Sentiment Flow for Video Interestingness Prediction

Sejong Yoon and Vladimir Pavlovic
Rutgers University
Piscataway, NJ 08854, USA
{sjyoon, vladimir}@cs.rutgers.edu

ABSTRACT

Computational analysis and prediction of digital media interestingness is a challenging task, largely driven by subjective nature of interestingness. Several attempts were made to construct a reliable measure and obtain a better understanding of interestingness based on various psychological study results. However, most current works focus on interestingness prediction for images. While the video affective analysis has been studied for quite some time, there are few works that explictly try to predict interestingness of videos. In this work, we extend a recent pilot study on the video interestingness prediction by using a mid-level representation of sentiment (emotion) sequence. We evaluate our proposed framework on three datasets including the datasets proposed by the pilot study and show that the result effectively verifies a promising utility of the approach.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis, Perceptual reasoning; Representations, data structures, and transforms*

Keywords

Video Interestingness, Temporal Feature, Fisher Vector

1. INTRODUCTION

Can we predict how interesting a video clip is? While the term "interest" can be interpreted as "an emotional state that attracts caution and keep focused", at least three critical questions are much harder to answer: (a) What factors trigger human interest in general? (b) How can we computationally measure the interestingness? and (c) What is a good representation (or feature) of a video that correlates well with the measure of interestingness? Early work by [2] suggests that the interestingness is affected by multiple factors. For example, unusual, complex and surprising events and/or objects will increase the level of interest. Cognitive studies found that emotions can be the source of interestingness [3]. Moreover, it has been reported that some scene categories such as natural environments are more prefered over man-made scenes [3]. Unlike the closely related concept of aesthetic beauty of images [6, 19], the computational prediction of interestingness has not been studied extensively. Based on above psychological findings, several recent attempts were made to directly predict interestingness [7, 22]. One of the most recent works by [10] considers three factors: unusualness (novelty), aesthetics, and general preferences for certain scene types (e.g., outdoor vs. indoor). They found that in a controlled environment such as fixed webcam image sequences, unusualness is important but as the constraints become relaxed, general preference such as the scene category becomes the dominant factor of interestingness.

Many prior studies consider interestingness prediction in the context of images, i.e. static scenes or objects. Few works computationally address the interestingness in image sequences [16, 9], but both regard videos as sets of discrete frames without considering temporal cues or video semantics that dynamically change over time. Recently, [13] conducted a pilot study on the video interestingness prediction. They formulated the prediction problem as a ranking problem between pairs of videos and using the videos collected from Flickr and YouTube, demonstrated that the fusion of multimodal features including low level visual, audio and high level semantic features can effectively predict the relative interestingness of a video.

While promising, the result of [13] leaves several open questions. First, although they argued that the style attribute based features (e.g., color composition or the rule of third) are not as effective as other features, we consider this may be because the variability of those videos is too large for style attribute based features, which are more meaningful in controlled, professionally edited media content, e.g., music video clips. Second, except for audio features, the other features they considdered are static low level features that do not take time into account. Even the audio features such as mel-frequency cepstral coefficients (MFCC) were quantized into a single histogram per video. Therefore, [13] took no explict modeling of time dependent features into account.

We hypothesize that the temporal trend of emotional states can be a factor for the interestingness of a video. Thus, temporal modeling of such trend is the key that may lead to improved representation for the video interestingness. Most methods for affective video analysis, e.g., [26], use low level

HuEvent'14, November 7, 2014, Orlando, FL, USA.

ACM 978-1-4503-3120-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2660505.2660513.

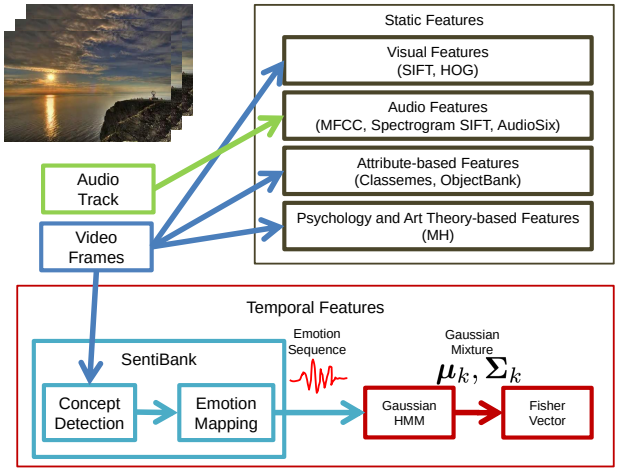


Figure 1: Our Feature Extraction Process for Video Interestingness Prediction.

audio-visual features or learn the feature representation automatically using, e.g., convolutional neural networks [1]. Instead, we adopt a mid-level representation that tends to produce more robust estimates. One such representation was proposed [29], but we modify it in two ways. First, instead of building a dynamic (HMM) model where the number of states matches the number of basic emotions[1], we train a single HMM with a context selected state size and then use the Fisher information score to correlate the flow of sentiment with global relative interestingness for the video sequence, as shown in Fig. 1. Second, we cluster the emotion distribution obtained by SentiBank [4] into the selected HMM states. More details will be provided in Section 2.

It is worth noting that the video interestingness we consider is the interestingness as a holistic measure of the video. Therefore, it clearly contrasts with prior works on video highlight detection methods, e.g. [28]. Moreover, our video interestingness measurement setting is also different from studies using human responses, e.g. [25], as we do not benefit from physiological human responses to measure the interestingness of the video.

In this paper, we (a) extend the study of [13] to get a better understanding on the topic and (b) propose a video interestingness prediction framework that includes a mid-level temporal representation that can effectively capture the interest induced by mixtures of emotions. The rest of the paper is organized as follows. In Section 2, we explain the features we used and introduce our temporal feature based framework for video interestingness prediction. We describe the dataset we used and present the experimental result in Section 3 and draw our conclusion in Section 4.

2. FRAMEWORK FOR VIDEO INTERESTINGNESS PREDICTION

In this section, we briefly explain the low level, semantic, and temporal features employed and then describe the proposed video interestingness prediction framework.

[1]In [29], they defined 5 emotions, *Fear, Angry, Happy, Sad, Neutral*. We consider 8 emotions defined by the Plutchik's wheel of emotions [21].

2.1 Static Features

A large number of static features were considered in [13] and we selected the features reported effective in video interestingness prediction therein. They are categorized into three groups: visual, audio, and semantic. For the visual features, the best combination was identified as SIFT [17], HOG [5] and SSIM [24]. We used only SIFT and HOG because SSIM contributes little to improved performance while being the most computationally demanding of the three. For SIFT and HOG, we used the popular bag-of-words (BoW) representation with 500 codewords. After learning the codebook, each video is represented using histograms of 500 bins. For audio, [13] considered MFCC, Spectrogram SIFT and a set of audio statistics features called Audio-Six, including energy entropy, signal energy, zero crossing rate, spectral rolloff, spectral centroid, and spectral flux. The best combination includes all features and we resort to the same choice. For MFCC, we used a 32ms window and 50% overlap and extracted 12 cepstral coefficients and their first derivatives, resulting in a 24 dimensional feature vector for each video. Spectrogram SIFT is an audio feature that mimics a computer vision feature design approach. It extracts SIFT feature descriptors from constant-Q spectrogram of each video's audio track. For both MFCC and Spectrogram SIFT, the BoW representation using 500 codewords is used to extract the feature vector. Finally, for the semantic features, Classemes [27] and ObjectBank [15] were identified as the best combination in [13]. We used the default parameters provided by the authors to extract the features. For these features, we use the average value over all frames in a video to describe the entire video.

Attributes based on the photographic styles (e.g. color-based features, rule-of-thirds, vanishing points, etc.) were also considered in [13] but were reported not as effective as other high level features. They concluded that the video interestingness is mostly determined by high level semantics rather than low level color-based or spatial attributes. While the claim is reasonable, we suspect that in some cases, e.g. professionally edited music videos, such spatial and color composition based features might help predict interestingness of the videos. Therefore, we consider another set of image based features based on art theory and psychological studies [18]. We will show the utility of this feature in the experiments. A total of 114 features were extracted and we denote this set of features as MH (Machajdik and Hanbury) features for brevity.

2.2 Temporal Features

To capture the trend of emotional states for the video interestingness prediction, we need a computationally measure for emotional states induced by the video. However, to the best of our knowledge, there are few established ways of achieving the task. In this section, by exploiting SentiBank [4], we propose a way to encode the emotional flow as a temporal mid-level feature.

We take concept detection response values of the SentiBank encoded as 1,200 dimensional vector $\mathbf{y}_{i,t}$ for each sequence i of duration T_i. Next, we inverse map the response vectors into the basic emotion distribution using SentiBank. The inverse mapping is available on SentiBank Visual Sentiment Ontology web interface[2]. This will map each 1,200

[2]http://visual-sentiment-ontology.appspot.com

dimensional vector $\mathbf{y}_{i,t}$ into an 8 dimensional emotion vector $\mathbf{x}_{i,t}$. In order to deal with the time-dependent emotional change, we learn a Hidden Markov Model (HMM) using N videos from the training set. Since we do not know the ground truth emotional state change on a frame-by-frame basis, the HMM will act as a mixture model. After learning the means and covariances of the states using an EM algorithm, we build a normalized Fisher vector [20] as the emotional flow representation for each video i.

The key intuition of this approach is based on the notion that in professionally edited videos, e.g., documentary films or music videos, the sequences of frames were deliberately edited to induce emotional movement of the audience, triggering interestingness and keeping the audience focused on the content. Therefore, for such videos, instead of directly predicting and tracking interestingness, we can predict and track the sequence of emotional mixtures to find the sequential emotional state pattern that correlates well with the interestingness of the video.

2.3 Combined Framework

We rely on kernels to combine different static and dynamic features. For BoW representation features, we use the χ^2 kernel and for the semantic features and AudioSix, we use the radial basis function (RBF) kernel. For the temporal feature, we computed the Fisher kernel using the Fisher vectors [12]. Given the set of kernels, the principled way to combine them is by using a multiple kernel learning (MKL) framework [8]. In this work, we used equal weights for all kernels, which is often shown to work very well in practice. We found that in our case, both the sum and the product kernel are equally effective. Specifically, we use the product of kernels for the features of the same "kind", e.g., $\mathcal{K}_{\text{Vis}} = \mathcal{K}_{\text{SIFT}} \circ \mathcal{K}_{\text{HOG}}$, where $\circ$ denotes the Hadamard product, and the sum of kernels to combine different types of features, e.g. $\mathcal{K}_{\text{Combined}} = \mathcal{K}_{\text{Vis}} + \mathcal{K}_{\text{Aud}}$.

For the prediction algorithm, following [13], we employed Ranking SVM [11]. This choice is due to the nature of the current difficulty in exact computation and prediction of the interestingness score. In other words, given a pair of videos, we predict a relative ranking that determines which video is more interesting than the other, instead of directly predicting interestingness as an absolute score.

3. RESULTS

We tested our framework on three datasets. The first dataset is DEAP [14], a collection of video and physiological signal recordings of human subjects watching one minute highlights of music videos. The dataset provides emotion assessment scores for 120 videos collected by self-assessment survey of the participants, each rated by 14-16 people. The rating criterion consists of two criteria: one includes valence, arousal and dominance and the other uses the emotional wheel of named emotion categories [23] and the corresponding intensity of each category. We used the mean rating of the category for each video as the ground truth interestingness. Out of 120 videos, we collected 73 videos that were still available on YouTube at the time of this research. Fig. 2 shows the accuracy of pair ranking for each video, separating the case when the video was chosen as more interesting and the case when it was chosen as less interesting. As one can see in the top row, our proposed method can robustly rank videos regardless of pair choice in most cases.

Table 1: Ranking Accuracy on DEAP Dataset

Features	Accuracy
VisAudAtt	47.6 ± 7.5
MH	53.0 ± 7.2
SentiBank (concept, RBF)	52.2 ± 5.8
SentiBank (emotion, RBF)	55.4 ± 4.6
SentiBank (HMM + Fisher)	**55.8 ± 6.8**
SentiBank (HMM + Fisher / Partial)	53.7 ± 3.6
Features (with VisAudAtt and MH)	**Accuracy**
SentiBank (concept, RBF)	45.1 ± 7.4
SentiBank (emotion, RBF)	44.9 ± 7.2
SentiBank (HMM + Fisher)	48.3 ± 6.2
SentiBank (HMM + Fisher / Partial)	**50.8 ± 4.9**

To compare with prior work, we also used two datasets collected from Flickr and YouTube, introduced in [13]. The Flickr dataset is the top 400 videos retrieved by each of 15 keyword queries using the "interestingness" criterion provided by Flickr service. Only the top 10% and bottom 10% of the 400 videos were selected as interesting and uninteresting samples, resulting in 1,200 videos. The YouTube dataset consist of 30 advertisement video clips in each of 14 categories, totaling 420 videos. For each category, videos were ranked by 10 assessors from 1 (most interesting) to 30 (least interesting).

We used the experimental framework for the video interestingness prediction described in Section 2.3. All visual and semantic features were extracted from every fifth frame from each video. For all datasets, we used 2/3 of videos as training set and the rest for testing, across 20 random splits. We use accuracy as percentage of correct pairwise ranking of test samples for the performance measure. Since the Flickr dataset does not have fine-grained ranking, we used pairs of interesting and uninteresting samples. We found that $K = 8$ for the Gaussian mixture HMM worked best but the result was not overly sensitive to this choice.

3.1 DEAP Dataset

Table 1 shows the pairwise ranking accuracy of the test samples for DEAP dataset. One can see that the proposed `SentiBank`-based emotion features outperformed the original best combination[3]. Since all DEAP dataset videos are professionally edited, emotion inducing stimuli, this verifies our intuition of the relationship between the emotion flow and the interestingness. Interestingly, `MH` features also performed reasonably well on DEAP dataset. This confirms our hypothesis that the reason for the poor performance of the attribute features on the Flickr and YouTube dataset in [13] was the lack of clear emotional flows delierately emphasized in professional videos.

We also observe that when used as isolated feature, `SentiBank`-based methods and `MH` yield comparable results. However, if we combine all features, `SentiBank`-based methods using raw concept detector output degrades drastically. On

[3]In [13], `VisAudAtt`, except SSIM, worked best for Flickr while `VisAud` worked best for YouTube. However, the difference was very small. As we found no significant difference in performance for the two combinations for DEAP dataset, we only report `VisAudAtt` here.

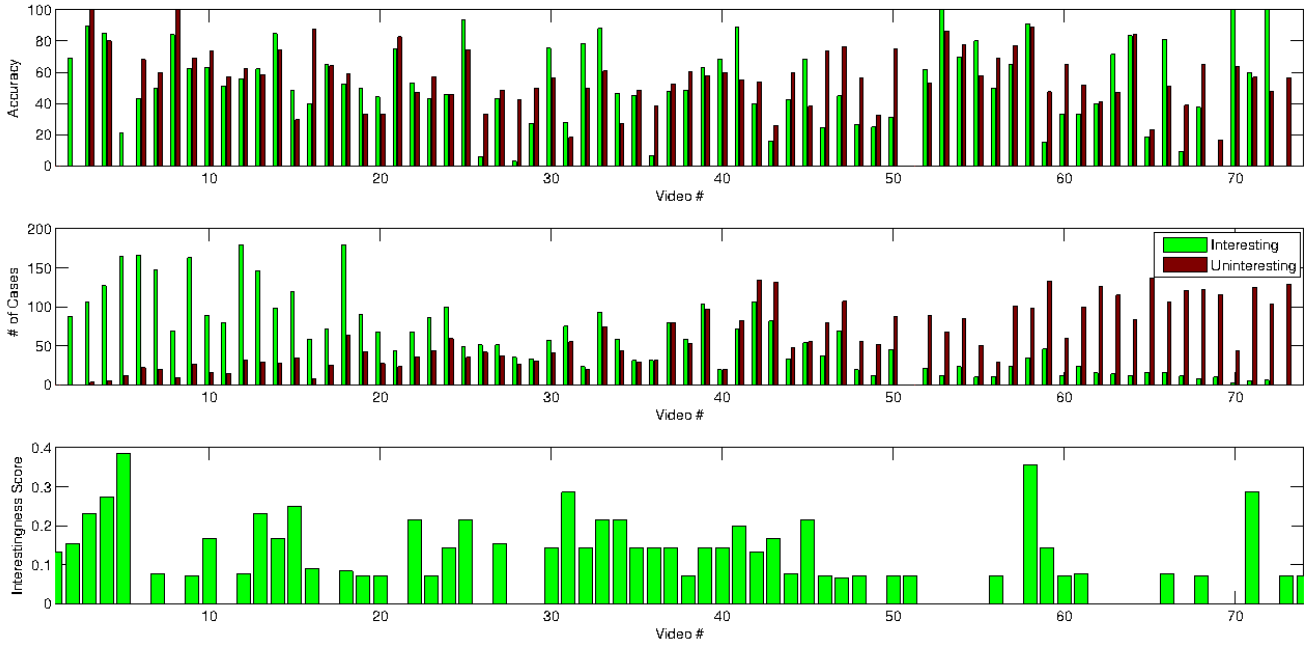


Figure 2: (Top) Accuracy of each video when chosen as test pair for interesting and uninterstingness. Please note that the video ID #51 was omitted due to frame corruption. (Middle) Number of cases chosen for each condition. This shows the general interestingness score of each video, i.e. interesting videos are more tested as interesting pair and uninteresting videos are tested as uninteresting pair. (Bottom) Groundtruth interestingness score.

the other hand, `SentiBank`-based methods using the converted emotion-based representation showed more stable performance. With the Fisher kernel, the mean accuracy remained reasonable 50+% while yielding the smallest variance. Since it is reasonable to assume that we do not know which feature would work well on the new test videos, the robust combination of `SentiBank (HMM)` or emotion mixture model is obviously more preferable.

3.2 Flickr and YouTube Dataset

Table 2 show the result on the Flickr and YouTube datasets. In both datasets, the combinations `VisAudAtt`, `VisAudAtt + MH + Sentibank` and `VisAudAtt + MH + Sentibank (HMM)` show virtually identical performance. This is *not* surprising as the SentiBank is partially dependent on ObjectBank, which is included in `VisAudAtt`. More importantly, unlike DEAP, the information captured by the sequential model of `SentiBank (HMM)` is not prominent in Flickr and YouTube datasets. Flickr videos are user created, thus one cannot expect the emotional flow effect as in DEAP professional music videos. Videos in the YouTube dataset are professionally edited advertisements. However, for the purpose of the advertisement, inducing interestingness here depends more on high level semantic contentext such as the conversation content between actors or the texts displayed on the screen. Thus the amount of emotional flow that `SentiBank (HMM)` can capture is limited. Moreover, since SentiBank concepts are adjective noun pairs (ANP), it is possible that the correctly detected concepts may have a totally opposite meaning in the advertisement. Nevertheless, the combination `VisAudAtt + MH + Sentibank (HMM)` effectively rivals the original `VisAudAtt` and `VisAudAtt + MH + Sentibank` in both datasets. This implies that with our framework one can effectively predict interestingness of professionally emotion induced videos such as music videos in DEAP without diminishing performance in general cases such as Flickr or YouTube datasets. Note that in Table 2, `VisAudAtt` is not the same as [13] possibly because (a) we did not fine-tuned the RBF kernel parameters and always used the rough estimate of the parameter $\gamma = 1/D$ where D is the number of dimensions for the feature type we used and (b) we used linear SVM with precomputed kernel to approximate full kernel-based SVM. Fig. 4 and Fig. 3 depict examples where our approach produces better predictions than the competing methods and Fig. 5 shows an example that our method suffered.

4. CONCLUSION AND FUTURE WORK

In this paper, we describe a video interestingness prediction framework that includes a mid-level emotion flow as an interestingness determinant. We tested our framework on three datasets. In DEAP dataset with all music video clips, we confirmed that emotion flow successfuly captures the sequential pattern that correlates with video interestingness that cannot be easily found by traditional low level features. In contrast, datasets containing amateur videos such as Flickr lack dominant emotional flow, rendering the feature a less significant factor for video interestingness prediction.

Table 2: Ranking Accuracy on Flickr & YouTube Datasets

Category	VisAudAtt	MH	SentiBank	SentiBank +VisAudAtt +MH	SentiBank (HMM)	SentiBank (HMM) +VisAudAtt +MH
Flickr Dataset						
basketball	68.6 ± 10.4	59.9 ± 7.5	45.2 ± 9.0	$\mathbf{71.2 \pm 9.3}$	42.1 ± 11.3	60.6 ± 12.0
beach	74.3 ± 9.6	65.6 ± 11.1	72.7 ± 8.2	$\mathbf{74.3 \pm 8.6}$	71.4 ± 8.3	74.2 ± 8.4
bird	$\mathbf{73.0 \pm 6.9}$	63.3 ± 7.3	67.0 ± 9.7	72.5 ± 7.6	62.0 ± 8.9	69.1 ± 8.3
birthday	70.5 ± 12.2	71.2 ± 11.2	63.3 ± 13.8	$\mathbf{75.9 \pm 11.3}$	66.6 ± 9.5	74.9 ± 8.7
cat	60.9 ± 9.8	55.7 ± 9.1	54.9 ± 8.5	59.9 ± 8.5	58.7 ± 8.8	$\mathbf{61.6 \pm 8.4}$
dancing	66.9 ± 7.3	60.3 ± 12.4	61.0 ± 12.4	67.7 ± 8.9	64.0 ± 11.1	$\mathbf{68.6 \pm 10.1}$
dog	61.0 ± 8.6	44.1 ± 8.2	$\mathbf{68.4 \pm 10.8}$	57.4 ± 9.5	51.8 ± 11.7	58.7 ± 10.1
flower	$\mathbf{83.0 \pm 6.8}$	73.4 ± 9.4	81.3 ± 5.8	82.9 ± 8.4	66.3 ± 10.3	80.7 ± 8.3
graduation	75.1 ± 11.2	64.4 ± 8.0	76.2 ± 8.3	73.6 ± 7.8	79.8 ± 7.3	$\mathbf{81.3 \pm 6.5}$
mountain	79.0 ± 9.0	75.8 ± 9.3	67.9 ± 6.7	82.0 ± 7.8	67.1 ± 9.9	$\mathbf{82.3 \pm 6.1}$
music performance	65.0 ± 8.2	62.2 ± 10.7	59.2 ± 8.7	$\mathbf{65.3 \pm 6.1}$	42.9 ± 10.8	55.3 ± 8.4
ocean	66.1 ± 8.5	52.4 ± 12.1	63.4 ± 7.9	$\mathbf{66.5 \pm 8.9}$	60.5 ± 9.8	63.1 ± 9.6
parade	$\mathbf{70.7 \pm 7.9}$	58.5 ± 10.4	61.8 ± 11.3	68.9 ± 8.7	62.6 ± 9.3	69.2 ± 9.8
sunset	$\mathbf{83.3 \pm 7.2}$	67.5 ± 8.5	69.8 ± 10.0	82.1 ± 6.9	57.9 ± 9.1	78.8 ± 7.6
wedding	75.6 ± 7.5	61.6 ± 8.6	64.2 ± 10.2	75.1 ± 9.2	67.5 ± 5.6	$\mathbf{77.0 \pm 6.8}$
Overall	71.5 ± 7.0	62.4 ± 8.1	65.1 ± 8.7	$\mathbf{71.7 \pm 7.6}$	61.4 ± 10.0	70.3 ± 8.9
YouTube Dataset						
accessories	66.3 ± 11.7	67.6 ± 11.0	62.9 ± 9.8	$\mathbf{69.1 \pm 9.2}$	62.6 ± 12.2	64.3 ± 10.5
clothing&shoes	64.0 ± 12.0	$\mathbf{70.4 \pm 8.6}$	67.7 ± 10.9	69.3 ± 10.4	63.2 ± 10.0	68.1 ± 9.5
computer&website	63.3 ± 10.7	66.0 ± 8.3	66.3 ± 7.9	$\mathbf{68.8 \pm 9.3}$	56.2 ± 11.3	60.8 ± 9.9
digital products	64.4 ± 10.5	62.8 ± 10.3	52.7 ± 11.1	$\mathbf{68.1 \pm 11.0}$	44.0 ± 8.8	58.0 ± 13.5
drink	$\mathbf{63.8 \pm 7.2}$	50.0 ± 8.3	58.0 ± 8.7	60.8 ± 9.7	44.9 ± 11.6	55.7 ± 6.9
food	59.0 ± 10.5	54.1 ± 10.7	56.8 ± 8.1	58.8 ± 10.6	$\mathbf{62.7 \pm 8.9}$	60.8 ± 6.1
house application	51.3 ± 15.0	62.4 ± 10.7	56.8 ± 9.4	58.6 ± 14.2	57.2 ± 8.0	$\mathbf{62.9 \pm 9.3}$
houseware&furniture	$\mathbf{74.1 \pm 7.0}$	57.4 ± 12.9	58.6 ± 10.3	73.2 ± 8.3	59.9 ± 8.3	65.1 ± 7.9
hygienic products	65.0 ± 8.6	65.7 ± 11.7	63.7 ± 8.1	$\mathbf{68.0 \pm 9.6}$	45.1 ± 11.3	53.0 ± 13.5
insurance&bank	53.7 ± 11.3	$\mathbf{61.7 \pm 8.0}$	47.6 ± 11.6	61.4 ± 8.4	49.6 ± 8.1	52.6 ± 10.6
medicine	58.6 ± 12.3	57.9 ± 11.3	56.7 ± 10.2	$\mathbf{61.0 \pm 11.8}$	42.6 ± 7.0	58.3 ± 11.0
personal care	62.6 ± 10.2	50.9 ± 13.6	66.2 ± 10.4	60.3 ± 10.7	60.9 ± 11.1	$\mathbf{63.1 \pm 11.3}$
phone	48.8 ± 10.0	58.6 ± 7.7	50.7 ± 10.6	52.8 ± 10.6	59.0 ± 6.9	$\mathbf{59.7 \pm 8.7}$
transportation	$\mathbf{62.3 \pm 10.2}$	54.0 ± 12.3	59.4 ± 12.9	61.7 ± 11.8	59.4 ± 11.2	58.4 ± 12.0
Overall	61.2 ± 6.6	60.0 ± 6.3	58.9 ± 6.0	$\mathbf{63.7 \pm 5.7}$	54.8 ± 7.8	60.1 ± 4.5

Several investigation avenues remain open. First, static emotion prediction methods other than SentiBank can be considered. While the `MH` features were originally introduced for emotion prediction they were not shown as effective in our experiments so far. Second, instead of using ranking SVM with kernels, one may directly apply ranking HMM such as [30]. This may also be an interesting approach but due to the nature of HMM learning, the size of the training data to learn a reliable HMM in this case would need to be significantly higher.

5. REFERENCES

[1] E. Acar. Learning Representations for Affective Video Understanding. In *ACM MM*, 2013.

[2] D. Berlyne. *Conflict, arousal, and curiosity*. McGraw-Hill, 1960.

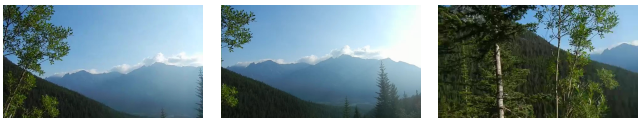
[3] I. Biederman and E. Vessel. Perceptual Pleasure and the Brain. *American Scientist*, 2006.

[4] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *ACM MM*, 2013.

[5] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. In *IEEE CVPR*, 2005.

[6] R. Datta, D. Joshi, J. Li, and J. Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *ECCV*, 2006.

[7] S. Dhar, V. Ordonez, and T. Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. In *IEEE CVPR*, 2011.



(a) Uninteresting example

(b) Interesting example

Figure 3: Easy examples in Flickr dataset: Category is `mountain`. This is an easy category for both low level features [13] and our model. As suggested by [10], natural scenes are highly correlated with interestingness thus they are easier to predict the interestingness than the others categories.

(a) Uninteresting example

(b) Interesting example

Figure 4: Easy examples in Flickr dataset: Category is `graduation`. This is an example category where our method is better than [13]. The interesting video is inducing interestingness by showing the bike riding and the change of surrounding views. On the other hand, the less interesting video shows a lot of people and playing in a natural environment, all seem to positively affect the interestingness, but has little change in semantic context thus lacks emotional stimulus than the interesting one.

(a) Uninteresting example

(b) Interesting example

Figure 5: Hard examples in Flickr dataset: Category is `music performance`. This is one of the hardest category where our method suffered compared to [13]. This category is supposed to be similar to DEAP music videos, but it turns out the semantic gap between the less constrained user created videos and the professionally edited videos were too large.

[8] M. Gönen and E. Alpaydin. Multiple Kernel Learning Algorithms. *JMLR*, 12:2211–2268, 2011.

[9] H. Grabner, F. Nater, M. Druey, and L. V. Gool. Visual interestingness in image sequences. In *ACM MM*, 2013.

[10] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool. The interestingness of images. In *ICCV*, 2013.

[11] R. Herbrich, T. Graepel, and K. Obermayer. *Large Margin Rank Boundaries for Ordinal Regression*, chapter 7, pages 115–132. MIT Press, 2000.

[12] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1998.

[13] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang. Understanding and predicting interestingness of videos. In *AAAI*, 2013.

[14] S. Koelstra, C. Mühl, 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. *IEEE Trans. Affective Computing*, 3:18–31, 2012.

[15] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification. In *NIPS*, 2010.

[16] F. Liu, Y. Niu, and M. Gleicher. Using Web Photos for Measuring Video Frame Interestingness. In *IJCAI*, 2009.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, Nov. 2004.

[18] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010.

[19] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *IEEE CVPR*, 2012.

[20] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[21] R. Plutchik and H. Kellerman. *Emotion: Theory, Research, and Experience Volume 1*. Academic Press, 1980.

[22] T. Schaul, L. Pape, T. Glasmachers, V. Graziano, and J. Schmidhuber. Coherence Progress: A Measure of Interestingness Based on Fixed Compressors. In *Artificial General Intelligence*, 2011.

[23] K. R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44:695–729, 2005.

[24] E. Shechtman and M. Irani. Matching Local Self-Similarities across Images and Videos. In *IEEE CVPR*, 2007.

[25] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun. Affective ranking of movie scenes using physiological signals and content analysis. In *Proc. of the 2nd ACM Workshop on MM Semantics*, MS '08, pages 32–39, New York, NY, USA, 2008. ACM.

[26] R. Srivastava, S. Yan, T. Sim, and S. Roy. Recognizing emotions of characters in movies. In *IEEE ICASSP*, 2012.

[27] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient Object Category Recognition Using Classemes. In *ECCV*, 2010.

[28] J. Wang, C. Xu, E. Chng, and Q. Tian. Sports highlight detection from keyword sequences using hmm. In *IEEE ICME*, volume 1, pages 599–602 Vol.1, June 2004.

[29] M. Xu, J. Jin, S. Luo, and L. Duan. Hierarchical movie affective content analysis based on arousal and valence features. In *ACM MM*, 2008.

[30] Q. Zhang and B. Li. Relative hidden markov models for evaluating motion skills. In *IEEE CVPR*, 2013.