A MULTIMODAL APPROACH FOR PREDICTING THE 
INDUCED AFFECT CONTENT IN VIDEOS

ANKIT GOYAL 

12120 

ASHUTOSH KUMAR 

11168 

GROUP NO.- 22 

Project Mentor 

DR. VINAY P. NAMBOODIRI 

Course Instructor 

DR. AMITABHA MUKHERJEE 

Date of Submission: April 15th, 2016
1 INTRODUCTION

Affective computing is an burgeoning field of research, with researchers trying to predict the expressed and/or induced emotions from various multimedia data like music, movies, speech among others. According to Wikipedia,”Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. It is an interdisciplinary field spanning computer science, psychology, and cognitive science.” Such emotion recognition systems are very useful like they can be used for building mood driven video recommendation systems, video summarization systems etc. In some sense, this can be used to make the machines to feel empathy bringing them closer to humans.

We have focused our attention to find the induced affect content in videos. The induced affect content can be the amount of excitation that a video induced in a viewer. According to standard emotion literature, any emotion can be broken down into three orthogonal axes i.e. affect, valence and tension, out of them first two are often considered to be more important. The affect axis describes the amount of excitation/agitation, which can be there for both positive as well as negative emotions. We have used regression analysis on a variety of audio-video features extracted from the samples to predict their affect content.

We have build on the work done by Ankit et al Goyal et al. (2016). We have used the audio and video features for emotion prediction suggested by Ankit et al. Along with their features, we have also extracted CNN features. We used various regression models for doing the affect prediction. To conduct out experiments we have used the dataset proposed in Baveye et al. (2015). We have used a reduced version of this dataset which has around 3500 video clips with arousal and valence ratings.

2 FEATURES

As mentioned earlier we used the features proposed by Ankit et al for emotion prediction. Along with these hand crafted features, we also used CNN features. Most of the work of feature extraction was done in Matlab with help of open source libraries. The following sections describe the features we have extracted.

2.1 Audio Features

2.1.1 MFCC

The Mel-frequency cepstrum (MFC) represents the power spectrum of a sound. It is based on a linear cosine transfer of a log power spectrum on a nonlinear mel scale of frequency. MFCC are the coefficients that make up the MFC. MFCC long with Chroma has been widely used in emotion recognition models. We extracted MFCC feature of each of the video at an interval of 25 ms with a window of 10ms. The derivatives of MFC and Chroma wrt time were also extracted. To sum it all, (mean, min, max and variance) of all the features were calculated.

2.1.2 Audio Compressibility

Data encoding is an integral part of the signal processing. It helps reduce transmission bandwidth and resource usage and if the compression is lossless there is no loss of data too. The Audio compressibility feature has been shown to be highly correlated to human emotion and by extension to valence and arousal values. Audio Compressibility feature can be defined as the ratio of the size of losslessly compressed audio clip to raw audio
clip. We have compressed the raw audio using FLAC lossless codec, ffmpeg and used the ratio of size of compressed audio clip to that of the original as a feature.

2.1.3 Harmonicity

According to Srinivasan et. al, "When a body vibrates with a repetitive period, its vibration give rise to an acoustic pattern in which the frequency components are multiples of a common fundamental. This behavior is termed Harmonicity." The different harmonics are known to simulate different emotional responses from humans, that what makes it an excellent feature to be used. Harmonicity for the sample is taken as the ratio of clips that have a pitch to the total number of clips.

2.2 Video Features

2.2.1 Shot Frequency

Shot frequency is the frequency at which the shots, frames recorded continuously, are changed in the video. It is one of the critical factors used by cinematographers to set in the mood of the scene. A lot of cuts put the action to fast paced with the protagonist facing some imminent peril while a long duration shot is used for peaceful and calm scenes. To make use of this we count the total number of shots in the video using ffmpeg.

2.2.2 Histogram of Optical Flow (HOF)

The amount of activity on the scene has an impact on the emotion of the viewer. This can be captured using Histogram of Optical Flow. The Lukas-Kanade Optical Flow for all frames except the ones near a shot boundary are calculated. We didn’t do so for the frames near a shot boundary as due to the discontinuity they would exhibit a error-prone high optical flow. A 8-bin histogram for the each frame was made. For each \([x;y]\), optical flow vector, in a frame, its angle with the x axis can be computed as \(tan^{-1}\left(\frac{x}{y}\right)\). The \(i^{th}\) bin contains angles ranging from \((\frac{i-1}{4}\phi\) to \((\frac{i}{4}\phi\). The size of each bin is calculated by using the L2 norm of each of the optical flow vectors inside id, i.e. \(\sqrt{x^2+y^2}\). A complete statistical analysis was done later.

2.2.3 Video Compressibility

We characterize a video compressibility highlight to catch parts of movement and change in a video. Video data redundancy over time is exploited by these video compression algorithms, utilizing movement and change prediction. We utilize video compressibility as a reduced component to consolidate the impacts of such variables over a clasp, because they are correlated with the apparent emotional ratings, providing us another good measure.

2.2.4 3d Hue Saturation Value (HSV)

Shading, hue and color play a important role stimulating the human feelings. A lot of work has been done showcasing the correlation between color and behavior of people in reaction to them. This data, hue, color and saturation, is extracted utilizing the 3d HSV feature. The frames are changed from RGB to HSV color space and the resulting values are put in one of the 64 bins corresponding to 4 range of values for hue, saturation and value each. At the end, we do a statistical analysis across all the frames.

2.2.5 Histogram of Facial Area (HFA)

Facial expressions in a close-up shots are very useful in making an impact or delivering the solemnness or joy of the moment. We endeavor to use this data utilizing the HFA highlight. We did face location in every one of the
frame utilizing a Deep Convolutional Network based face locator which are further binned to build a histogram. We develop a 3 bin histogram, with the bins speaking to little, medium and large sized faces.

3 CONVOLUTIONAL NEURAL NETWORK FEATURES

Convolutional Neural Networks are feed forward deep learning architecture mimicking a very simplified version of the neurons in the brain. The architecture used in the project consists five convolutional layers, followed by two fully connected layers (fc6 and fc7) and a final output layer. Yadav (2015)

3.1 Application of Convolutional Neural Networks

An another interesting way to look at the problem to predict the induced affect content can be by using Convolutional Neural Networks. Some considerable work over the same has been done by Srinivas et al. (2016) and Donahue et al. (2013). We take 6 screen-shots from each video at an interval of 2 seconds, thus spanning 12 seconds and then use these images to get the CNN feature vectors. First the frames are re-sized to a 256x256 fixed size by doing bi linear interpolation. The CNN features are then extracted using Caffe. Caffe is an open framework with models and worked out examples for deep learning developed by Berkley Vision an Learning Center Jia et al. (2014). Next, we get output for each frame entered from the seventh layer (fc7) of the CNN model in the form of a 4096 dimensional vector. So now we have a 6 such vectors for each image. Thereafter, we have used the mean of the six to perform further linear regression.

4 SYSTEMS FOR PREDICTION

After extracting feature from the video clips, we try various regression models in order to do the arousal and valence prediction. We learn distinctive regression models which attempt to anticipate the arousal and valence values. We have divided the data into train and test set with 3500 train and 414 test sets.

4.1 Only Audio, Only Video, Early Fusion Model

First we attempt to learn a regression model using just the audio features or video features only. Other regression models like Support Vector Regression (SVR) and Gaussian Process Regression (GPR) were also tried, however there was very little change in the forecast. In the early fusion model, the audio and the video features were joined to learn a regression model.

4.2 Late Fusion Model

This is basically the combination of the two independent models, one from audio only and another from video only. If \( y(v) \) is the prediction from the video features and \( y(a) \) is that from the audio features, then final prediction \( y(pre) \) is given by Equation(1). The value of \( \alpha \) is chosen to maximize the correlation of actual to predicted values and remains same across all samples.

\[
y(pre) = \alpha y(v) + (1 - \alpha)y(a), \alpha \in [0, 1]
\]
4.3 Mixture of Experts Based Fusion Model

It is similar to the late fusion model, but here the two independent models, one each from audio and video, are gated to produce the final model prediction. We also learn the parameter, which is not constant across samples, for this gating function. The last forecast for the MoE-based model is fundamentally the same to the Late Fusion model with the exception of the way that here we don’t have a constant $\alpha$ across all samples. The estimation of $\alpha$ relies on upon the sound and video highlights of the present specimen. Hence, it can be though of as trying to predict an additional value for each sample which gives the weight for the video and audio feature of that sample. Here we use pre-trained models that have been trained on a humongous amounts of images and hence the features extracted with their help prove to be very helpful.

5 EVALUATION AND RESULTS

To evaluate our models we did a 5-fold cross validation on the dataset. We predict the arousal and valence values for new unseen clips using the trained model that we make. We report the mean absolute error in arousal and valence across all the five folds. The following table summaries our results.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE Aro</th>
<th>MSE Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR Audio</td>
<td>0.811</td>
<td>0.519</td>
</tr>
<tr>
<td>SVR Audio</td>
<td>0.803</td>
<td>0.493</td>
</tr>
<tr>
<td>GPP Audio</td>
<td>0.812</td>
<td>0.514</td>
</tr>
<tr>
<td>LR Video</td>
<td>0.817</td>
<td>0.507</td>
</tr>
<tr>
<td>SVR Video</td>
<td>0.816</td>
<td>0.469</td>
</tr>
<tr>
<td>GPR Video</td>
<td>0.814</td>
<td>0.479</td>
</tr>
<tr>
<td>LR Early Fusion</td>
<td>0.790</td>
<td>0.496</td>
</tr>
<tr>
<td>SVR Early Fusion</td>
<td>0.791</td>
<td>0.465</td>
</tr>
<tr>
<td>GPR Early Fusion</td>
<td>0.801</td>
<td>0.476</td>
</tr>
<tr>
<td>LR Late Fusion</td>
<td>0.802</td>
<td>0.502</td>
</tr>
<tr>
<td>SVR Late Fusion</td>
<td>0.787</td>
<td>0.462</td>
</tr>
<tr>
<td>GPR Late Fusion</td>
<td>0.803</td>
<td>0.479</td>
</tr>
<tr>
<td>MoE</td>
<td>0.791</td>
<td>0.478</td>
</tr>
<tr>
<td>SVR CNN</td>
<td>0.85</td>
<td>0.53</td>
</tr>
</tbody>
</table>

6 CONCLUSION

In this work we try to predict the arousal and valence value for a video clip. We extract various features, suggested in the literature, which are suitable for task. With the help of these features, we try to model the arousal and valence values using various machine learning algorithms. In particular, we find that fusion models perform slightly better than audio only or video only models which indicate that audio and video have contain complementary information, which can help in emotion prediction. Among the linear models, the mixture of experts model which dynamically models the contribution of audio and video performs the best.
REFERENCES


