Understanding and Predicting Interestingness of Videos

Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, Hanfang Yang
School of Computer Science, Fudan University, Shanghai, China

Abstract

The amount of videos available on the Web is growing explosively. While some videos are very interesting and receive high rating from viewers, many of them are less interesting or even boring. This paper conducts a pilot study on the understanding of human perception of video interestingness, and demonstrates a simple computational method to identify more interesting videos. To this end we first construct two datasets of Flickr and YouTube videos respectively. Human judgements of interestingness are collected and used as the ground-truth for training computational models. We evaluate several off-the-shelf visual and audio features that are potentially useful for predicting interestingness on both datasets. Results indicate that audio and visual features are equally important and the combination of both modalities shows very promising results.

Introduction

The measure of *interestingness* of videos can be used to improve user satisfaction in many applications. For example, in Web video search, for the videos with similar relevancy to a query, it would be good to rank the more interesting ones higher. Similarly this measure is also useful in video recommendation, where users will certainly be more satisfied and as a result the stickiness of a video-sharing website will be largely improved if the recommended videos are interesting and attractive.

While great progress has been made to recognize semantic categories (e.g., scenes, objects and events) in videos (Xiao et al. 2010; Laptev et al. 2008), a model for automatically predicting the interestingness of videos is elusive due to the subjectiveness of this criterion. In this paper we define interestingness as a measure that is collectively reflected based on the judgements of a large number of viewers. We found that—while a particular viewer may have a very different judgement—generally most viewers can reach a clear agreement on whether a video is more interesting than another. The problem we pose here is: given two videos, can a computational model automatically analyze their contents and predict which one is more interesting? To answer this challenging question, this paper conducts an in-depth study using a large set of off-the-shelf video features and a state-of-the-art prediction model. Most selected features have been shown effective in recognizing video semantics, and a few are chosen particularly for this task based on our intuitions. We investigate the feasibility of predicting interestingness using two datasets collected from Flickr and YouTube respectively, both with ground-truth labels generated based on human judgements (see example frames in Figure 1).

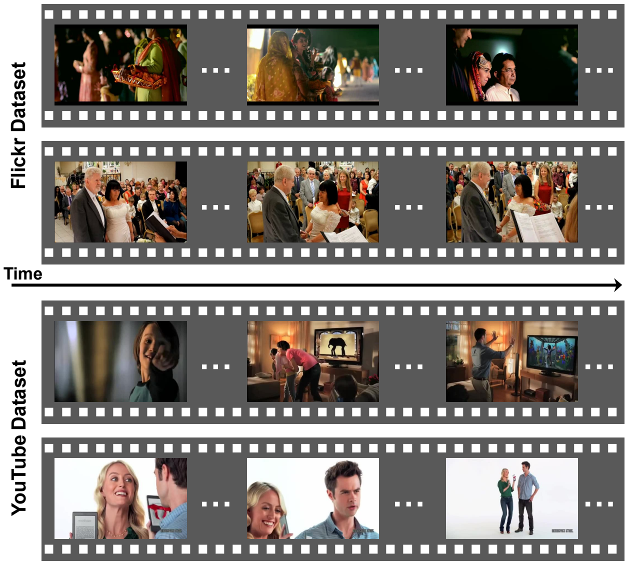

Figure 1: Example frames of videos from the Flickr and YouTube datasets we collected. For each dataset, the video on top is considered more interesting than the other one according to human judgements.

This paper makes two important contributions:

- We construct two benchmark datasets with ground-truth labels to support the study of video interestingness[1]. We analyze the factors that may reflect interestingness on both datasets, which are important for the development of a good computational model.

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

- We evaluate a large number of features (visual, audio, and high-level attributes) through building models for predicting interestingness. This provides critical insights in understanding the feasibility of solving this challenging problem using current techniques.

Notice that our aim is to have a *general* interestingness prediction that most people would agree with, not specifically tailored for a particular person. While the latter might also be predictable with a sufficient amount of training data and supporting information like browsing history, it is out of the scope of this work.

Related Work

There have been a number of works focusing on predicting aesthetics of images (Datta et al. 2006; Ke, Tang, and Jing 2006; Murray, Marchesotti, and Perronnin 2012), which is a related yet different goal. Several papers directly investigated the problem of image interestingness. In (Katti et al. 2008), the authors used human experiments to verify that people can discriminate the degree of image interestingness in very short time spans, and discussed the potentials of using this criterion in various applications like image search. The use of interestingness was also explored in (Lerman, Plangprasopchok, and Wong 2007) to achieve improved and personalized image search results. A more recent work in (Dhar, Ordonez, and Berg 2011) presented a computational approach for the estimation of image interestingness using three kinds of attributes: compositional layouts, content semantics, and sky-illumination types (e.g., cloudy and sunset). The authors found that these high-level attributes are more effective than the standard features (e.g., color and edge distributions) used in (Ke, Tang, and Jing 2006), and further combining both of them can achieve significant improvements.

Another work that is perhaps the most related to ours is (Liu, Niu, and Gleicher 2009), where the authors used Flickr images to measure the interestingness of video frames. Flickr images were assumed to be mostly interesting as compared to many video frames since the former is generally well-taken and selected. A video frame is considered interesting if it matches (using image local features) to a large number of images. Promising results were observed on travel videos of several famous landmarks. Our work is fundamentally different since predicting the interestingness of a video frame is essentially the same as that for images. Entire video-level prediction in this paper demands serious consideration of several other clues such as audio. To the best of our knowledge, there is no existing works studying computational approaches for the modeling of video-level interestingness.

Datasets

To facilitate the study we need benchmark datasets with ground-truth interestingness labels. Since there is no such kind of dataset publicly available, we collected two new datasets.

The first dataset was collected from Flickr, which has a criterion called "interestingness" to rank its search results. While the exact way of computing this is unknown, it is clear that various subjective measures of human interactions are considered, such as the number of views, tags and comments, popularity of video owners, etc. We used 15 keywords to form 15 Flickr searches and downloaded top-400 videos from each search, using the interestingness-based ranking. The keywords include "basketball", "beach", "bird", "birthday", "cat", "dancing", "dog", "flower", "graduation", "mountain", "music performance", "ocean", "parade", "sunset" and "wedding". We found that the interestingness ranking of videos is not locally stable, i.e., the orders of nearby videos may change from time to time. This may be due to the fact that the criterion is frequently updated based on most recent user interactions. We therefore follow a previous work on image interestingness analysis (Dhar, Ordonez, and Berg 2011) to use the top 10% of the videos from each search as "interesting" samples, and the bottom 10% as "uninteresting" samples. Therefore the final dataset contains 1,200 videos ($15 \times 400 \times$(10%+10%)), with an average duration of around 1 minute (20 hours in total).

We underline that although the Flickr search engine already has the function of estimating video interestingness, it relies on human interactions/inputs which become available only after a certain time period. A computational model based on purely content analysis is desired as it can predict interestingness immediately. Besides, content-based interestingness estimation can be used in other domains where human interactions are not available.

The second dataset consists of advertisement videos downloaded from YouTube[2]. Practically it may be useful if a system could automatically predict which ad is more interesting. We collected ad videos in 14 categories: "accessories", "clothing&shoes", "computer&website", "digital products", "drink", "food", "house application", "houseware&furniture", "hygienic products", "insurance&bank", "medicine", "personal care", "phone" and "transportation". 30 videos were downloaded for each category, covering ads of a large number of products and services. Average video duration of this dataset is 36 seconds.

After collecting the YouTube data, we assigned an "interestingness" score to each video, defined by the average ranking provided by human assessors after viewing all videos in the corresponding category. Ten assessors (five males and five females) were involved in this annotation process. Each time an assessor was shown a pair of videos and asked to tell which one is more interesting (or both are equally interesting). Labels from the 10 assessors were consolidated to form fine-grained interestingness rankings of all the videos from each category.

We analyzed the videos and labels in order to understand if there are clear clues that are highly correlated with interestingness, which would be helpful for the design of the computational models. While the problem was found to be very complex and challenging as expected, one general observation is that videos with humorous stories, attrac-

[2] We chose YouTube because Flickr has much fewer ad videos, whose interestingness rankings are also unreliable due to the limited number of views received by those videos.

tive background music, or better professional editing tend to be more interesting. Figure 1 shows a few example video frames from both datasets.

Predicting Interestingness

With the datasets we are now able to develop and evaluate computational models. A central issue in building a successful model is the selection or design of video features, which are fed in machine learning algorithms for prediction. A large number of features are studied, ranging from visual, audio, to high-level semantic ones (a.k.a. attributes). These features are selected based on past experiences on video analysis (e.g., object/scene recognition) and intuitions on what may be useful for predicting interestingness. We briefly describe each of them in the following subsection.

Features

Color Histogram in HSV space is the first feature to be evaluated, as color is an important clue in visual perception.

SIFT (Scale Invariant Feature Transform) feature has been popular for years, producing outstanding performance in many applications. We adopt the standard pipeline as described in (Lowe 2004). Since SIFT computation on every frame is computationally expensive and nearby frames are visually very similar, we sample a frame per second. The same set of sampled frames are used in computing all the other features except the audio ones. The SIFT descriptors of a video are quantized into a bag-of-words representation (Sivic and Zisserman 2003) using a spatial pyramid with a codebook of 500 codewords. This representation has been widely used in the field of image/video content recognition.

HOG (Histogram of Gradients) features (Dalal and Triggs 2005) are effective in several computer vision tasks such as human detection. Different from SIFT which is computed on sparsely detected patches, we compute HOG descriptors over densely sampled patches (on the same set of frames). Following (Xiao et al. 2010), HOG descriptors in a $2{\times}2$ neighborhood are concatenated to form a descriptor as higher dimensional feature is normally more discriminative. The descriptors are converted to a bag-of-words representation for each video, in the same way as we quantize the SIFT features.

SSIM (Self-Similarities) is another type of local descriptor, which is computed by quantizing a correlation map of a densely sampled frame patch within a larger circular window (Shechtman and Irani 2007). We set the patch size and the window radius as 5×5 and 40 pixels respectively. Similarly, SSIM descriptors are also quantized to a bag-of-words representation.

GIST is a global visual feature, capturing mainly the global texture distribution in a video frame. Specifically, it is computed based on the output energy of Gabor-like filters (8 orientations, 4 scales) over a 4×4 image grids (Oliva and Torralba 2001). The final descriptor is 512-d ($8 \times 4 \times 4 \times 4$). We use the averaged GIST descriptor over the sampled frames to represent a video.

MFCC: Research in neuroscience has found that multiple senses work together to enhance human perception (Stein and Stanford 2008). Audio information is intuitively a very important clue for our work. The first audio feature we consider is the well-known mel-frequency cepstral coefficients (MFCC), which are computed for every 32ms time-window (an audio frame) with 50% overlap. Multiple MFCC descriptors of a video are mapped to a bag-of-words representation, which is the same as we quantize the visual features.

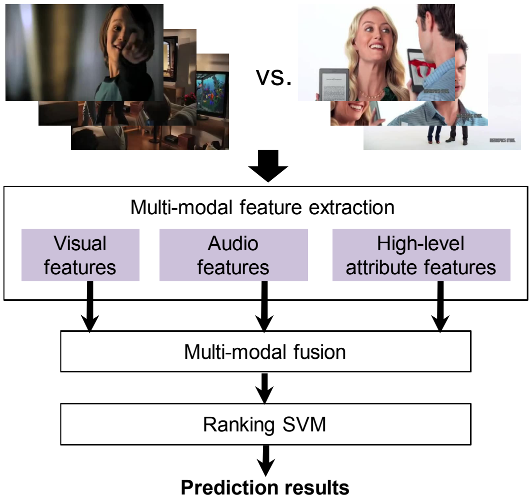

Figure 2: Overview of our method for predicting video interestingness.

Spectrogram SIFT is an audio feature motivated by recent computer vision techniques. First, an auditory image is generated based on the constant-Q spectrogram of a video's sound track, which visualizes the distribution of energy in both time and frequency. SIFT descriptors are then computed on the image and quantized into a bag-of-words representation. Similar ideas of using vision techniques for audio representation were explored in (Ke, Hoiem, and Sukthankar 2005).

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. This six-dimensional feature is expected to be useful as we observed that music seems to have a certain degree of positive correlation with video interestingness.

Classemes (Torresani, Szummer, and Fitzgibbon 2010) is a high-level descriptor consisting of prediction scores of 2,659 semantic concepts (objects, scenes, etc.). Models of the concepts were trained using external data crawled from the Web. Different from the visual and audio features listed above, each dimension of this representation has a clear semantic interpretation (often referred to as an *attribute* in recent literatures). The authors of (Dhar, Ordonez, and Berg 2011) reported that high-level attributes (e.g., scene types and animals) are effective for predicting image interestingness. Like the GIST descriptor, an averaged vector of the Classemes descriptors of all the sampled frames is used to represent a video.

ObjectBank (Li et al. 2010) is another high-level attribute descriptor. Different from Classemes which uses frame-level concept classification scores, ObjectBank is built on the local response scores of object detections. Since an object may appear in very different sizes, detection is performed in multiple image (frame) scales. 177 object categories are adopted in ObjectBank.

Style Attributes: We also consider the photographic style attributes defined in (Murray, Marchesotti, and Perronnin 2012) as another high-level descriptor, which is formed by concatenating the classification outputs of 14 photographic styles (e.g., Complementary Colors, Duotones, Rule of Thirds, Vanishing Point, etc.). These style attributes were found useful for evaluating the aesthetics and interestingness of images, and it is therefore interesting to know whether similar observations could extend to videos.

Among these representations, the first five are low-level visual features, the next three are audio features, and the last three are high-level attribute descriptors. Due to space limitation, we cannot elaborate the generation process of each of them. Interested readers are referred to the corresponding references for more details.

Classifier and Evaluation

As it is difficult to quantify the degree of interestingness of a video, our aim in this work is to train a model to *compare* the interestingness of two videos, i.e., to tell which one is more interesting. Given a set of videos, the model can be used to generate a ranking, which is often sufficient in practical applications. We therefore adopt Joachims' Ranking SVM (Joachims 2003) to train prediction models. SVM is used here because it is the most successful algorithm in many video analysis applications, and the Ranking SVM algorithm designed for ranking-based problems has been popular in information retrieval. For the histogram-like features such as the color histogram and the bag-of-words representations, we adopt the χ^2 RBF kernel, which has been observed to be a good choice. For the remaining features (GIST and the attribute descriptors), Gaussian RBF kernel is used.

Training data are organized in the form of video pairs, with ground-truth labels indicating which one is more interesting for each pair. Since the Flickr dataset does not have stable fine-grained ranking, similar to (Dhar, Ordonez, and Berg 2011), we generate training and test pairs between the top 10% and bottom 10% of videos (the total number of pairs is $600 \times 600/2$). For both datasets, we use 2/3 of the videos for training and the rest for testing. The datasets are randomly partitioned ten times and a model is trained for each train-test split. We report both mean and standard deviation of the prediction accuracies, which are computed as the percentage of the test video pairs that are correctly ranked.

Fusing multiple features is important since many features are complementary. We use kernel-level fusion which linearly combines kernels computed from each single feature to form a new kernel. Equal weights are adopted for simplicity, while it is worth noting that dynamic weights learned from cross-validation or multiple kernel learning might further improve performance.

Results and Discussions

In this section, we train models using the multi-modal features and measure their accuracies on both datasets. We first report results of the visual, audio, and attribute features separately, and then evaluate the fusion of all the features.

Visual feature prediction: We first examine the effectiveness of the visual features. Results are summarized in Figure 3 (a). Overall the visual features achieve very impressive performance on both datasets. For example, the SIFT feature produces an accuracy of 74.5% on the Flickr dataset, and 67.0% on the YouTube dataset. Among the five evaluated visual features, Color Histogram is the worst (68.1% on Flickr and 58.0% on YouTube), which indicates that—while color is important in many human perception tasks—it is probably not an important clue in evaluating interestingness. Intuitively this is reasonable since it is mainly the semantic-level story that really attracts people. Although other features like SIFT and HOG do not have direct semantic interpretations, they are the cutting-edge features for recognizing video semantics.

We also evaluate the combination of multiple visual features. Since there are too many feature combinations to be evaluated, we start from using only the best single feature and then incrementally add in new features. A feature is dropped if it does not contribute in a fusion experiment. As shown in Figure 3 (a), fusing visual features does not always improve the results. In addition to Color Histogram, GIST also degrades the performance, probably because it is a global descriptor and is not robust to content variations like object scale changes.

Audio feature prediction: Next, we experiment with the three audio-based features. Figure 3 (b) gives the results. We see that all the audio features–including the very compact Audio-Six descriptor–are discriminative for this task. This verifies the fact that the audio channel conveys very useful information for human perception of interestingness. Spectrogram SIFT produces the best performance on Flickr (mean accuracy 74.7%) and MFCC has the highest accuracy on YouTube (64.8%). These individual audio features already show better performance than some of the evaluated visual features.

The three audio features are also very complementary, which is not surprising since they capture very different information from video soundtracks. Through combining them, we get an accuracy of 76.4% on Flickr and 65.7% on YouTube.

Attribute feature prediction: The last set of features to be evaluated are attribute-based. As shown in Figure 3 (c), while the attribute features are all effective, they do not work as well as we expected. Compared with the audio and visual features, the results are similar or slightly worse. This may be because the models for detecting these attributes were trained using external data (mostly Web images), which are visually very different from the video data used in our work. The data domain difference largely limits the performance of this set of features. Therefore we conjecture that significant performance gains may be attained with new carefully designed attribute models trained using video data.

Comparing the three attribute features, Classemes and

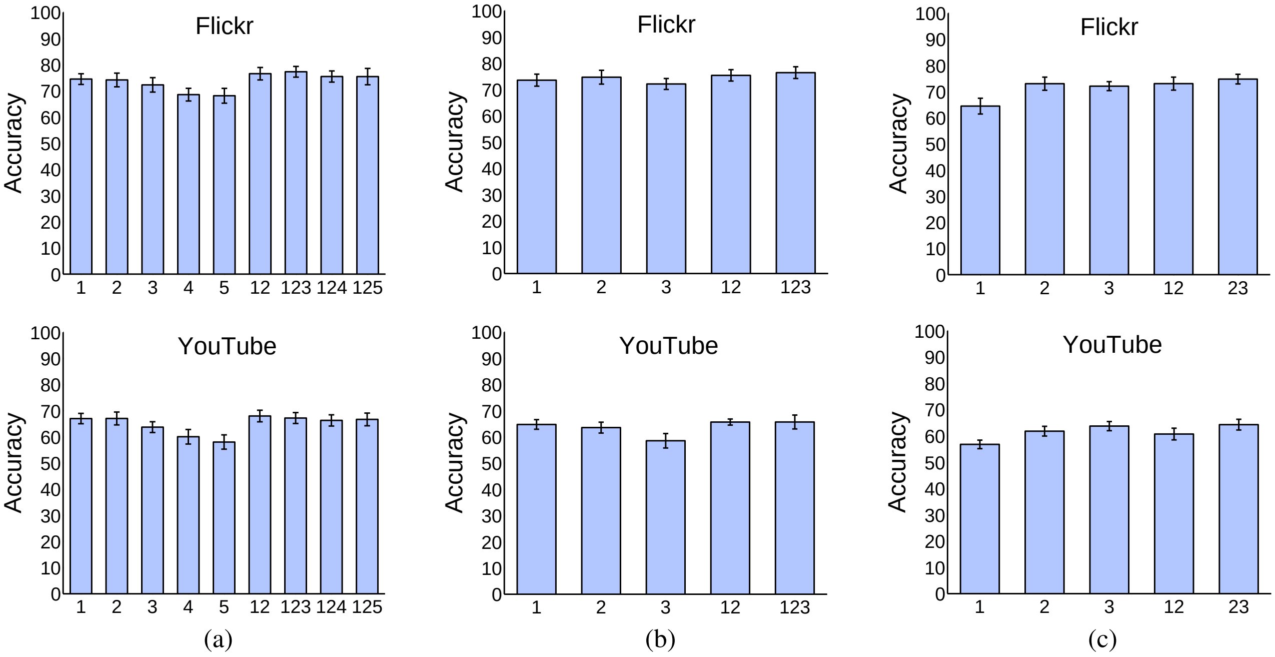

Figure 3: Accuracies (%) of interestingness prediction on both Flickr and YouTube datasets, using models trained with individual features and their fusion. (a) Visual features (1. SIFT; 2. HOG; 3. SSIM; 4. GIST; 5. Color Histogram). (b) Audio features (1. MFCC; 2. Spectrogram SIFT; 3. Audio-Six). (c) Attribute features (1. Style Attributes; 2. Classemes; 3. ObjectBank). Notice that a feature is dropped immediately if it does not contribute to a fusion experiment, and therefore not all the feature combinations are experimented. The best feature combinations are "123", "123" and "23" in the visual, audio and attribute feature sets respectively.

ObjectBank are significantly better than the Style Attributes, which is an interesting observation since the latter was claimed effective in predicting image aesthetics and interestingness (Dhar, Ordonez, and Berg 2011; Murray, Marchesotti, and Perronnin 2012). This verifies the fact that video interestingness is mostly determined by complex high-level semantics, and has little correlation with the spatial and color compositions like the Rule-of-Thirds.

Fusing visual, audio and attribute features: Finally, we train models by fusing features from all the three sets. In each set, we choose the subset of features that show the best performance in the intra-set fusion experiments (e.g., Classemes and ObjectBank in the attribute set, as indicated in Figure 3). Table 1 (the bottom "Overall" row) gives the accuracies of each feature set and their fusion. Fusing visual and audio features always leads to substantial performance gains (76.6%→78.6% on Flickr and 68.0%→71.7% on YouTube), which clearly show the value of jointly modeling audio and visual features for predicting video interestingness.

Attribute feature further boosts the performance from 78.6% to 79.7% on Flickr but does not improve the results on YouTube. This is perhaps because most Flickr videos were captured in less constrained environments by amateur consumers. As a result, contents of the Flickr videos are more diverse, for which the large set of semantic attributes may be more discriminative.

Generally the results on the Flickr dataset are better than that on the YouTube dataset, because the training and test video pairs in Flickr are generated only between top and bottom ranked videos (see discussions in Section "Datasets"), which are apparently easier to predict.

Per-category performance and analysis: We also analyze the prediction accuracy of each category in both datasets, to better understand the strength and limitation of our current method. Results are listed in Table 1. We see that the performance is very good for some of the categories like "mountain" in Flickr and "medicine" in YouTube. Many highly rated "mountain" and "sunset" videos contain background music, which is the main reason that audio features alone already produce very impressive results. In contrast, videos from categories like "graduation" are very diverse and do not have a very clear audio clue. Overall we observe similar performance from the audio and visual features on both datasets. While the fusion of audio-visual features shows clear increases in performance of most categories, we emphasize that the current fusion method, i.e., average kernel fusion, is obviously not an optimal choice. Deeper investigations on joint audio-visual feature design might lead to a big performance leap.

Figure 4 shows some challenging examples. Our method has difficulties in handling the cases where the videos have similar visual and audio clues. For example, both videos in the "food" category contain people sitting around a dining table, while talking about something related to the products. The main difference between these videos is the sense of

Flickr						YouTube					
Category	Visual	Audio	Attribute	Vis.+Aud.	Vis.+Aud.+Att.	Category	Visual	Audio	Attribute	Vis.+Aud.	Vis.+Aud.+Att.
basketball	73.9±11.6	70.5±7.5	64.8±8.3	73.5±10.2	**74.2±9.6**	accessories	**77.8±8.6**	63.6±8.8	73.8±7.4	74.0±9.1	74.9±7.5
beach	**87.3±5.1**	77.0±4.1	81.0±7.5	82.6±4.8	83.3±5.3	clothing&shoes	67.3±9.8	73.3±9.0	53.3±5.9	**76.0±7.3**	73.1±7.9
bird	74.3±6.9	78.6±8.4	75.3±6.5	80.9±7.9	**82.4±8.3**	computer&website	71.8±4.5	70.7±8.8	70.9±9.5	72.2±8.7	**74.9±7.0**
birthday	86.9±4.5	83.8±6.6	80.8±4.5	87.1±6.4	**88.0±5.7**	digital products	66.2±9.5	61.6±15.5	71.8±11.9	74.4±7.5	**75.6±8.7**
cat	**68.7±8.1**	63.4±10.2	60.3±7.3	64.9±7.6	66.9±7.0	drink	70.4±13.0	73.3±13.0	60.4±12.6	**75.1±10.3**	74.2±10.9
dancing	62.6±11.1	73.2±9.9	**74.0±8.1**	65.6±13.5	70.2±12.6	food	53.6±12.5	57.8±8.5	55.1±8.9	57.8±8.0	**59.8±8.8**
dog	71.9±10.3	72.2±6.0	68.5±6.9	**74.0±7.8**	72.9±6.3	house application	**74.2±7.6**	68.7±9.0	68.2±9.1	74.0±8.8	72.2±7.9
flower	82.6±6.7	82.2±5.2	81.7±5.5	85.5±3.4	**86.2±2.3**	houseware&funiture	60.4±7.6	64.2±8.3	59.8±7.6	**68.9±7.6**	66.7±6.6
graduation	76.4±11.9	69.7±6.1	71.9±11.7	**78.9±7.0**	74.7±8.8	hygienic products	63.3±10.6	72.9±4.7	55.6±10.8	**73.8±8.3**	72.7±6.8
mountain	83.4±5.4	88.1±3.0	81.0±6.6	89.4±3.0	**90.4±2.4**	insurance&bank	**70.4±8.4**	56.2±11.2	64.9±13.3	66.0±12.5	62.9±13.9
music perf.	81.0±5.0	77.4±6.5	80.7±8.5	83.3±6.4	**86.0±4.7**	medicine	72.4±7.9	71.8±7.6	67.1±9.4	74.0±9.3	**76.0±8.1**
ocean	68.4±7.1	76.3±7.4	74.0±9.5	76.2±6.6	**78.0±6.5**	personal care	62.0±11.4	67.8±9.4	59.6±12.4	**72.0±7.6**	71.6±7.6
parade	**77.8±8.1**	72.1±5.9	76.6±6.5	71.3±8.9	76.0±9.1	phone	65.3±8.8	70.7±8.6	67.1±10.0	**75.1±5.8**	73.3±6.2
sunset	71.6±9.6	**82.6±6.8**	69.0±7.2	82.3±7.5	81.3±7.8	transportation	**76.4±7.3**	53.8±11.4	72.7±7.0	70.2±9.4	71.1±8.0
wedding	81.5±8.1	79.2±8.9	82.1±7.1	83.5±7.0	**84.9±6.6**						
Overall	76.6±2.4	76.4±2.2	74.8±1.8	78.6±2.5	**79.7±2.7**	*Overall*	68.0±2.2	65.7±2.7	64.3±2.0	**71.7±3.5**	71.4±2.8

Table 1: Per-category prediction accuracies (%), using visual, audio, attribute features, and their fusion. The best result of each category is shown in bold.

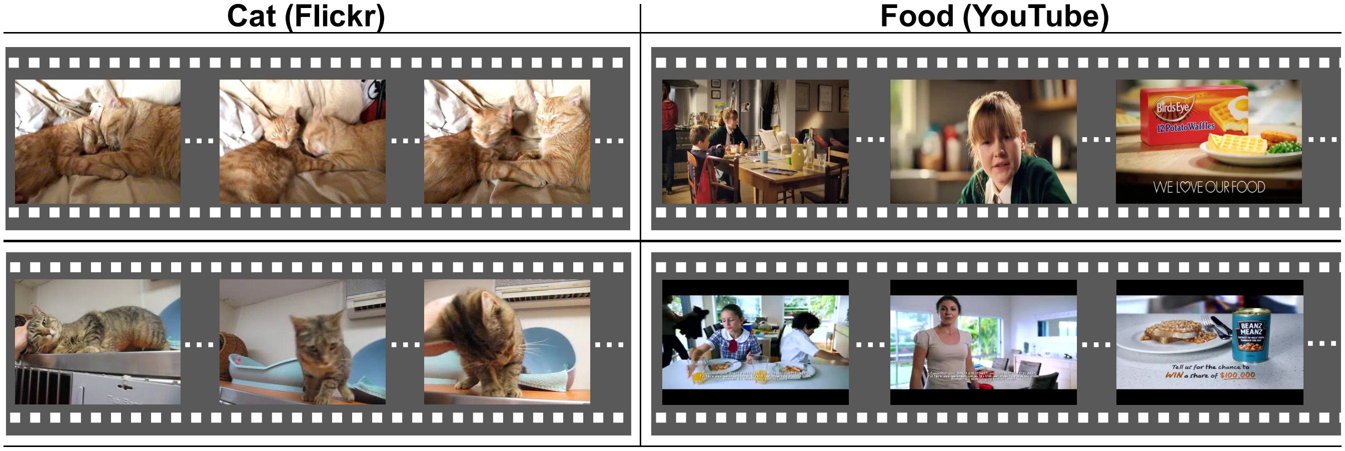

Figure 4: Challenging examples from two categories that have low prediction accuracies using our method. The upper video is more interesting than the lower one according to human judgements, but both of them contain very similar visual and audio features.

humor, or attractiveness of the story, which cannot be captured by our current features. One promising direction to tackle this challenge is to design a better set of high-level visual/audio attributes for interestingness prediction. In addition, speech recognition and natural language processing techniques may be adopted to help understand the conversations, which will be very useful.

Conclusions

We have conducted a pilot study on video interestingness, a measure that is useful in many applications such as Web video search and recommendation. Two datasets were collected to support this work, with ground-truth labels generated based on human judgements. These are valuable resources and will be helpful for stimulating future research on this topic. We evaluated a large number of features on the two datasets, which led to several interesting observations. Specifically, we found that current audio and visual features are effective in predicting video interestingness, and fusing them significantly improves the performance. We also observed that some findings from previous works on image interestingness estimation do not extend to the video domain (e.g., the style attributes are not helpful for this task).

While our initial results are very encouraging, the topic deserves future investigations as there is still space for performance improvement. Promising directions include the joint modeling of audio-visual features, the design of more suitable high-level attribute features, and speech recognition and analysis with natural language understanding. It would also be interesting to deploy and rigorously evaluate the interestingness prediction models in real-world applications like video search and recommendation.

Acknowledgments

This work was supported in part by two grants from the National Natural Science Foundation of China (#61201387 and #61228205), a National 973 Program (#2010CB327900), two grants from the Science and Technology Commission of Shanghai Municipality (#12XD1400900 and #12511501602), and a National 863 Program (#2011AA010604), China.

References

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

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *Proc. of European Conference on Computer Vision*.

Dhar, S.; Ordonez, V.; and Berg, T. L. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

Joachims, T. 2003. Optimizing search engines using clickthrough data. In *Proc. of the ACM Conference on Knowledge Discovery and Data Mining*.

Katti, H.; Bin, K. Y.; Chua, T. S.; and Kankanhalli, M. 2008. Pre-attentive discrimination of interestingness in images. In *Proc. of IEEE International Conference on Multimedia and Expo*.

Ke, Y.; Hoiem, D.; and Sukthankar, R. 2005. Computer vision for music identification. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

Laptev, I.; Marszalek, M.; Schmid, C.; and Rozenfeld, B. 2008. Learning realistic human actions from movies. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

Lerman, K.; Plangprasopchok, A.; and Wong, C. 2007. Peronalizing image search results on flickr. In *Proc. of AAAI Workshop on Intelligent Techniques for Web Personlization*.

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 *Advances in Neural Information Processing Systems*.

Liu, F.; Niu, Y.; and Gleicher, M. 2009. Using web photos for measuring video frame interestingness. In *Proc. of International Joint Conference on Artificial Intelligence*.

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

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*.

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.

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.

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

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*.