Predicting Emotions in User-Generated Videos

Yu-Gang Jiang, Baohan Xu, Xiangyang Xue
School of Computer Science, Fudan University, Shanghai, China
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China

Abstract

User-generated video collections are expanding rapidly in recent years, and systems for automatic analysis of these collections are in high demands. While extensive research efforts have been devoted to recognizing semantics like "birthday party" and "skiing", little attempts have been made to understand the emotions carried by the videos, e.g., "joy" and "sadness". In this paper, we propose a comprehensive computational framework for predicting emotions in user-generated videos. We first introduce a rigorously designed dataset collected from popular video-sharing websites with manual annotations, which can serve as a valuable benchmark for future research. A large set of features are extracted from this dataset, ranging from popular low-level visual descriptors, audio features, to high-level semantic attributes. Results of a comprehensive set of experiments indicate that combining multiple types of features—such as the joint use of the audio and visual clues—is important, and attribute features such as those containing sentiment-level semantics are very effective.

Introduction

Automatic techniques for understanding the emotions in diverse user-generated videos on the Web are helpful for many applications. For example, the emotions contained in videos about a new electronic product may be used by a company to improve the product and perform targeted marketing, i.e., promotion on a particular group of customers who (or, some of whom) expressed positive emotions in their videos. Governments can also utilize this function to better understand people's reactions about hot events or new policies.

In this paper, we present a comprehensive computational approach for predicting emotions purely based on video content analysis. While significant progress has been made on the computational inference of emotions in images (Joshi et al. 2011), previous research on video emotions has mostly been conducted on movie data (Wang and Cheong 2006). To the best of our knowledge, there is no existing work investigating this problem on user-generated videos, which have more diversified contents with little quality control and post-editing. One important issue that has limited the needed progress of emotion analysis in the user-generated videos is the scarcity of well-defined datasets with manual annotations. To prompt research on this interesting and important problem, we first construct and publicly release a benchmark dataset[1] based on videos downloaded from YouTube and Flickr (see Figure 1 for several example frames). A large set of features are then extracted from this dataset, covering not only audio and visual descriptors that were popularly used in the works on movie video analysis, but also new attribute features that have semantic meanings in each dimension. Using a state-of-the-art prediction model, we provide a comprehensive analysis of the effect of each individual feature and their combinations, leading to several interesting observations.

Figure 1: Example frames of four emotion categories from the dataset we collected.

This work makes two important contributions:

- In establishing a good benchmark for emotion analysis in user-generated videos, we construct a dataset with eight manually annotated emotions. We analyze and identify

[1] Available at www.yugangjiang.info/research/VideoEmotions/.

potentially helpful clues for emotion recognition on this dataset, which are important for the design of a good computational model.

- We compute and evaluate a large set of audio-visual features, and introduce the use of semantic attributes for emotion prediction. Several valuable insights are attained from extensive evaluations, which set the foundation for future research of this challenging problem.

Notice that the emotion carried by a video is not necessarily the same with the emotion of a particular person after viewing the video. While the latter could be highly subjective, the dominant emotion expressed by the content of a video, or that intended to be delivered by the owner of the video, can be considered relatively more objective, and therefore it is possible to develop computational models to predict it.

Related Works

The computational inference of emotions in images has been studied extensively, partly stimulated by the availability of the International Affective Picture System (IAPS) benchmark (Lang, Bradley, and Cuthbert 2008). In (Yanulevskaya et al. 2008), the authors designed a system based on holistic image features to predict emotions. They also showed the potential of applying their models trained on the IAPS to images of masterpieces. More advanced features inspired by psychology and art theory were utilized in (Machajdik and Hanbury 2010), where color, texture, composition, and faces were extracted for emotion prediction. Lu et al. further investigated the relationship between shape features and image emotions (Lu et al. 2012). The authors proposed a method to compute features that can model several shape characteristics like roundness and angularity, which were shown to be very complementary to the traditional low-level features and the combination of all of them led to state-of-the-art results on the IAPS. It is also worthwhile mentioning here that many approaches have been proposed to model the aesthetics and interestingness aspects of images or videos (Murray, Marchesotti, and Perronnin 2012; Jiang et al. 2013), which were occasionally studied together with emotions (Joshi et al. 2011).

Existing works on video emotion recognition mostly focused on the movie domain. In (Kang 2003), Kang proposed to use Hidden Markov Model for affect analysis in movies based on low-level features such as color and motion. The authors of (Rasheed, Sheikh, and Shah 2005) adopted several visual features in a mean shift based classification framework to identify the mapping between the features and six movie genres. One similar observation from these works is that combining multiple visual features is effective. More recently, in addition to purely using visual features, the authors of (Wang and Cheong 2006; Xu et al. 2012; Teixeira, Yamasaki, and Aizawa 2012) emphasized the importance of jointly using audio and visual features, and showed promising results on a set of Hollywood movies. Audio features are intuitively effective as some emotions like "joy" may contain clear auditory clues (e.g., the cheering sound). Besides, a few researchers have investigated this problem on other types of data like meeting (Jaimes et al. 2005) and sports (Ren, Jose, and Yin 2007) videos.

Our work in this paper is different from the previous studies in that we focus on user-generated videos, which have several unique characteristics compared with movies. First, the user-generated videos are normally very short (e.g., a few minutes) and thus there could be a single dominant emotion per video. This is different from movies where many emotions co-exist and emotion recognition has to be done on segment level. Second, the analysis of the user-generated videos is more challenging as the contents are highly diversified with almost no quality control. Unlike movies, they are mostly from amateur consumers and thus do not follow professional editing rules or styles. In addition to a comprehensive computational system, to facilitate this research, we constructed a benchmark dataset, which is also considered as a contribution. Public benchmarks have played very important roles in advancing many artificial intelligence problems, but a public dataset for video emotion analysis has been elusive until very recently the work of (Baveye et al. 2013), which was, however, built on movie videos.

The Dataset

We constructed a dataset based on videos downloaded from the Web. Eight emotion categories are considered according to the well-known Plutchik's wheel of emotions (Plutchik 1980), including "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", and "trust". These categories were popularly adopted in the existing works on emotion analysis.

To download a sufficient amount of videos for manual annotation, we used a finer-grained emotion categorization also defined in (Plutchik 1980), so that more searches can be formed and more relevant videos may be found. Each of the eight categories was extended to three sub-classes, e.g., "apprehension", "fear" and "terror" for the broader category of "fear". This allows us to use 24 YouTube and 24 Flickr searches. Notice that two search engines were used to download more videos, and the videos from both websites were merged after annotation to form a single dataset.

We downloaded the largest allowed number of videos from each search, leading to 4,486 videos from YouTube and 3,215 from Flickr. These videos were manually filtered by 10 annotators (5 males and 5 females), who were instructed about the detailed definition of each emotion before performing the task. After careful annotations of all the videos by each annotator separately, a group discussion was held to finalize the categories of the videos with inconsistent initial labels. The final dataset contains 1,101 videos, with a minimum number of 100 videos per category and an average duration of 107 seconds. Table 1 summarizes more details.

We looked into the videos manually to see if there are computable clues highly correlated with each emotion category, which, if existed, would be very helpful for the design of a good prediction system. While the problem was found to be very complex, as expected we had the feeling that both audio and visual information are important. In addition, we also observed that some emotions share high correlations with certain semantics like the existence of a par-

Category	# Flickr videos	# YouTube videos	Total
Anger	23	78	101
Anticipation	40	61	101
Disgust	100	15	115
Fear	123	44	167
Joy	133	47	180
Sadness	63	38	101
Surprise	95	141	236
Trust	44	56	100
Ave. duration	54s	175s	107s

Table 1: The number of videos per emotion category in our dataset.

ticular event or object. For example, the emotion of "joy" may frequently co-occur with events like parties and kids playing. This observation motivated us to propose the use of semantic attributes for video emotion analysis, which will be described later. Figure 1 gives a few example video frames from the dataset.

The Computational System

This section introduces a comprehensive computational system for emotion prediction. Figure 2 shows the emotion prediction framework of our system. Similar to many other video content recognition problems, the most important component in the system is feature representation which converts the original videos into fixed-dimensional feature vectors based on certain computable rules. For this, we consider three groups of features, covering a wide range of popular visual and audio descriptors, as well as several newly developed semantic attribute representations. The effectiveness of jointly using visual and audio features has been justified by prior works on movie emotion analysis, but the audio-visual feature set used in this work is more comprehensive and we expect that some features never used before are helpful. Due to space limit, we briefly introduce each of the features below. Interested readers may refer to the corresponding references for more details.

Visual and Audio Features

Dense SIFT (Scale Invariant Feature Transform) is a powerful visual feature in many image and video content recognition tasks. The SIFT descriptors are computed following the original work of (Lowe 2004), except that the local frame patches are densely sampled instead of using interest point detectors. Since there can be many SIFT descriptors extracted from a single video frame, we quantize them into a fixed-dimensional bag-of-words representation, which has been popular for over a decade (Sivic and Zisserman 2003). A codebook of 300 codewords is used in the quantization process with a spatial pyramid of three layers (Lazebnik, Schmid, and Ponce 2006). Since neighboring frames are similar and feature extraction is computationally expensive, we sample a frame per second. These frames are used for computing all the other features except the audio ones, for which the entire soundtrack is used.

HOG (Histogram of Gradients) descriptor was originally

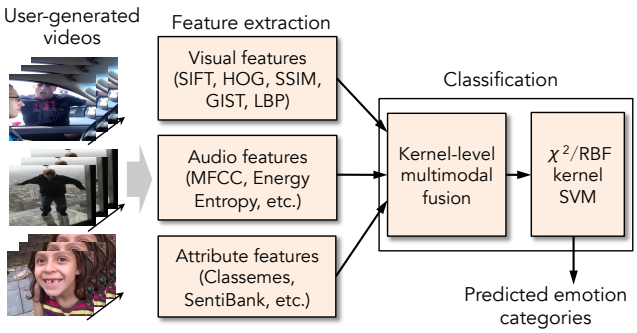

Figure 2: The emotion prediction pipeline of our computational system. See texts for more details.

proposed for human detection in images (Dalal and Triggs 2005), and has been widely adopted as a type of local feature in many visual recognition applications since then. Like the dense SIFT based representation, the HOG descriptors are computed on densely sampled frame patches, which are then converted to a bag-of-words representation for each video in the same way as the SIFT descriptors.

SSIM (Self-Similarities) is also a type of local visual descriptors (Shechtman and Irani 2007). Different from the gradient based descriptors like the SIFT, SSIM is obtained by quantizing the correlation map of a densely sampled patch in a larger circular window around the patch. The SSIM descriptors from each video are also quantized into a bag-of-words representation.

GIST is a global feature that mainly captures the texture characteristics of a video frame. It is computed based on the output energy of several Gabor-like filters (8 orientations and 4 scales) over a dense frame grids (Oliva and Torralba 2001). The averaged GIST feature over all the sampled frames is used to represent a video.

LBP (Local Binary Patterns) (Ojala, Pietikainen, and Maenpaa 2002) is another popular texture feature capturing different visual aspects. It uses binary numbers to label each frame pixel by comparing its value with that of its neighborhood pixels. The averaged representation of all the frames is used as the video feature. All the aforementioned visual features are extracted using the codes from the authors of (Xiao et al. 2010).

MFCC: Neuroscientists have found that human perception often relies on the use of multiple senses (Stein and Stanford 2008). In addition to the visual features, audio clues are an important complement to reach our goal in this work. The first audio feature being considered is the mel-frequency cepstral coefficients (MFCC), which is probably the most well-known audio representation in the field. An MFCC descriptor is computed over every 32ms time-window with 50% overlap. The descriptors from the entire soundtrack of a video are also converted to a bag-of-words representation using vector quantization.

Audio-Six: We also include another compact audio feature consisting of six basic audio descriptors that have been frequently adopted in audio and music classification,

including Energy Entropy, Signal Energy, Zero Crossing Rate, Spectral Rolloff, Spectral Centroid, and Spectral Flux. These descriptors are expected to be complementary to the MFCC as they capture different aspects of an audio signal.

Attribute Features

Unlike the low-level audio-visual features, attribute features contain rich semantics that are potentially very useful, since emotions often occur under certain (semantically interpretable) environments. We therefore propose to use attributes for emotion prediction, and adopt three kinds of attribute descriptors that cover a wide variety of semantics.

Classemes (Torresani, Szummer, and Fitzgibbon 2010) is an attribute descriptor generated by models trained on images from the Web. It consists of automatic detection scores of 2,659 semantic concepts (mainly objects and scenes). Each dimension of this representation corresponds to one semantic category. A Classemes representation is computed on a video frame, and similar to some of the visual descriptors, the averaged representation of all the frames is used as the video feature.

ObjectBank (Li et al. 2010) is another high-level attribute descriptor originally proposed for image classification. Different from the Classemes which uses *global* frame-level concept classification scores, ObjectBank uses the response scores of *local* object detections. Since objects may appear in very different scales, the detection is performed on multiple image (frame) resolutions. There are 177 object categories in the ObjectBank representation.

SentiBank: We also consider a new attribute representation based on emotion related concepts (Borth et al. 2013). There are 1,200 concepts in SentiBank, and each is defined as an adjective-noun pair, e.g., "scary dog" and "lonely road", where the adjective is strongly related to emotions and the noun corresponds to objects and scenes that are expected to be automatically detectable. Models for detecting the concepts were trained on Flickr images. This set of attributes is intuitively effective for the problem in this work as the emotion-related objects and scenes are very helpful clues for determining the emotion of the user-generated videos.

Classification

With the video features, emotion prediction models can be easily trained. We adopt the popular SVM due to its outstanding performance in many visual recognition tasks. For the kernel option of the SVM, we adopt the χ^2 RBF kernel for all the bag-of-words representations, because it is particularly suitable for histogram-like features. The standard Gaussian RBF kernel is used for the remaining features. We follow the one-against-all strategy to train a separate classifier for each category, and a test sample is assigned to the category with the highest prediction score.

As the selected features are from complementary information channels, combining them is very important for achieving outstanding performance. We adopt kernel-level fusion, which linearly combines kernels computed on the individual features. Equal fusion weights are used in our experiments for simplicity. It is worth noting that dynamic weights predicted by cross-validation or multiple kernel learning techniques may produce slightly better results, according to existing works on other visual recognition tasks. Nevertheless, a different and more complex fusion strategy is not expected to change the major conclusions gained from our analysis.

Experiments

We now introduce experimental settings and discuss the results. In addition to using the entire dataset of eight emotion categories, we also discuss results on a subset of four emotions ("Anger", "Fear", "Joy", and "Sadness"), which have been more frequently adopted in the existing works. For both the entire set and the subset, we randomly generate ten train-test splits, each using 2/3 of the data for training and 1/3 for testing. A model is trained on each split for each emotion, and we report the mean and standard-deviation of the ten prediction accuracies, which are measured as the proportions of the test samples with correctly assigned emotion labels. In the following we first report results of the visual, audio and attribute features separately, and then discuss their fusion performance.

Visual features: Results of the five visual features are summarized in Figure 3 (a) for the subset, and Figure 4 (a) for the entire dataset. Overall the results are fairly good, with an accuracy around 50% on the subset and nearly 40% on the entire dataset. Among the five features, dense SIFT and HOG are consistently the top performers, followed by SSIM. SIFT and HOG features are computed based on local pixel gradients. Although it is difficult to explain why gradients can be used to better infer emotions, both of them are the state-of-the-art features in recognizing image/video semantics, and have been frequently shown to be more effective than features like GIST and LBP.

We also discuss the results of fusing multiple visual features. As there are too many different feature combinations to be reported in detail, we select only a subset of assumingly important combinations based on the following strategy, which has been found effective empirically. We start from the best visual feature and incrementally include new features (ordered by their individual feature performance). A newly added feature is discarded if fusing it does not improve the results. As shown in Figure 3 (a) and Figure 4 (a), combining more features does not always lead to better results. The fusion of SIFT and HOG (indicated by "12" in both figures) is clearly useful, but adding LBP does not contribute to the results on both the subset and the entire set.

Audio features: Figure 3 (b) and Figure 4 (b) visualize the results of the audio features. Both MFCC and Audio-Six are discriminative for emotion prediction, confirming the fact that the audio soundtracks contain useful information. However, their performance is lower than that of all the visual features, which indicates that the visual channel is more important. These overall results of the audio features are not very competitive because the prediction accuracies of the audio features are very low for some emotions with weak audio clues (e.g., "sadness"). We will discuss per-category performance later. In addition, we observe that the two audio features are very complementary. A performance improvement of over 9% is obtained from their fusion on both the sub-

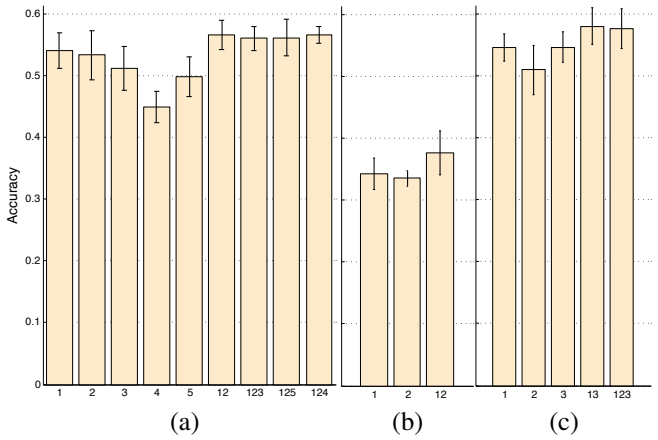

Figure 3: Prediction accuracies on the subset of four emotion categories, using models based on individual features and their fusion. (a) Visual features (1. Dense SIFT; 2. HOG; 3. SSIM; 4. GIST; 5. LBP). (b) Audio features (1. MFCC; 2. Audio-Six). (c) Attribute features (1. Classemes; 2. ObjectBank; 3. SentiBank). Notice that, in the fusion experiment, not all the feature combinations are reported. A feature is dropped immediately if adding it does not improve the results (see texts for more details). The best feature combinations are "124", "12" and "13" within the visual, audio and attribute feature sets, respectively.

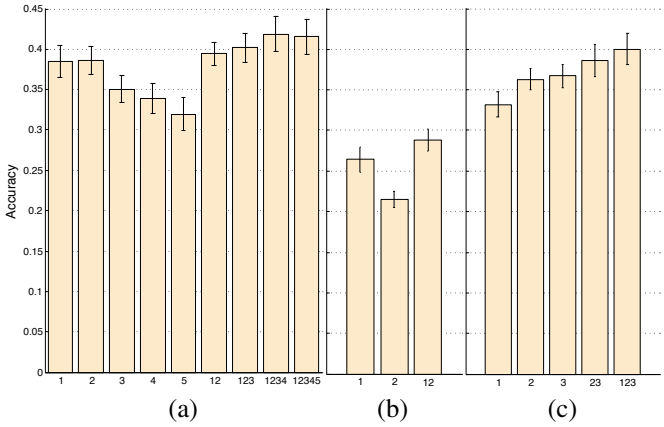

Figure 4: Prediction accuracies on the entire dataset of eight emotion categories, presented following the same strategy given in the caption of Figure 3. The best feature combinations are "1234", "12" and "123" within the visual, audio and attribute feature sets, respectively.

set and the entire set, over the higher of the two individual features.

Attribute features: Next we discuss results of the attribute features, which, to our knowledge, have never been used for video emotion analysis. As shown in Figure 3 (c) and Figure 4 (c), the attribute features demonstrate very strong performance, similar to or even higher (on the four-class subset) than the visual features. This confirms our conjecture that semantic-level clues are effective for predicting

Category	Visual	Audio	Attribute	Vis.+Aud.+Att.
Anger	52.4±4.1	7.9±2.1	59.4±2.5	**64.2**±2.4
Fear	60.7±2.5	45.5±11.0	60.5±4.0	**62.7**±3.1
Joy	68.0±2.7	62.2±11.2	**71.9**±2.4	69.5±1.6
Sadness	34.2±1.2	10.6±2.0	28.2±1.6	**37.0**±3.0
Overall	*56.7±1.3*	*37.7±3.6*	*58.1±2.9*	*60.5±2.3*
Anger	49.4±2.9	12.1±2.8	37.6±2.5	**53.0**±1.8
Anticipation	3.0±0.9	3.3±1.3	3.9±1.4	**7.6**±1.9
Disgust	35.1±3.2	14.9±1.9	33.1±3.1	**44.6**±2.3
Fear	45.0±2.6	12.5±4.0	**49.5**±3.1	47.3±3.0
Joy	44.8±2.6	35.7±4.4	41.2±4.1	**48.3**±2.7
Sadness	**23.5**±2.2	0.0±0.0	13.2±3.5	20.0±2.3
Surprise	75.6±4.4	75.1±6.3	77.5±3.3	**76.9**±4.8
Trust	10.6±1.1	24.8±2.1	8.8±1.5	**28.5**±1.6
Overall	*41.9±2.2*	*28.8±1.4*	*40.0±1.9*	*46.1±1.7*

Table 2: Prediction accuracies (%) of each emotion category, using the visual, audio and attribute features, and their fusion. *Top*: results on the subset of four categories. *Bottom*: results on the entire set of eight categories. The highest accuracy of each category is shown in bold.

emotions. It is important to notice that the models used for generating the attribute features were all offline trained using Web images, which have significant data domain difference from the user-generated videos. Therefore we expect that the performance of the attribute features can be largely improved if the attributes could be detected by models directly trained on the videos.

Among the three attribute features, there is no clear winner. The SentiBank is consistently competitive, indicating that emotion related attributes are very suitable for this task. In addition, the three features are very complementary. Substantial improvements are obtained from fusion.

Combining visual, audio and attribute features: The last experiment is to fuse the features from the three different groups. Within each group, we select the feature combination that demonstrates the best result in the intra-group fusion experiments (indicated in Figure 3 and Figure 4). Table 2 (the two italic "Overall" rows) gives the results of each feature group and the fusion of all the three groups, on both the subset and the entire set. We see that the fusion of features from the three groups clearly improves the results. On the entire dataset of eight emotions, the accuracy is significantly improved from 41.9% to 46.1%. When only fusing the visual and audio features, the accuracies are 56.4% on the subset and 44.9% on the entire set, which are clearly lower than that from fusing all the three groups. This indicates that the attribute features are consistently effective and complementary to the audio-visual features.

Comparing results across the subset and the entire set, the performance on the subset is higher because the chance of confusion is lower. Audio does not improve the overall results on the subset (visual only: 56.7%; visual+audio: 56.4%) as its performance is very low for a few categories, which will be discussed below.

Per-category results: The predication accuracies of each emotion category on both sets are listed in Table 2. We see that some categories such as "joy" have very high accura-

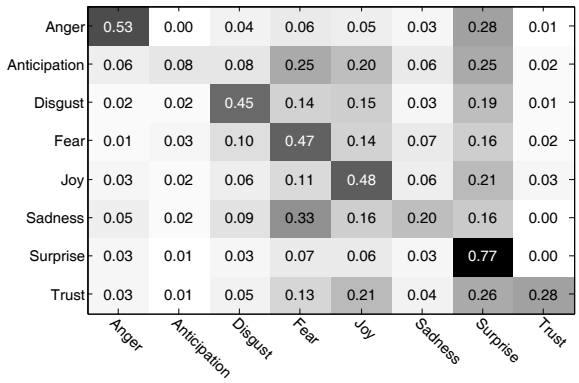

Figure 5: Confusion matrix of the entire set, based on the fusion of the selected features in all the three groups.

cies, while a few emotions like "anticipation" are very difficult to predict. This is not surprising as the emotion of "anticipation" does not have clear audio-visual clues, compared with the "easy" categories.

Audio is good at predicting emotions like "joy" and "surprise", but does not perform as well as we expected for "anger" and "sadness". This is because some user-generated videos expressing "anger" and "sadness" were captured far away from the major subjects, and as a result the volume of sounds from the subjects is very low or dominated by other "noises" like musics. For instance, there are videos about traffic accidents and drivers arguing captured by another driver in his own car. This characteristic of user-generated videos is generally different from professionally captured videos like the movies. Another interesting observation is that audio is much better than the visual and attribute features for the "trust" emotion. This is due to a fact the many user-generated videos expressing the trust emotion are about *trust tests* with laughing and cheering sounds.

Similar to the visual features, attributes show strong performance for many categories. In addition, as shown in the table, fusion leads to top results for most of the categories, which again verify the importance of using multiple features for emotion prediction. The confusion matrix of the fusion results is shown in Figure 5.

Figure 6 further shows some easy and difficult examples. The success examples share some common audio/visual characteristics like the dark environments in videos under the "fear" category, or the laughing sounds in those under "joy". A few "joy" videos were wrongly classified as "fear", which is probably because of the dark lighting and the screaming-like sounds. Some failure examples of the "anger" and "sadness" categories only contain slight facial expressions, which are difficult to be captured by the current set of features.

Conclusions

We have presented a comprehensive computational framework for video emotion analysis. While previous studies were mostly conducted on movie videos, this work focuses

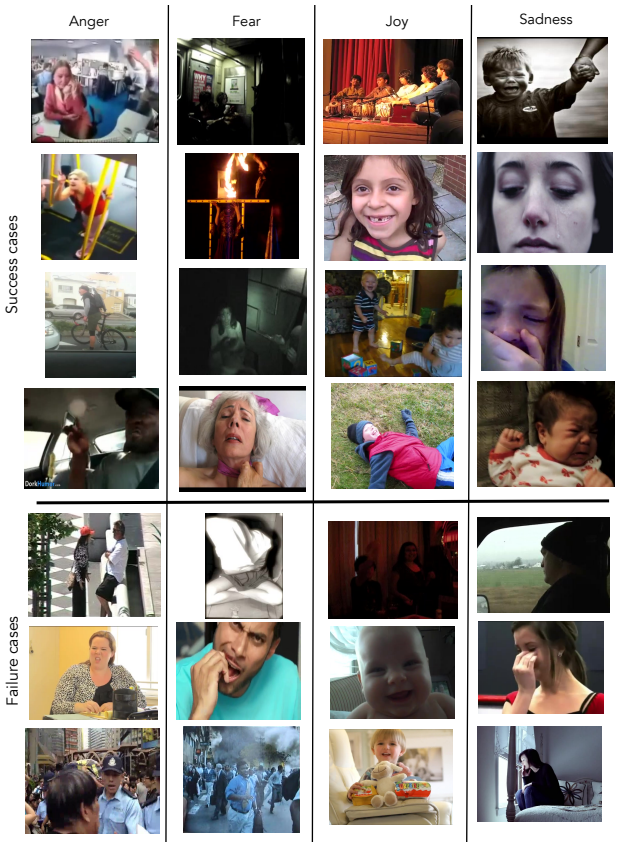

Figure 6: Success and failure examples of four emotion categories, produced by models using the fusion of all the features.

on user-generated videos, the dominant type of videos on the Web. To facilitate the study, we constructed a benchmark dataset with manual annotations, which is valuable for future investigations on this interesting and challenging problem.

In addition to evaluating a large set of low-level audio and visual features, we also proposed to use attributes, a high-level representation with semantic meanings in each dimension. Results from a large set of experiments have shown that models based on the attribute features can produce very competitive performance. The features are also highly complementary—combining attributes with the audio-visual features shows very promising results.

While the results are encouraging, there are several directions deserving future investigations. First, the audio features in the current framework are limited, and using more advanced features may significantly improve the results. In particular, the speech analysis community has developed several features like the voicing related descriptors, which have demonstrated promising results in audio-based emotion prediction (Schuller et al. 2013). In addition, the attributes were computed using models trained on Web images, which have significant domain difference from videos. Therefore, training a new set of attribute models specifically for this task, using user-generated videos, is a promising direction.

Acknowledgments

This work was supported in part by two grants from the National Natural Science Foundation of China (#61201387 and #61228205), a National 863 Program (#2014AA015101), two grants from the Science and Technology Commission of Shanghai Municipality (#13PJ1400400 and #13511504503), and a New Teachers' Fund for Doctoral Stations, Ministry of Education (#20120071120026), China.

References

Baveye, Y.; Bettinelli, J.-N.; Dellandrea, E.; Chen, L.; and Chamaret, C. 2013. A large video data base for computational models of induced emotion. In *Proc. of Humaine Association Conference on Affective Computing and Intelligent Interaction*.

Borth, D.; Chen, T.; Ji, R.-R.; and Chang, S.-F. 2013. Sentibank: Large-scale ontology and classiers for detecting sentiment and emotions in visual content. In *Proc. of ACM Multimedia*.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

Jaimes, A.; Nagamine, T.; Liu, J.; and Omura, K. 2005. Affective meeting video analysis. In *Proc. of IEEE International Conference on Multimedia and Expo*.

Jiang, Y.-G.; Wang, Y.; Feng, R.; Xue, X.; Zheng, Y.; and Yang, H. 2013. Understanding and predicting interestingness of videos. In *Proc. of AAAI Conference on Artificial Intelligence*.

Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.-T.; Wang, J. Z.; Li, J.; and Luo, J. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28(5):94–115.

Kang, H.-B. 2003. Affective content detection using HMMs. In *Proc. of ACM Multimedia*.

Lang, P. J.; Bradley, M. M.; and Cuthbert, B. N. 2008. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical Report A-8. University of Florida, Gainesville, FL.*

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

Li, L.; Su, H.; Xing, E.; and Fei-Fei, L. 2010. Object Bank: A high-level image representation for scene classification semantic feature sparsification. In *Proc. of Advances in Neural Information Processing Systems*.

Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision* 60:91–110.

Lu, X.; Suryanarayan, P.; Adams, R. B.; Li, J.; Newman, M. G.; and Wang, J. Z. 2012. On shape and the computability of emotions. In *Proc. of ACM Multimedia*.

Machajdik, J., and Hanbury, A. 2010. Affective image classication using features inspired by psychology and art theory. In *Proc. of ACM Multimedia*.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

Ojala, T.; Pietikainen, M.; and Maenpaa, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7):971–987.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision* 42:145–175.

Plutchik, R. 1980. *Emotion: Theory, Research, and Experience*. Number Volume 1. Academic Press.

Rasheed, Z.; Sheikh, Y.; and Shah, M. 2005. On the use of computable features for film classification. *IEEE Trans. on Circuits and Systems for Video Technology* 15(1):52–64.

Ren, R.; Jose, J.; and Yin, H. 2007. Affective sports highlight detection. In *Proc. of European Signal Processing Conference*.

Schuller, B.; Steidl, S.; Batliner, A.; and et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. of Interspeech Conference*.

Shechtman, E., and Irani, M. 2007. Matching local self-similarities across images and videos. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

Sivic, J., and Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proc. of IEEE International Conference on Computer Vision*.

Stein, B. E., and Stanford, T. R. 2008. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience* 9:255–266.

Teixeira, R. M. A.; Yamasaki, T.; and Aizawa, K. 2012. Determination of emotional content of video clips by low-level audiovisual features. *Multimedia Tools and Applications* 61(1):21–49.

Torresani, L.; Szummer, M.; and Fitzgibbon, A. 2010. Efficient object category recognition using classemes. In *Proc. of European Conference on Computer Vision*.

Wang, H.-L., and Cheong, L.-F. 2006. Affective understanding in film. *IEEE Trans. on Circuits and Systems for Video Technology* 16(6):689–704.

Xiao, J.; Hays, J.; Ehinger, K.; Oliva, A.; and Torralba, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

Xu, M.; Xu, C.; He, X.; Jin, J. S.; Luo, S.; and Rui, Y. 2012. Hierarchical affective content analysis in arousal and valence dimensions. *Signal Processing* 93(8):2140–2150.

Yanulevskaya, V.; van Gemert, J. C.; Roth, K.; Herbold, A. K.; Sebe, N.; and Geusebroek, J. M. 2008. Emotional valence categorization using holistic image features. In *Proc. of IEEE International Conference on Image Processing*.