Emotion and Violence Detection in Movie Clips

Multimedia Search and Recommendation Group 8

Nivedita Prasad
4712099
TU Delft
N.Prasad@student.tudelft.nl

Chia-Lun Yeh
4718836
TU Delft
c.yeh-1@student.tudelft.nl

Fenglu Xu
4579976
TU Delft
F.Xu-1@student.tudelft.nl

Jiahui Li
4734769
TU Delft
J.Li-27@student.tudelft.nl

ABSTRACT

Movies carry a wide range of emotional information, Being able to automatically categorize the emotions in the movies have wide applications. It can not only be used to recommend movies of certain moods to users, but also prevent people from seeing inappropriate scenes. In this report, we analyze how affective class and violence can be detected in movies. We explore visual and audio features that can work with simple models. From initial experiments, we do not observe significant improvements. However, we present more thoughts on future directions.

1 INTRODUCTION

Affective content is information about mood, emotion and feeling that is stimulated when watching a video. The source and content of videos have increased dramatically, making it crucial to automatically categorize the affective content of each video in order to help users search videos that meet their need. We investigated 2 related question:

- Can we predict the extent of induced valence and arousal of a video?
- Can the predicted valence and arousal, together with other information from the video, give good information about whether the video contains violent scenes?

The first research question aims to assist users to search for certain affective content while the second question aims to prevent children from being exposed to harmful video content. The dataset used is the LIRIS-ACCEDE dataset [**?**]. It contains 9800 video clips that last 8 to 12 seconds each. Each clip is labeled with the arousal class, valence class, and the violent or nonviolent labels. Several features from still image, audio, and video have also been pre-extracted, such as color contrast, spectral energy, motion, etc. To answer the research questions, we aim to extract additional features and study whether these features improve predictions. Therefore, the pre-extracted features would be used as the baseline model upon which we compare our new features.

MediaEval'18, 13-15 September 2017, Dublin, Ireland

2 RELATED WORK

2.1 What are valence and arousal?

Human emotions can be modeled using valence, arousal and control, where valence and arousal account for most of the variances in the affective content [6]. Figure 1 shows the 2 dimensional space spanned by valence and arousal. Valence accounts for the positivity or negativity of the emotion while arousal accounts for the extent or how active you want to express it. Although emotions evoked from videos are subjective, the amount and type of affect that is expected to be evoked in the user is largely objective, which is what we try to predict.

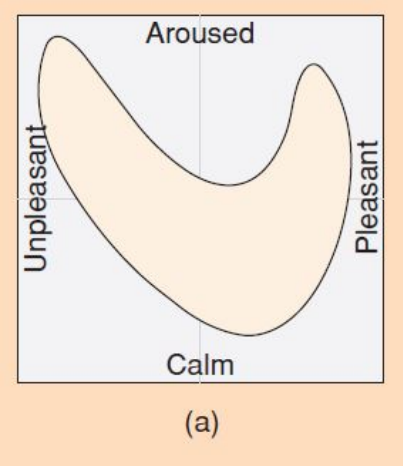

Figure 1: Illustration of the 2D affect space.

2.2 Violence Detection

One early study on violence detection appears in [3]. They use several high-level intuitive features such as blood detector, flame detector, and soundtrack classification with Gaussian models. The work by Eyben et al. aims to optimize the whole pipeline, including segmentation of the frames and extracting features [4]. Giannakopoulos et al. uses a two-level architecture where low level features are later on combined with high-level events.

3 INDUCED AFFECT DETECTION

As mentioned in section 2, emotion can be modeled with valence and arousal. Therefore, we predict the valence (3 classes, negative, neutral, and positive) and arousal (calm, neutral, and excited) for each video. This is a multi-class classification problem. We train three different classifiers, logistic regression (LR), random forest (RF), and nearest neighbor (NN), with the provided features. We normalize all features before training. 2450 excerpts are used as the training set, 2450 excerpts as the validation set to choose the parameters for the models, and the remaining 4900 excerpts as the testing set. The accuracy of the testing set is shown in table 1. The number in the parentheses indicate the parameter chosen using the validation set. We train the final model with this parameter and all data available (training and validation together). Then we test it on the held out test set. Without using more advanced tuning or other machine learning techniques, the classification accuracy is around 40% for the valence prediction, and more than 60% for the arousal prediction.

Classifier	Valence accuracy	Arousal accuracy
Logistic Regression	0.429	0.636
Random Forest	0.403 (20)	0.645 (80)
Nearest Neighbor	0.380 (7)	0.610 (7)

Table 1: The results of valence and arousal class classification.

We also plot the relationship between the arousal and valence class with the violence class in figure 2. The left axis represent valence, where -1 means negative and 1 means positive; while the bottom axis represents arousal, where -1 means calm and 1 means aroused. The value plotted is the percentage of violent samples in each class. We expect violence to occur with negative and aroused emotions. As expected, the combination has the highest proportion of violent labels.

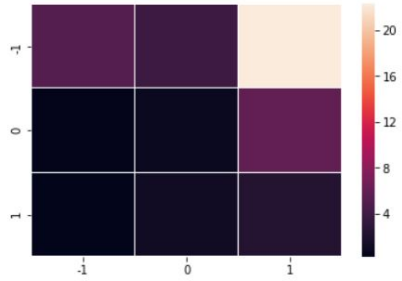

Figure 2: The heatmap shows the correlation of valence, arousal and violence.

4 VIOLENCE DETECTION

The combination of valence and arousal values gives rise to labels like Angry, Happy, Neutral and Sad. The emotions according to the author of [6] is a parabolic curve known as the affect curve that has arousal and valence value plotted on the affect feature space. If the arousal has negative values and valence is also negative then the emotion that can be detected is sad, bored or sleepy but if the valence is positive then the mood predicted would be relaxed, peaceful and calm. On the other hand, if the arousal has positive values and the valence is negative then the mood detected is nervous, angry or annoyed and on the opposite hand if the valence is positive then the mood is Happy, peaceful or calm. The combined values of the A and V contribute to two classes *Violent* which is given a class value as "1" and *Non - Violent* which has a class value of 0. We applied K-nearest neighbors (k = 3) to receive an accuracy of only 68% ; Logistic regression gave an accuracy of 75% and for a black box random forest classifier with 200 trees as limit gave very good classification accuracy of 79%/ as expected. We additionally performed SMOTE process to oversample the minority class and found that accuracy improved to 82% compared to 80% without smoteing.

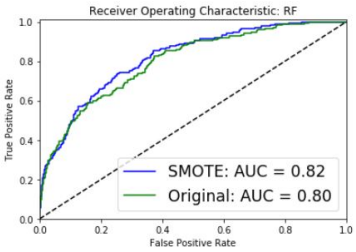

Figure 3: ROC for valence and arousal - Random forest

Hence the classification is an important task to understand the amount violent and non violent videos we will be dealing with. We additionally experimented with only the features of arousal and valence separately to check if the violence and non violence could be classified better without combining the features of both arousal and valence. The results were pretty similar interestingly.

5 APPROACH

Our approach to improve the prediction is to find more discriminative features instead of finding better models. We focus on image and audio analysis.

5.1 Face detection & Emotion analyses

The content of the movie could be well represented by humans' faces expression, as this is the most important semantic information. Human emotions are mental states of feelings that arise spontaneously rather than through conscious effort and are accompanied by physiological changes in facial muscles which implies expressions on face. Some of the critical emotions are happy, sad, anger, disgust, fear, surprise etc.
We first train an emotion classifier on another dataset (fer2013 [2], Fig 4, each image is greyscale image and of size 48*48) and use the trained model to classify our dataset latter.

We then extract faces in every video with a sampling rate of 3 frame per second. Figure 5 is our sample images. From Fig 4 5, we see that the images extracted are quite different from the standard dataset, the video images are often in ill lighting condition.

Figure 4: Sample images from fer2013 dataset

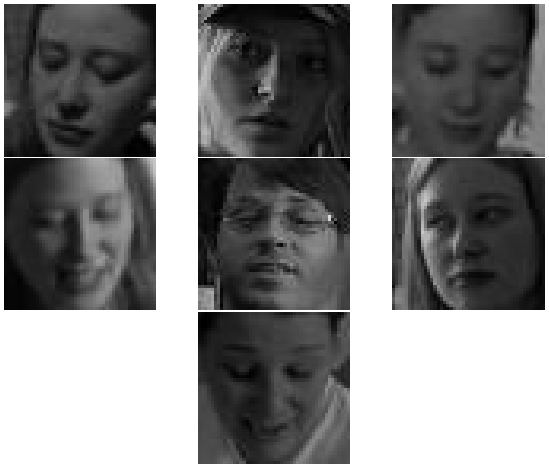

Figure 5: Sample images extracted from videos

5.2 Audio

For audio features, we use the Yaafe audio features extraction [1] to extract some selected features. To simplify the process, we extract one value per feature per video clip. These features include minimum and maximum loudness, minimum and maximum perceptual sharpness, and minimum and maximum perceptual spread. We use a window size of 1024 samples and a 50% overlap between frames.

6 RESULTS AND ANALYSIS

To evaluate the system that we have created to classify emotions and to predict the violence the performance is measured using Mean Square Error and Pearson correlation coefficient for valence and arousal prediction subtask.

[1] http://yaafe.sourceforge.net/

6.1 Emotion classification

For this task, we first split the fer2013 dataset into 8:2 for training and test. We did not perform any feature extraction or other preprocessing method. We use a naive SVM classifier to train on raw dataset ($C = 10, \gamma = 10^{-7}$), this simple method gets **67.82%** accuracy on test set. Fig 6 shows our normalized confusion matrix. But when we test on images extracted from videos, the classifier

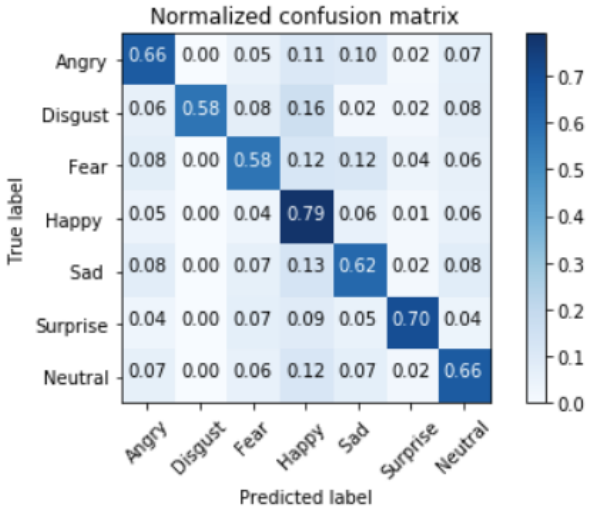

Figure 6: Confusion Matrix for fer2013 dataset

always predict it to be 'Happy' image. This could be explained by these two dataset is too different: extracted images are tend to be darker. Hence the naive liner SVM classifier is not robust to a new dataset, and we should find features that are invariant to light, brightness, scale and rotation.

Due to the poor performance, we do not add this feature to compare with the baseline models.

6.2 Audio features

After adding the features mentioned in section 5.2, we use the same classifiers as section 3 to perform affective class prediction. The accuracy is shown in table 2. We see that using minimum and maximum values computed from the whole clip is not very effective. Some results improve a little, but in general there is no significant improvement. Further experiments are required to decide which features are more useful. Moreover, we think that we need higher-level features, such as analysis of speech or music genre.

Classifier	Valence accuracy	Arousal accuracy
Logistic Regression	0.431	0.636
Random Forest	0.411 (60)	0.646 (60)
Nearest Neighbor	0.372 (7)	0.612 (7)

Table 2: The results of valence and arousal class classification.

7 DISCUSSION & FUTURE

7.1 Audio

Here are some possible future work to distinguish valence and arousal in field of audio.

7.1.1 separate voice and music. As we inspect, lots of audio we extracted from the video are mixture of speech and background music. Some of them are only music or human voice. Currently, the audio analysis without separation of voice and music may loss some important features contained only in voice or music but now they are mixed up together. Thus, to explore the internal correlation between human voice/music and valence/arousal, next step we want to separate voice and background music, possibly by using proper filter or any more advance technique. After the separation of voice and music, we can extract features from them respectively. Or more further, we can do speech recognition in the humane voice and music genre detection, which may help to tell valence and arousal.

7.1.2 Image classification on MFCC spectrogram. : Since currently the technique of auditory perception has not quite caught up to computer vision, we can transform audio problem into image classification. For example, we may find the spectrogram of MFCC per audio and then do classification. To train the model, we can take an advantage pretrained CNN model such as Inception which is trained on Imagenet and released by Google. And then we retrain the last few layers of it for our new categories(valence and arousal), which is so-called transfer learning.

7.1.3 Wavenet and transfer learning. : We can also use WaveNets (a deep generative model of raw audio waveforms released by DeepMind) as pretrained model and retrain it for our new categories.

7.2 Emotion classification

There are basically 2 ways to improve our emotion classification method: preprocessing and invariant feature extraction.

7.2.1 Pre-processing. Expression representation can be sensitive to translation, scaling, and rotation of the head in an image. Hence we could perform geometric normalization of the head. The aim of the pre-processing phase is to obtain images which have normalized intensity, uniform size and shape, and depict only a face expressing certain emotion. The pre-processing procedure should also eliminate the effects of illumination and lighting.

7.2.2 Convolutional Neural Network. When we talk about large scale image classification, the first method that comes up will be Deep Learning and Convolutional Neural Network (CNN), this method is very powerful that it does not need any prior information, and not need to design exact filters as it can be learned automatically. Then we can directly feed the extracted CNN features into a classifier (like SVM).

7.2.3 Geometric Features Extraction. As in [1], our facial features are highly structured, different emotions exhibit different geometric features, such as the shape of mouth, eyes and orientation of eyebrow as shown in Fig 7. There are 19 features are extracted from the face image, which is implemented by segment the face image into three regions: mouth, nose and two eyes and two eyebrows and then located the facial characteristic points (FCPs) in each face. Finally, the distance between FCPs are calculated using Euclidean distance. Instead of preprocessing the images to make them in the same form,feature extraction method can extract certain features that are robust to different conditions. This process could be viewed as dimension reduction.

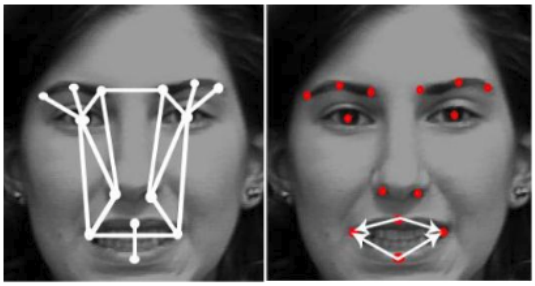

Figure 7: Geometric features: (a) geometric lengths, (b) mouth angles

REFERENCES

[1] Ali K. K. Bermani, Atef Z. Ghalwash and Aliaa A. A. Youssif, âĂIJAutomatic Facial Expression Recognition Based on Hybrid ApproachâĂİ International Journal of Advanced Computer Science and Applications(IJACSA), 3(11), 2012.

[2] Goodfellow, Ian J., et al. "Challenges in representation learning: A report on three machine learning contests." International Conference on Neural Information Processing. Springer, Berlin, Heidelberg, 2013.

[3] Nam J, Alghoniemy M, Tewfik AH (1998) Audio-visual content-based violent scene characterization. In: Proc. of 1998 International Conference on Image Processing (ICIP) 1998. IEEE, volume 1, pp. 353âĂŞ357.

[4] Florian Eyben, Felix Weninger, Nicolas Lehment, BjÃűrn Schuller, Gerhard Rigoll. Affective Video Retrieval: Violence Detection in Hollywood Movies by Large-Scale Segmental Feature Extraction.(2013) https://doi.org/10.1371/journal.pone.0078506

[5] Giannakopoulos T, Makris A, Kosmopoulos D, Perantonis S, Theodoridis S (2010) Audio-Visual Fusion for Detecting Violent Scenes in Videos. In: Konstantopoulos S, Perantonis S, Karkaletsis V, Spyropoulos C, Vouros G, editors, Artificial Intelligence: Theories, Models and Applications, Springer Berlin/Heidelberg, volume 6040 of Lecture Notes in Computer Science. pp. 91âĂŞ100.

[6] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv", IEEE Signal Processing Magazine, vol. 23, no. 2, pp. 90-100, March 2006.