2014 IEEE International Symposium on Multimedia

# Predicting Evoked Emotions in Video

Joseph G. Ellis*, W. Sabrina Lin†, Ching-Yung Lin†, Shih-Fu Chang*
*Dept. of Electrical Engineering, Columbia University, New York, New York
†Network Science and Big Data Analytics, IBM Research, Yorktown Heights, New York

***Abstract*—Understanding how human emotion is evoked from visual content is a task that we as people do every day, but machines have not yet mastered. In this work we address the problem of predicting the intended evoked emotion at given points within movie trailers. Movie Trailers are carefully curated to elicit distinct and specific emotional responses from viewers, and are therefore well-suited for emotion prediction. However, current emotion recognition systems struggle to bridge the "affective gap", which refers to the difficulty in modeling high-level human emotions with low-level audio and visual features. To address this problem, we propose a mid-level concept feature, which is based on detectable movie shot concepts which we believe to be tied closely to emotions. Examples of these concepts are "Fight", "Rock Music", and "Kiss". We also create 2 datasets, the first with shot-level concept annotations for learning our concept detectors, and a separate, second dataset with emotion annotations taken throughout the trailers using the two dimensional arousal and valence model for emotion annotation. We report the performance of our concept detectors, and show that by using the output of these detectors as a mid-level representation for the movie shots we are able to more accurately predict the evoked emotion throughout a trailer than by using low-level features.**

***Keywords*-multimedia; affective computing; video processing; multimodal; emotion analysis; computer vision; audio processing; signal processing; movie analysis; movie**

## I. Introduction

Understanding the evoked emotions that are induced by a particular multimedia document (text, audio, video) has always been an extremely difficult task for machines to perform. Industry and academia have made great advances in sentiment and emotion analysis, but much of the work has been targeted towards the text domain. Many exciting applications have been realized, including obtaining public opinion towards particular products from Twitter posts, and predicting the outcome of elections from the vast amount of text data available on-line in recent years. While advances have been made in the processing and analysis of text, videos also can portray strong emotion. The amount of videos on-line and available has been increasing greatly and should continue to do so. However, due to the computational complexity for processing and difficulty in analyzing video content they have been largely ignored for emotion analysis. Recently, advances have been made in detecting a person's emotional reaction to a video based on facial gestures and auditory signals. However, predicting the emotional reaction based on the video itself has remained a largely unsolved problem.

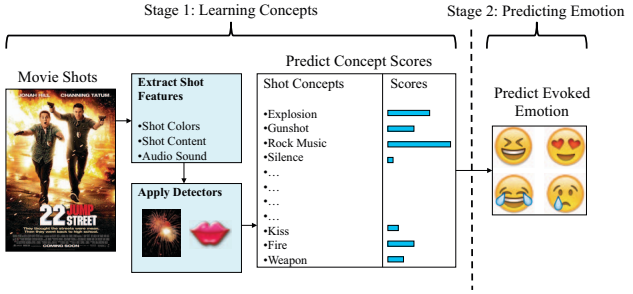


Figure 1. Proposed Emotion Analysis System. The entire emotion analysis system can be thought of in 2 stages. First, we learn the concepts that appear in each shot within the movie. Second, we predict the emotional content of sections of the movie or movie trailer.

Affective Computing is the research field that addresses building systems that have the capability to understand human emotions. Much of the difficulty that appears in affective computing, has been attributed to what is known as the "affective gap". Which refers to the disconnect between low-level visual and audio features and high-level affective concepts such as human emotions. To help bridge the affective gap we propose to learn a mid-level feature representation for multimedia content that is grounded in machine detectable "concepts", and then model human emotions based on this representation. This approach presents particular advantages as compared to current state-of-the-art practices. First, our concepts are concrete detectable items, which have distinct visual and audio properties, and can therefore be assumed to be reliably detectable. For example, while two or more scenes that are associated with the same emotion, such as "frightening", can look and sound very different, our concepts such as "gunshot" should have consistent audio and visual appearance within video. Therefore, we argue that it is more reasonable to build a gunshot detector and then link that gunshot to an evoked emotion, such as frightening, rather than detecting that the

978-0-7695-5437-2/14 $31.00 © 2014 IEEE
DOI 10.1109/ISM.2014.69

scene is frightening given the low-level features. Secondly, Our mid-level representation also models the scene with a much lower dimensional feature, allowing for storage and processing gains over traditional low-level feature based methods.

In this work we focus on predicting the evoked emotion at particular times within a movie trailer. Desired emotions play an important role in determining the type of content that is most appealing to a given consumer. This holds particularly true for determining the movie that a movie-goer would most prefer. For example, a person looking for a funny and pleasant experience may choose to see a comedy movie such as "Caddyshack" or "The Hangover". Whereas a movie-goer searching for an intense and sometimes frightening experience would tend to choose movies such as "The Purge" or "The Conjuring". Movies are carefully crafted to elicit particular emotional responses from the viewers, and emotions in movies are typically less ambiguous than other video domains such as television or social videos. In particular, movie trailers are heavily crafted to elicit a very specific emotional response from viewers, and can contain many different desired emotional responses in a very short amount of time. Entire movies will experience lulls in which the desired evoked emotion is neutral. Movie trailers do not generally exhibit this trait. Therefore, in this work we target predicting the emotion within movie trailers under the assumption that the intended evoked emotion will be more obvious and consistent across viewers.

The organization of the rest of the paper is as follows. Section 2 describes the related work that has been performed in this domain in the past. Section 3 presents an overview of our full video emotion analysis pipeline. Section 4 details how we learn our mid-level concept representation. Finally, Section 5 explains how we use the mid-level feature to predict the intended evoked emotion at particular points within the movie trailers. In Section 6 we show the results of our emotion prediction pipeline, and compare it to other low-level feature based methods. We conclude the work in Section 7, and then detail future work we hope to explore in Section 8.

The contributions of this work are as follows:

- A novel mid-level feature approach to predicting emotion analysis in video.
- A framework and website for collecting shot-level concept annotations, and "section" level emotion annotations from videos.
- An analysis of the performance of our mid-level feature in predicting evoked emotions in video as compared to low-level feature approaches, where we show that our mid-level feature significantly outperforms the low-level approaches in both speed and performance.

## II. Related Work

Recently, much work has been completed in the field of text sentiment and emotion analysis. A thorough overview of the state of the art technologies and work can be seen in [1]. Less work has been completed on evoked emotions from videos, and the field is less mature than text-based sentiment and emotion prediction. A related field, in which progress has been made in the past decade is the analysis of human emotion given videos of their reactions [2]–[4], but this is not the problem addressed in our work.

One of the pioneering works in this area was [5], which first addressed emotional analysis of video content. [6] built off of the original work in [5], and applied similar techniques to videos and pictures creating an affective analysis system for creation of a personalized television platform. Some recent work has appeared that takes art theory into account when predicting the evoked emotions within images and has shown promising results [7]. Very recently, work has investigated how cinematographic theory can enhance feature extraction algorithms for the use of emotional understanding within movies [8]. There is also a large amount of recent work on using biological signals, such as EEG, heart rate, gaze distance, etc. to attempt to understand evoked emotion within videos [9], [10]. Another interesting study can be found in [11]. The authors proposed to map the raw video features to a connotative space they defined before attempting to predict viewer emotion. [12], showed that object and generic higher level features, such as [13], [14] are effective for predicting the emotion in user-generated videos. Our approach differs in that we have specifically tailored a mid-level concept feature for use in movie trailers, instead of using already implemented generic mid-level feature detectors. The authors of [15] presented a system and algorithm for bridging the affective gap in movie scenes using audio-visual cues, in particular they used low-level audio features to create an Audio Scene Affect Vector (SAV). This vector represents the probability that 7 different emotion labels would be applied to a scene given only the audio components. The authors then combined this vector with the low-level features extracted from video to predict emotion in the video. Our work differs from [15] in that our concepts are not emotionally defined, but are instead machine detectable semantic concepts. We define their visual features as low-level, because they do not have semantic meaning. Some examples of their extracted visual features are shot duration, visual excitement based on motion, and lighting.

Our work in large part is inspired by [14], where the authors used an ontology to predict sentiment of flickr images. We take a similar approach in our work by using a mid-level concept based ontology for movie classification, but make use of all of the multi-modal data available in videos that is not available in static images. Our work is


Figure 2. Screenshot of the Concept Annotation Website.

differentiated from other related works, because we have created a mid-level concept based emotion prediction system in videos that is tailored with emotionally-related concepts specific to the video domain, and utilize multimodal features to build our concept detectors. To the best of our knowledge this is the first work that shows the usefulness of this type of mid-level representation for predicting emotions within videos.

## III. System Overview

In this section we provide a brief overview of the entire system, and discuss how our methodology for emotion classification is different than existing works. A pictorial view of our entire system pipeline from raw video to emotion prediction can be seen in Figure 1.

Our pipeline can be split into two separate stages, and this separation is seen in Figure 1. The first stage of the pipeline is concerned with learning and predicting concepts from the shots within each movie trailer. We extract low-level audio and visual features from the movie shots, and then build concept detectors based on these low-level shot features. The scores from the concept detectors within this stage of the pipeline are used as the mid-level feature for the movie shots. In the second stage we use the concept detector scores as our mid-level representation to predict the emotion within particular portions of the movie trailers.

Emotion Detection systems generally utilize a similar pipeline, but instead of using the mid-level representation for the shots others learn a model from the low-level extracted features directly to emotions. The results of our mid-level feature approach as compared to the traditional low-level approach are detailed in Section 6.

## IV. Learning Concepts

We propose a mid-level feature based on concepts that evoke strong human emotional responses. We believe that concepts are tied more closely to human emotional responses than low-level features, and therefore our mid-level feature will be able to model evoked emotions better than low-level features.

### A. Concept Annotations

To train our concept detectors we have created a dataset of movie trailer shots, annotated for which concepts appear within them. Shots are defined as visually consistent sections of video. We chose to annotate our concepts on the shot level, because this eliminates the ambiguous case for annotators in which a concept may appear on screen then disappear all in one annotation unit. We segment the videos into shots using a commonly utilized shot-detection framework,[1] which utilizes change of colors and motion between successive frames to detect shot-changes. We defined 36 different concepts, which appear frequently within movie trailers and in our opinion have a high-level of emotional importance. Among the 36 concepts that we defined we only received more than 25 annotations for 23 of the concept classes, which we deemed sufficient to build reliable concept detectors. A full list of the concepts can be seen below, and the italicized concepts are the 23 concepts for which we have sufficient annotations to build reliable detectors.

- *gunshot*, *explosion*, *speaking*, *screaming*, sex, kissing, *car*, meeting, *weapon*, animals, *rock music*, *jazz music*, boat, *inside car*, *up-close face/body*, *fight*, *fly-over*, *night scene*, *fire*, crying, *dialogue*, *supernatural*, blues music, *chorus*, *city*, neighborhood/school, *natural scene*, *robot/machinery*, police, beach/pool, slow-motion, *daily life*, killing, beeping, drumming, *silence*

To collect these annotations we have built a website for movie shot trailer annotation. An image of this website is shown in Figure 2. The annotator is presented with the 36 concepts below, and a clip of the particular movie shot plays automatically. The annotator chooses from up to 3 concepts which he/she believes describe the movie shot the best, and then moves on to the next shot in this particular trailer. Each movie trailer can exhibit anywhere from 30-130 shots. If the annotator does not believe that a shot is adequately described by the available concepts they can choose to skip this shot, and it will not be used in training or testing of the detectors.

Using this website we were able to collect 3,018 movie shots. These shot annotations were obtained from 37 unique

[1] http://johmathe.name/shotdetect.html

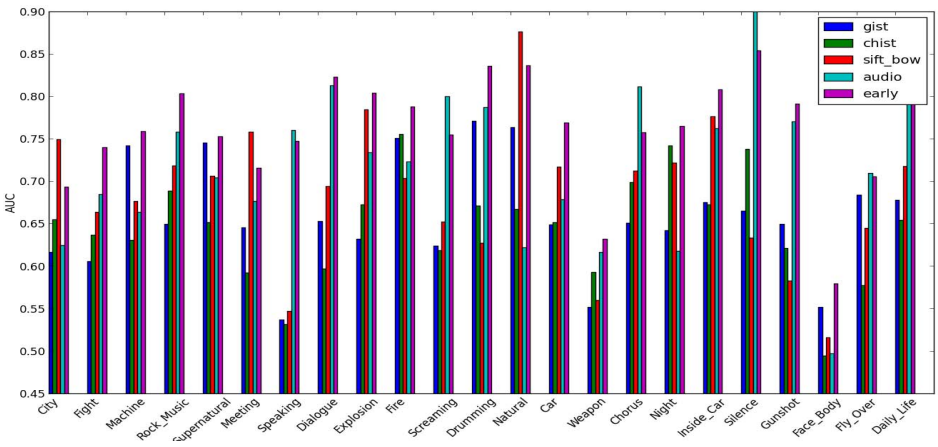


Figure 3. Concept Detector Performance. This figure shows the performance of each concept detector over four-fold cross validation. We can also see the performance of each feature per concept. The early fusion feature is used to build our final concept detectors, and the output of these detectors will be used as the mid-level feature.

movies trailers. From these 3,018 trailer shots we collected 23 different concepts with 25 or more annotations which we deemed a sufficient amount of positive examples to train concept detectors.

### *B. Shot Feature Extraction*

We extract low-level audio and visual features from each of the shots and use them for building the concept detectors. To extract the audio features from the shots we utilized the popular audio feature extraction tool, openSMILE [16]. openSMILE extracts a variety of low-level audio features (Mel Frequency Cepstral Coefficients, pitch, noise energy, and more) from 10 millisecond windows within the audio track, and then applies a variety of functionals (mean, standard deviation, etc.) on the values of the features across the entire shot. The openSMILE audio feature set is specifically tuned for emotion recognition in audio, and more detailed explanation of the system can be seen in [16].

We model our visual feature extraction based on other popular object detectors, such as [17] and [14]. We extract a keyframe from each of the shots, by choosing the frame that appears in the middle of the shot time frame. The visual features extracted from this shot keyframe are used as the visual representation for the shot. We represent each shot by concatenating these feature descriptors into one feature vector: GIST [18], Color Histogram, SIFT [19] Bag of Words (BoW), number of faces that appear, and average hue and saturation of the image. The GIST feature is a holistic image descriptor generally used for scene recognition. Color histograms are extracted over all 3 primary colors within the image (RGB), and convey color information about the scene. We use SIFT BoW feature with a 1000-dim codebook to encode local features within the images. Finally, the number of faces that appear and the average hue and saturation of the images are extracted, and have been shown in the literature to be related to emotion. When all of the audio and visual features are concatenated together we arrive at our 3717 dimensional early-fusion shot feature. We extract a wide variety of different audio and visual features because some features are most discriminative for detecting particular concepts. For example, "silence" is detected most accurately using the audio features that are extracted, where as a "natural scene" is detected using visual features, in particular SIFT BoW.

### *C. Concept Detector Performance*

We trained Support Vector Machines using libSVM to detect our concepts [20]. We used polynomial kernel with degree up to 3 for each of the concept detectors. Four-fold cross validation was used, and we optimized the area under the curve (AUC) metric to set the parameters for each concept SVM. We chose to use AUC as the optimization metric, because we will be using the output scores of the concept detectors as our feature and not the binary yes or no detection value. AUC takes into account the output score of the detector. During training we used 3 times as many negative examples as positive, and the negative examples were sampled randomly from shots that were annotated with other concept labels, and not the target concept.

The results of the average AUC across all four folds for each concept detector can be seen in Figure 3. The output of the SVMs using the early-fusion feature is used as our mid-level feature. In Figure 3, we also present the performance using only particular features to see what features are the most useful in detecting each particular concept. AUC for random guess is 0.5, and perfect recall and precision for each of the concept detectors would result in an AUC of 1.

In Figure 3 we have some concepts that are difficult to

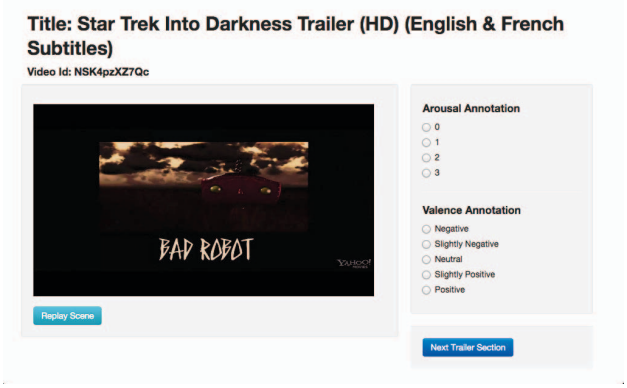


Figure 4. Emotion Annotation Website. This is a screen capture of the website that was used to collect movie trailer section emotion annotations from annotators.

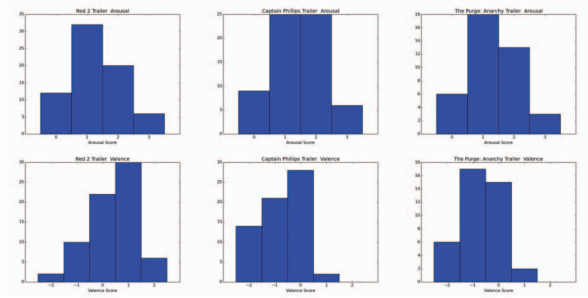

Figure 5. Arousal and Valence Annotation Histograms. This figure shows the histogram of all annotation scores for arousal and valence that we received from the annotators from 3 movies: Red 2, Captain Phillips, and The Purge.

detect such as "weapon". "Weapon" is a highly variable class and often takes up only a very small portion of the screen. We also have some concept detectors that perform very reliably, such as "natural" and "silence". It's also interesting to note which features perform well for each particular concept, and the useful features agree with our intuition. Finally, The numerical score from each concept detector are concatenated together to generate our mid-level feature for each movie trailer shot.

## V. Emotion Prediction

Now that we have generated a mid-level feature for representing movie shots, we will use this feature to predict the evoked emotion within particular portions of the video. First, to develop our emotion prediction framework we have created an annotation website for collecting ground-truth evoked emotion, using the two-dimensional arousal and valence emotion model, detailed in Sec.V-A. Once the annotations are gathered we build a regression model to predict the evoked emotion from each portion of the video.

### A. Emotion Model

For our emotion model we adopt the popular two-dimensional arousal and valence model that has been discussed and utilized [21]–[23]. This model maps human emotions onto a three-dimensional plane, where the orthogonal directions within the plane are defined as "arousal", "valence", and "dominance". Most of the emotional information is captured in just the arousal and valence dimensions of the model, and therefore it is common to simply ignore the dominance dimension when attempting to computationally model human emotions. Arousal corresponds to the intensity of reaction to a given stimuli, and valence corresponds to the positive or negative sentiment in reaction to a stimulus. For example, emotions like excitement will have a high arousal and valence score, where as contentment will have a lower arousal score but high valence. Conversely, boredom will have a very low arousal score and a medium valence value, and scared would have low valence and high arousal values. A detailed description of the emotional model that we used can be found in [5].

### B. Emotion Annotations

We collect emotion annotations on 7 different movie trailers, using the arousal-valence emotion model. The movie trailers used are The Purge, The Conjuring, Riddick, Red 2, Iron Man 3, Captain Phillips, and The Hunger Games. We separate each of the trailers into "sections" to collect the annotations. We define a movie section as a portion of the movie that exhibits a consistent and coherent evoked emotion. We manually cut the movie trailers into sections, based on the definition above. The trailers contain approximately 12 sections each, and the sections last anywhere from 3-30 seconds. We chose to manually create the movie sections, because we wanted the annotators task to be as simple as possible with consistent emotion throughout the section.

We created a website for the collection of emotion annotations. The viewer was shown a movie section, and then was asked to rate their evoked emotion from the given content in the arousal and valence scale. The arousal was annotated on a scale of zero to three, and the valence was annotated on a scale of negative two to two. For both annotation dimensions only integer annotations were allowed. A screen-shot of the emotion annotation website can be seen in Figure 4.

Using this website we were able to collect 511 different annotations from 6 annotators. Each movie section was annotated by 4-6 annotators. The average absolute deviation from the mean across the annotators within the dataset for the arousal and valence annotations was $\mathbf{0.55}$ and $\mathbf{0.57}$ respectively. This shows that there was a reasonably high level of agreement between the annotators on the arousal and valence that was evoked by a given movie section.

Figure 5 shows histograms of all of the annotations gathered for arousal and valence of the sections within 3 distinct trailers: Red 2, The Purge, and Captain Phillips.

Table I
AROUSAL PREDICTION PERFORMANCE.

| | Concept Feature | low-level (PCA) | low-level (PCA) | low-level (Raw) |
|---|---|---|---|---|
| Feat. Dim. | 26 | 26 | 103 | 3720 |
| The Purge | **0.45** | 0.66 | 0.58 | 0.52 |
| The Conjuring | **0.69** | 1.08 | 0.74 | 1.29 |
| Riddick | **0.32** | 0.56 | 0.61 | 1.16 |
| Red 2 | **0.25** | 1.01 | 0.69 | 2.38 |
| Iron Man 3 | **0.51** | 0.65 | 0.74 | 1.36 |
| Capt. Phillips | **0.51** | 0.70 | 0.87 | 1.20 |
| Hunger Games | **0.25** | 0.35 | 0.35 | 1.54 |
| **Overall** | **0.43** | 0.71 | 0.65 | 1.35 |

Table II
VALENCE PREDICTION PERFORMANCE.

| | Concept Feature | low-level (PCA) | low-level (PCA) | low-level (Raw) |
|---|---|---|---|---|
| Feat. Dim. | 26 | 26 | 103 | 3720 |
| The Purge | 0.60 | 0.92 | 0.80 | **0.44** |
| The Conjuring | 0.84 | 1.38 | **0.48** | 0.78 |
| Riddick | **0.23** | 0.56 | 0.58 | 0.57 |
| Red 2 | **0.80** | 2.12 | 1.42 | 2.29 |
| Iron Man 3 | **0.45** | 0.85 | 0.63 | 0.91 |
| Capt. Phillips | 0.78 | 1.18 | **0.60** | 0.97 |
| Hunger Games | **0.45** | 0.70 | 0.53 | 0.69 |
| **Overall** | **0.59** | 1.10 | 0.72 | 0.95 |

Each of these 3 trailers represent a genre of movie: Captain Phillips is a thriller, The Purge is a horror movie, and Red 2 is an action/comedy movie. We can see from Figure 5 that each of emotion annotations have different histogram shapes, and the distribution matches with our intuition. For example, within the valence annotations Red 2 has the most positive annotations and The Purge has the most negative annotations. This makes sense, because horror movies are typically frightening, where as action/comedy movies are funny and exciting. The thriller movie, Captain Phillips, falls somewhere in between. In the arousal space the distribution of annotations exhibit more similarity across movies. Each of the movies that we used in our experiment had some action in the trailers, and therefore each had a high level of arousal during particular portions of the movie.

### C. Emotion Prediction Framework

Our emotion annotations are taken on a section level, but our mid-level feature is defined on the shot-level. Therefore, we can have multiple shots within a section. To create one feature for each of the sections we perform max-pooling over all of the extracted shot features within the section. We also tried averaging over the shot features, but the max-pooling performs better in emotion prediction. We believe this is because the max-pooling allows us to assess the most prominent appearance of a concept in a section. When using averaging over the features within a section, the contribution of one particular shot can be diluted due to many shots appearing within that section. At the end of each section feature we also append three extra section-specific features. These features have been shown to be useful for emotion analysis in the specific domain of movies, and they are the number of shots in the section, the average shot length, and the section length.

To predict the arousal and valence within each section we utilize a form of regularized linear regression, known as ridge regression, with the cost parameter set to one. When predicting the emotions within the trailers our ground truth label for each section is the annotator mean. We predicted the arousal and valence annotator means separately, although we believe that a joint formulation that learns a model for both quantities at the same time may improve performance. However, learning separate models helps to give us insight into the current performance of our emotion prediction and where we can alter our pipeline to improve performance in the future.

## VI. RESULTS

To test our emotion prediction pipeline we utilize leave-one-out cross validation for training and testing. This means that from the seven annotated videos, we train our emotion predictor using six of the videos, and then test the emotion prediction on the held out video. We present the performance on each of the videos as well as the performance across the entire dataset for both arousal and valence prediction. We compare our concept feature to the raw low-level features extracted from the movie shots, and the low-level features after PCA transformation keeping the most significant 23 and 100 dimensions. The emotion prediction pipeline and cross-validation was completed in exactly the same way for each model for all four groups of features. The metric that we used is the average absolute deviation from the annotator mean, and smaller numbers denote better performance. We trained our ridge regression using the Shogun Machine Learning Toolbox [24], and ran our experiments using an Intel Xeon E5-2609 processor with 8 GB of RAM. Average training time for one-fold of the cross-validation for our mid-level feature was 3 ms, and the average training time over one-fold for our full-dimensional low-level feature was 10900 ms. This represents a significant speed-up in computation time if video scenes are represented using our mid-level feature.

Our performance in predicting the evoked arousal of each section can be seen in Table I. We can see that the arousal prediction using our mid-level concept feature outperforms prediction using the low-level features. For arousal prediction the mid-level feature actually outperforms the low-level features both across the aggregate leave-one-out cross validation testing and in the testing of each particular movie. This demonstrates that our defined concept detectors are able to pick up on things throughout the movie sections that the

low-level features may miss, allowing us to more accurately model the evoked arousal of the viewers.

Our performance for predicting the evoked valence within each movie section can be seen in Table II. Once again our pipeline using the mid-level concept feature performs better than the classic low-level feature approach across all of the movies in valence prediction. However, when predicting the valence for some movies the low-level features more accurately predict the evoked valence. We believe this could be occurring because as defined many of our concepts may not be discriminative on the valence domain. For example, fight concepts appear in both Red 2 and in The Purge, but the fights in the action/comedy movie Red 2 are humorous and evoke positive valence in the viewers and the fight scenes in The Purge are frightening and evoke negative valence in the viewers. In our current implementation both types of fight scenes are grouped into one concept detector, "fight". With more granularly defined concepts we believe that the prediction of the viewer valence can be improved.

We can also look at the arousal and valence performance within particular movies, and see how well we are able to predict the emotion within each section for each feature set used. In Figure 6 we can see the arousal and valence predictions for Red 2 and The Conjuring. The line denoted with "Early (PCA)" in the legend of each subgraph denotes the prediction using the most important 100 PCA dimensions of our early fusion low-level feature. In Red 2 it is easy to see that our mid-level feature closely models the annotator mean for both arousal and valence, which is the target of our prediction. The mid-level feature outperforms the low-level features, and has a very similar shape to the actual annotator mean values. In the graph in the bottom left corner of Figure 6, it is shown that we are also able to closely model the arousal within The Conjuring as well. However, in the graph situated in the bottom right corner of Figure 6 our mid-level feature under-performs the low-level features. We can see that both the low-level features detect a drop in annotator valence between section 1 and 2. However, our mid-level feature based emotion prediction stays completely flat and even rises slightly. This shows that something is happening in this section that is not captured by our concept detectors. We hope to address this type of issue by training both a larger number and more granular concept detectors.

## VII. Conclusion

We have introduced a new mid-level concept based emotion prediction pipeline for videos. A framework and website for gathering annotations from on-line users for both shot-level video concept annotations, and manually defined shot level emotion annotations using the two dimensional arousal and valence model were given. Our mid-level feature was trained on over 3000 shot-level concept annotations, and we have shared the performance of each concept detector. Our concept based emotion prediction pipeline was tested

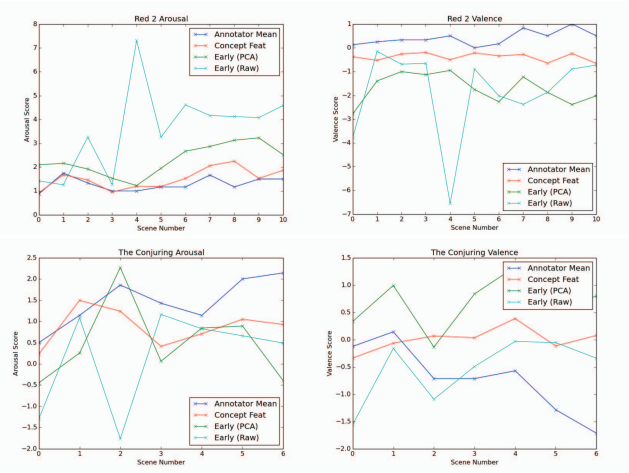


Figure 6. Specific Movie Emotion Prediction. This figure shows the arousal and valence predicted during each section within Red 2 and The Conjuring. From top left moving clockwise the graphs are as follows: Red 2 Arousal, Red 2 Valence, The Conjuring Valence, and the Conjuring Arousal.

against a seven movie trailer dataset with more than 500 emotion annotations, and we have shown that using our mid-level concept feature we are able to greatly improve emotion prediction over standard low-level feature approaches.

We believe that attempting to model an abstract concept such as human emotions using low-level visual and audio features is a near impossible task. However, by leveraging machine-detectable concepts that are closely tied to emotions we can move closer to bridging the affective gap. We believe that this work presents a proof-of-concept of what a robust mid-level concept based system would look like, and we hope that it paves the way for further study in emotion prediction from videos.

## VIII. Future Work

We hope this work brings to light the advantages of using machine detectable mid-level concepts as a means for bridging the affective gap. In particular we believe that our current model could be improved by using more concepts to represent the shots. We also believe that more granular concepts will be of use, in particular in improving the valence prediction. Our concept detector performance could be improved by taking into account motion features, in addition to the stationary image and audio features currently used. Finally, a prediction framework that takes into account the temporal occurrence of sections within a given video should be utilized, because the evoked emotion of each section is dependent on the context in which it is shown.

## ACKNOWLEDGMENTS

This research was sponsored in part by the U.S. Defense Advanced Research Projects Agency (DARPA) under the Social Media in Strategic Communications (SMISC) program, Agreement Number W911NF-12-C-0028. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Defense Advanced Research Projects Agency or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

This work was also sponsored in part by the National Science Foundation Graduate Research Fellowship Program.

## REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, 2008.

[2] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge (emotiw) challenge and workshop summary," in *ACM International Conference on Multimodal Interaction*, 2013.

[3] R. Navarathna, P. Lucey, P. Carr, E. Carter, S. Sridharan, and I. Matthews, "Predicting movie ratings from audience behaviors," in *IEEE Winter Conference on Applications of Computer Vision*, 2014.

[4] J. Hernandez, Z. Liu, G. Hulten, D. Debarr, K. Krum, and Z. Zhang, "Measuring the engagement level of tv viewers," in *IEEE Automatic Face and Gesture Recognition*, 2013.

[5] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transaction on*, 2005.

[6] A. Hanjalic, "Extracting moods from pictures and sounds: towards truly personalized tv," *Signal Processing Magazine, IEEE*, 2006.

[7] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the ACM International Conference on Multimedia*, 2010.

[8] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *Circuits and Systems for Video Technology, IEEE Transactions on*, 2013.

[9] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *Affective Computing, IEEE Transactions on*, 2012.

[10] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," *Affective Computing, IEEE Transactions on*, 2012.

[11] S. Benini, L. Canini, and R. Leonardi, "A connotative space for supporting movie affective recommendation," *Multimedia, IEEE Transactions on*, 2011.

[12] Y.-G. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," in *Proceedings of The AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

[13] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proceedings of the European Conference on Computer Vision*, 2010.

[14] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of ACM International Conference on Multimedia*, 2013.

[15] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *Circuits and Systems for Video Technology, IEEE Transactions on*, 2006.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 2010.

[17] L. jia Li, H. Su, L. Fei-fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in Neural Information Processing Systems*, 2010.

[18] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2003.

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2011.

[21] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (iaps): Technical manual and affective ratings," 1999.

[22] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[23] C. E. Osgood, *The measurement of meaning*. University of Illinois press, 1957, no. 47.

[24] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. d. Bona, A. Binder, C. Gehl, and V. Franc, "The shogun machine learning toolbox," *The Journal of Machine Learning Research*, 2010.