

This cluster represents an expression that has slightly raised eyebrows, a gaze straight forwards, and the upper torso tilted to the viewer's right.

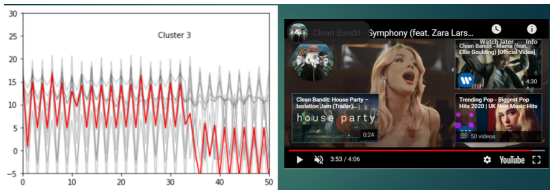

Figure 5: Cluster 3.

This social signal is rather harder to decide. The person is making an yelling expression (probably from singing) and have upper body slightly bend towards. Visually, it is difficult to be certain, but gathering contextual information from the music, it might be an social signal for feeling of longing or desire.

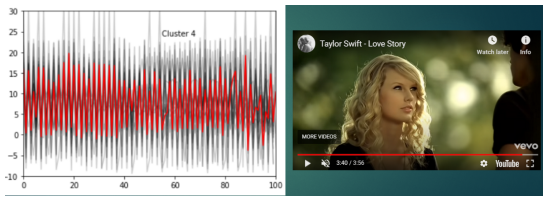

Figure 6: Cluster 4.

This cluster has the gaze looking up and to the left, with raised eyebrows and mouth slightly open. In terms of body pose coordinates, the upper torso is present but is roughly in a neutral position.

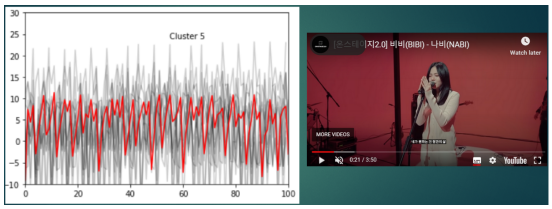

Figure 7: Cluster 5.

This is an social signal for calmness. You can see her eyes closed and almost no expression. The singer's hands are clasped together, like a prayer, which suggest that this is not a sad emotion.

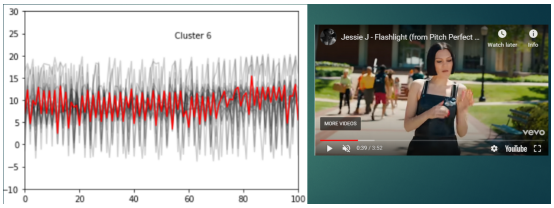

Figure 8: Cluster 6.

This is a social signal for sadness with weak intensity. She is looking down on the ground with no smile with slow body movements that appears with no energy in general.

Figure 9: Cluster 7.

From this video, the emotion is also sadness, but it appears a lot stronger in intensity as there a face with frown. Also, there is a shoulder movement that has a lot more power. More specifically faster movements although the the range in movement is small.

3 EXPERIMENTATION WITH MODELS

In this section we will explore the effectiveness of using different models to generate the emotion classifier. A general overview of the parameters used, training methodology and discussion of effectiveness.

3.1 Support Vector Machine (SVM)

SVM classification choose the hyperplane which maximizes the margin between clusters [7]. As we are dealing more than 700 parameters, if it very difficult to determine the clear boundaries.If there is a clear distinction between the clusters generated previously, it is likely to be a clear separation in spaces as well which make SVM a perfect choice for the classification.

3.2 k-Nearest Neighbours

K-Nearest Neighbours works by given a single point in a space, which of K-numbered nearest points belongs to which class[7]. Same as before, if we have a clear distinction between clusters this makes a good choice as well. Since K-nearest neighbours are relatively fast to predict as it only compares the nearest K-points, it makes it a favorable classification model to choose considering the given time constraint for this project.

3.3 Neural Networks

Using this technique was solely due to curiosity as Neural Network is because becoming more popular. We took this opportunity to explore the technique.

	KNN	Neural Network	SVM
Accuracy	0.70238	0.10714	0.55952
std	0.10310	0.11845	0.26163

Figure 10: K-fold Evaluation Result

From Figure 10, we note that k-NNs were the most effective classifier in the classification methods used. There was less standard deviation in the accuracy of classification, in addition to the general accuracy being higher by a significant margin.

3.4 Annotating the Data

Data was annotated based on the clusters generated by looking at the music videos associated with the most representative frames of the time series data. The annotation was done by the authors of the paper by manually reviewing the cluster data and selecting Representative samples of data from the dataset.

4 DISCUSSION

Using the information presented in section 2, we will analyze the overall effectiveness of our approach to emotion classification with musical pairs.

4.1 Interpreting the Results

Based on our classifier's accuracy scores, we can conclude that the k-NN method is the most effective with an accuracy score of 70% when tested on the validation dataset. 7 clusters were generated from our k-Means clustering approach, and each cluster represents a range of different emotions and subemotions. When used on a 200 frame sample from local webcam data, the k-NN classifier was much less effective, but still yielded consistent results in patterns that are not completely consistent with characteristics of facial-bodily expressions.

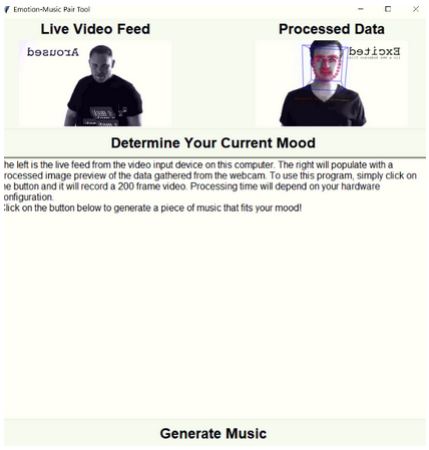

Figure 11: Emotion-Music Pair Classifier Application

4.2 Shortcomings

Classifying emotion was the primary objective of the project, however, the music pairings were the result of the classification. The music used in music videos do not always match up or provide trends that are similar due to the video aspect often having it's own progression of emotion. Choosing the music videos and acting samples presented many challenges in created a dataset that is representative of many emotions and also selected for data that fits the criteria of single person having a clear image of the person to allow for OpenPose and OpenFace to extract the data. During experimentation and preprocessing, we discovered that there was a lack of representation for many different expressions negative emotions such as crying for sadness or shaking for anger. This is potentially due to music videos being the predominant type of video in the dataset. Another area where the dataset may be lacking is the dataset containing only acted types of data. This may result in validation data sourced from the webcam demo being inaccurate for the user's intended portrayal of emotion.

4.3 Future Work and Open Research Questions

In order to generate a model that is more broad, the type of media used should contain a larger spectrum of emotions and contain naturalistic data. A separate music model can be generated which can provide music generation in accordance with the user's emotion. Neural networks are complex and require specific considerations in dataset and parameter tweaking to achieve optimal results. Given the industry's success with neural networks, an focused approach on neural networks may yield results that are more accurate than the ones shown in this paper.

5 CONCLUSION

Using music videos for the majority of the dataset proved to be a challenge in creating accurate models to classify emotions. Music videos contain frames that do not provide usable data due to it's purpose of artistic expression. This also has the effect of having visual flair that impedes OpenPose's and OpenFace's ability to extract data by reducing confidence values and assigning incorrect values for facial action units and body pose coordinates. However, we were still able to accomplish our task of creating a model with a reasonable degree of accuracy, and to create pairs with facial-bodily expressions and music.

6 APPENDIX OF MEDIA SOURCES IN TRAINING DATASET

(A) "train_data - YouTube", Various YouTube Channels. Available: https://www.YouTube.com/playlist?list=PLTdpbHHJbolpn-LCb1mUCtHrFmMhsk5bN

(B) "Best Pop Music Videos - Top Pop Hits Playlist (Updated Weekly 2021) - YouTube", Various YouTube Channels. Available: https://www.YouTube.com/playlist?list=PLMC9KNkIncKtGvr2kFRuXBVmBev6cAJ2u

6.1 Dataset Motivation

The YouTube playlists used in this paper were used to compile the dataset by extracting their facial-bodily expression data to generate our models. The two playlists used in this video were created by Yoonhong Lee (A) and #Redmusic: Best Hits, the first of which was created for the purposes of creating a compilation of suitable data, and the second was created as a list of popular music videos. The first dataset (A) has no funding associated with creating the playlist, and the second (B) is unknown, because it is user-generated content on YouTube.

6.2 Dataset Composition

The dataset is composed of scenes from movies, acting samples, and music videos. There are 22 videos in the first dataset (A) and 250 videos in the second dataset (B). The videos chosen have a roughly even mix of genders, but is primarily media from western, primarily American culture with some exceptions from Korean music videos. Each video file used was not modified beyond its preprocessing in applying body pose coordinates and facial action unit detection. Each video in the playlists were verified to be accessible on YouTube as of August 15, 2020, however due to YouTube's terms of service and the individual content creator's channel policy, the availability of the videos is not guaranteed. The data gathered for this video are converted into csv files, where the only association with the source video is it's filename being equal to the YouTube video ID. Each source video contains frames that do not contain any valid data to extract, which appears as noise to be filtered out when preprocessing the data. None of the videos require access privileges, and are intended for public consumption. The videos gathered do not focus on any subpopulations of people, however specific actors and musicians can be recognized from the video sources.

6.3 Collection Process

The specific process of collecting the data can be found in Section 2: Methodology. The videos making up the dataset were found prior and during the writing of this paper. This dataset has not been reviewed by an ethical review process. Data was collected from publicly available media from movies and music video, however direct consent was not given. Additionally there has been no impact analysis of the dataset as a whole used in this paper.

6.4 Dataset Uses

Thus far, the dataset as a whole has not been used outside training the models used for this paper. This dataset may have alternative uses such as extracting the audio and creating music generating neural networks, for example.

6.5 Dataset Cleaning & Preprocessing

The steps taken to clean and preprocess the data were outlined in Section 2: Methodology.

6.6 Distribution

The dataset itself will not be distributed, however it is accessible any time via YouTube for public use. The videos used will have the same license and copyright attributes that the original videos had in place and the intellectual property of the dataset belongs to the respective YouTube channels.

6.7 Maintenance

Questions about the dataset can be directed at the authors Orion Hsu and Yoonhong Lee. The dataset is in it's final iteration. Older or archived versions of the dataset are not available due to the licenses associated with the YouTube videos.

REFERENCES

[1] A. Heck, A. Chroust, H. White, R. Jubran, and R. S. Bhatt, "Development of body emotion perception in infancy: From discrimination to recognition," vol. 50, pp. 42–51.
[2] D. T. Cordaro, R. Sun, S. Kamble, N. Hodder, M. Monroy, A. Cowen, Y. Bai, and D. Keltner, "The recognition of 18 facial-bodily expressions across nine cultures," Publisher: American Psychological Association.
[3] H. Hwang and D. Matsumoto, "Evidence for the universality of facial expressions of emotion," in *Understanding facial expressions in communication: Cross-cultural and multidisciplinary perspectives.* (M. K. Mandal and A. Awasthi, eds.), pp. 41–56, Springer Science + Business Media.
[4] A. A. Sokolov, S. Krüger, P. Enck, I. Krägeloh-Mann, and M. A. Pavlova, "Gender affects body language reading," vol. 2.
[5] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," vol. 343, pp. 35–55.
[6] V. N. Gudivada, A. Apon, and J. Ding, "Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations," *International Journal on Advanced in Software*, vol. 10, no. 1, p. 21, 2017.
[7] J. Rogel-Salazar, "DATA SCIENCE AND ANALYTICS WITH PYTHON," in *DATA SCIENCE AND ANALYTICS WITH PYTHON*, p. 433.