Fast Summarization of User-Generated Videos Using Semantic, Emotional and Quality Clues

Baohan Xu, Xi Wang, Yu-Gang Jiang

Abstract—This paper introduces a novel approach for fast summarization of user-generated videos (UGV). Different from other types of videos where the semantic contents may vary greatly over time, most UGVs contain only a single shot with relatively consistent high-level semantics and emotional content. Therefore, a few representative segments are generally sufficient for a summary, which can be selected based on the segment-level semantic and emotional recognition results. In addition, due to the poor shooting quality of many UGVs, factors such as camera shaking and lighting condition are also considered to achieve more pleasant summaries.

Index Terms—Video Summarization, Semantic Recognition, Emotion Recognition, Quality, User-generated Videos

1 Introduction

A huge amount of videos are being produced and uploaded to the Internet everyday. Techniques for efficiently managing them are becoming increasingly important. Among the many practical needs, video summarization, which generates a short summary clip based on a long input video, is important in applications like efficient search result browsing. Video summaries are especially useful on mobile platforms, because users can preview the critical contents before downloading the entire video, so that the Internet bandwidth may be conserved.

One of the most important goals of video summarization is to produce a clip as short as possible, while preserving the most representative contents in the original video. Most existing works on video summarization focused on the professionally generated videos (PGVs), which are normally very long and contain multiple camera shots (e.g., [10], among others). In this paper, we focus on the summarization of user-generated videos (UGVs). Different from the PGV, the UGVs are normally much shorter and contain a single shot without professional editing. As most UGVs are captured by ordinary consumers with handheld devices like mobile phones, the quality of these videos is diverse and there often exists severe camera shaking. These unique characteristics of the UGVs demand specially designed summarization solutions. In addition, as the amount of the UGVs is extremely large, efficiency is an important factor in order to ensure that the related systems can be easily deployed in real-world applications.

To cope with these challenges, this paper proposes a novel approach for summarizing the UGVs by considering both the representativeness and the quality of the selected segments from an original video (see an example in Figure 1). Specifically, three important clues are considered and integrated: semantics, emotions and shooting quality. As the UGVs contain rich semantic and emotional contents, it is important to preserve both semantic and emotional representative contents. Based on our recent works on fast video content recognition [3], we can quickly recognize a large set of semantics and emotions in an input video. After that, a simple and effective scheme is designed to pick the representative segments that contain consistent semantics/emotions with the whole video. To ensure a good quality of the summary, a few quality measures including both motion and lighting conditions are computed and integrated with the semantic and emotional clues for segment selection.

The main contribution of this work is the unified framework integrating the three clues, and, in particular, the exploration of the semantic and emotional clues in video summarization, which are especially important for the UGVs. For example, in a video of a "birthday party" event, the summary is expected to highlight the main sub-events like singing the famous song and cutting cakes, which are not only semantically related, but also emotionally representative (very likely, "joy" in these scenes). These clues are extracted in a very efficient manner so that the entire system can be executed in real-time. To demonstrate the effectiveness of our approach, we conduct extensive experiments on 200 UGVs with both objective and subjective evaluations. Results show that our approach significantly outperforms alternative methods. This work extends upon a conference version [16] with further exploration of the emotional clues, experiments on more testing videos, and many extra discussions.

The rest of this paper is organized as follows. Related works are discussed in Section 2. We conduct a user study in order to understand what kinds of clues are important in a video summary. Results of the study are summarized in Section 3. The proposed approach is elaborated in Section 4 and experimental results are presented in Section 5. Finally, Section 6 concludes this paper.

- *This work was supported in part by a National 863 Program (#2014AA015101) and a grant from the NSF China (#61572134).*
Baohan Xu, Xi Wang and Yu-Gang Jiang (corresponding author) are with the Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China. Email:{bhxu14,xwang10,ygj}@fudan.edu.cn.

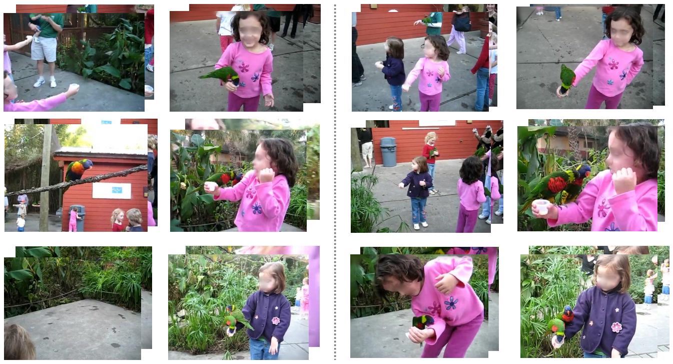

Fig. 1. Sampled frames from (left) an original UGV at 1/24 fps, and (right) the summary generated by our approach at 1/2 fps. The frames are ordered from left to right and top to bottom. Key contents related to the major semantics of this video (children playing with birds) are preserved in the summary. Discernible faces are blurred.

2 Related Works

We divide the discussions of related works into three subsections: video summarization, semantic recognition and emotion recognition.

2.1 Video Summarization

During the past two decades, significant progress has been made on video summarization. A complete review is beyond the scope of this paper and we refer the readers to [14] for works before 2007.

Generally, video summarization approaches can be divided into two groups based on the formats of the outputs. The first group is key-frame based summarization, which generates summaries by selecting a collection of static video frames. Early works in this group selected key-frames by measuring frame changes using features like color histogram [18]. A few researchers also investigated the ways to demonstrate the selected key-frames. For instance, Boreczky proposed an interactive comic book presentation for video browsing [1]. These key-frame based approaches, however, are not sufficient as the most important motion information is fully discarded.

The second group of works is often called dynamic summary, which selects several video segments from a video to produce a short summary clip. In this category, most prior works focused on the PGV. One representative effort is the TRECVid BBC rush video summarization task organized annually during 2005-2008 by the U.S. National Institute of Standards and Technology [10]. Many novel algorithms were proposed in this period. For instance, the authors of [12] introduced a hierarchical structure to model the shot, sub-shots and other level of contents of the BBC rushes videos. Representative shots were selected based on its length and the sum of activity level. Promising results were reported on the TRECVid task in 2008.

Sports videos are another type of PGV that have been popularly adopted in video summarization research. In [13], a unified summarization scheme combining audio and visual features was adopted for both content highlights and play-back scenes in the sports domain. In addition, summarization techniques can also help users quickly acquire the main contents of a movie [2]. Other types of data used in summarization include egocentric videos [8], the UGV (the focus of this work), etc. Khosla et al. [6] focused on summarizing the UGV by selecting key-shots visually similar to Web images, which are considered to contain rich and representative information as people often pick the best moment to take pictures.

Despite numerous efforts in this area, the results are still not satisfactory. The traditional low-level feature based approaches cannot fully capture the semantically meaningful segments. Furthermore, none of these approaches have considered using emotional clues to produce a video summary. As the UGVs often contain rich emotional contents, excluding this valuable dimension in the video summarization process may incur significant loss of precious moments. Our approach in this paper addresses these limitations by integrating multiple clues.

2.2 Semantic Recognition

Recognizing semantic categories in videos is a popular topic in multimedia and computer vision. In this paper, we adopt fast recognition methods so that the final summarization system can be deployed in large scale applications. In [15], the authors showed that the dense local descriptors, like dense SIFT or dense SURF, are as effective as the sparse detector-based local features and can be more efficiently extracted. In addition to the visual features, audio features have been found effective in semantic video recognition [7]. In [3], Jiang proposed a super fast framework using both visual and audio features for video event recognition, which is adopted in this work.

2.3 Emotion Recognition

Previous works on video emotion recognition focused more on the movie domain. For instance, the authors of [11] adopted several visual features to identify the mapping between the features and six movie genres. These works have demonstrated the effectiveness of visual features in emotion recognition. More recently, Jiang et al. [4] used both visual and audio features for emotion recognition in the UGVs instead of movies. This paper adopts a similar approach by replacing some expensive features with more efficient ones.

3 A User Study

We first conducted a user study to determine what kinds of video segments are favored in a UGV summary. Our proposed method is motivated by the observations from this study.

Two hundred UGVs from YouTube were adopted, with durations ranging from 2 to 10 minutes. Ten volunteers with different backgrounds and genders participated in this study. After watching each video, the participants were asked to select some short segments that *are representative and should be included in a summary*. They also need to briefly explain why each segment was selected.

After carefully inspecting all the results, we had the following major observations:

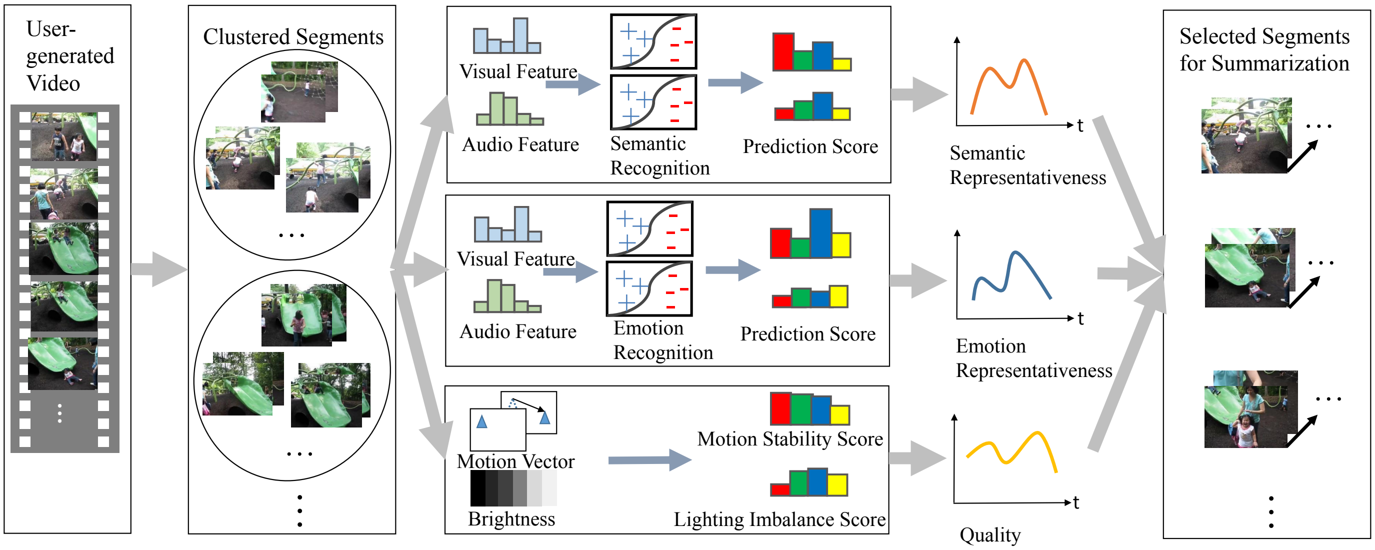

Fig. 2. Illustration of our proposed UGV summarization framework. Given an input video, it is first partitioned into short segments of fixed length, which are clustered into several groups. Semantic and emotion recognition models are applied to generate prediction scores on each segment based on efficient visual and audio features. Motion vector and brightness are also computed for quality evaluation. The three clues are integrated to select the most representative segments to generate the summary.

- Most of the chosen segments have strong connections to the dominate story of the original video. Many selected segments also convey strong emotions, which are also consistent with the dominate emotion of the original videos.
- All the selected segments from a video should, collectively, provide the needed information for users to understand the original story with little redundancy.
- Quality is important. Low quality segments (e.g., with severe camera shaking) should be avoided in the summary.

Considering the dominate semantics is intuitively important, which has been a key clue in several previous video summarization works. The observation of preserving representative emotional moments is interesting, which has rarely been used before. Besides, we also found that the audio clue is important for a couple of segment selections. For instance, some users selected a longer sub-clip that contains musics. However, the desirable length of the UGV summaries is short. Chopping the soundtracks into segments of a few seconds and then combining some selected discontinuous ones will not be understandable by the users. Therefore, audio was not considered in most video summarization approaches, including this work.

4 The Proposed Approach

According to the user study, we design an approach integrating semantic recognition, emotion recognition and quality estimation for UGV summarization. Figure 2 illustrates the whole pipeline. First, we partition the videos into several segments. K-means algorithm is then used to cluster the segments. After that, the three clues are integrated to select the most representative segments to generate the summaries. We introduce the details in the following.

4.1 Video Partitioning and Clustering

We first partition an input video into multiple short segments, among which representative ones are selected to generate the summary. Traditional video summarization methods usually used shot boundary detection to cut the video. However, most UGVs only contain a single shot. Therefore, we simply use uniform segmentation to partition the video (2 seconds each segment). Semantic and emotion recognition are then performed on each segment, which will be elaborated in the next two subsections.

Formally, a UGV is represented as a collection of N segments $\{s_1, s_2, \ldots, s_N\}$, where a segment s_j can be described by a K-dimensional semantic distribution vector e_j, containing prediction scores of K semantic classes. Note that other representations (including the emotion prediction scores) are also feasible here. K-means clustering is employed to group the segments into L clusters $C = \{C_1, C_2, \ldots, C_L\}$. Segment similarity is measured by the χ^2 distance of the semantic vectors, which has been found to be particularly suitable for histogram-like features. Because of the semantic features, segments within the same cluster are semantically similar, but may represent different stage of a story if they are temporally far way from each other. To ensure that we select temporally discontinuous segments but not the temporally adjacent ones within the same cluster, we further partition the L clusters into a set of segment groups $G = \{G_1, G_2, \ldots, G_T\}$, where $T \geq L$ and each group only contains a set of continuous segments from the same cluster. A group may contain only a single segment if the segment does not have temporal neighbors within the corresponding cluster. The generated segment groups are used to ensure a good diversity and coverage of the final summary.

4.2 Semantic Representativeness

Next, we define a criterion that helps us pick at most one segment from each group. The first clue to be considered is representativeness based on the semantic recognition results. As indicated by the user study, a good summary should highlight the major semantics in the original video. Following [3], twenty efficient one-vs-all SVM classifiers (see class names in [3]) are learned using a part of the Columbia Consumer Video database as training data [5]. To convert the video segments into fixed low-dimensional representations, we adopt the dense SURF (Speeded Up Robust Features) and MFCC (Mel-Frequency Cepstral Coefficients) features, which are briefly described below.

- SURF: SURF is a popular local image descriptor. Originally inspired by the well-known Scale Invariant Feature Transform (SIFT) descriptor, SURF can be more efficiently extracted. We quantize the SURF descriptors into a 4000-dimensional bag-of-words representation.
- MFCC: In addition to the visual feature, audio clues are an important complement for video content recognition. The well-known MFCC is adopted in our approach. An MFCC descriptor is computed over every 32ms time-window with 50% overlap. The descriptors from each video segment are also converted to a 4000-dimensional bag-of-words representation using vector quantization.

We adopt late fusion to combine the prediction scores of the visual and audio feature based SVM classifiers to form the semantic distribution vector e_i for each segment. Based on the vector, we define a *semantic representativeness score* of a segment s_i as:

$$\mathcal{E}(s_i) = e_i^T \cdot \mu \times \sqrt{\frac{1}{K}\sum_{j=1}^{K}(e_{i,j} - \bar{e_i})^2}, \tag{1}$$

where e_i^T is the transpose of the semantic vector e_i of s_i, $\mu = \frac{1}{N}\sum_i^N e_i$ is the semantic distribution of the entire video, and $\bar{e_i}$ denotes the average value of all the dimensions of e_i. The first part of this equation, i.e., $e_i^T \cdot \mu$, produces a high score if the segment is semantically consistent with the overall video. The second part prefers semantic prediction scores with a larger variance, which may reflect that the classifiers are more confident as the scores are likely to be either very high or very low. Since the values of the first part and the second part are in different scales, we adopt the product of them as the semantic representativeness score.

4.3 Emotional Representativeness

We also evaluate the representativeness of a segment based on its contained emotions. For this, we adopt similar models to [4], where SVM classifiers are trained to recognize eight emotions in the UGVs. The one-vs-all strategy is adopted to train an independent classifier for each class. The eight classes are selected according to the well-known Plutchik's wheel of emotions, including anger, anticipation, disgust, fear, joy, sadness, surprise and trust. The same dense SURF and MFCC based feature representations are also adopted here for efficiency reason, as we do not need to extract any extra features. We adopt the dataset provided by [4] as training data.

Same as the semantic component, an emotion distribution vector m_i of a segment is computed by fusing the prediction scores of the SVM classifiers trained using the two features. The *emotional representativeness score* of a segment s_i is defined as:

$$\mathcal{M}(s_i) = m_i^T \cdot \nu \times \sqrt{\frac{1}{K}\sum_{j=1}^{K}(m_{i,j} - \bar{m_i})^2}, \tag{2}$$

where m_i^T is the transpose of the emotion vector m_i of s_i, $\nu = \frac{1}{N}\sum_i^N m_i$ is the emotion distribution of the original video, and the average value of all the dimensions of m_i is denoted by $\bar{m_i}$. The higher the emotional representativeness score, the stronger emotion the segment contains and the more consistent the segment is with the entire video.

4.4 Quality Evaluation

The last clue considered in our approach is video quality, which is not easy to be evaluated. We adopt a simple and efficient way to measure quality, defined as:

$$\mathcal{Q}(s_i) = exp(-P_{s_i}/P) + exp(-\Delta L_{s_i}/L_{max}). \tag{3}$$

The first part measures motion stability based on P batches of frames extracted from a segment s_i, each containing three successive frames. P_{s_i} is the number of batches containing severe motion, determined by the angle between the global motion vectors of nearby frames [9]. Since most UGVs are in the H.264/MPEG-4 format, we can directly use the MPEG motion vector, which can be efficiently extracted. For videos not in this format, we may either convert the video format first or employ other methods such as [17] to estimate global motion.

The second part of Equation 3 measures lighting imbalance, where ΔL_{s_i} indicates the range of average brightness of the frames in the segment s_i, and L_{max} is the maximum brightness value of the original video [17]. We underline that, although these two simple measures are not ideal for video quality evaluation, they are generally sufficient for the amateur UGVs. More sophisticated methods can be hardly deployed when computational time is considered as an important constraint.

4.5 Integration

Recall the conclusions of the user study, all the three clues are important for generating a good summary. We adopt simple linear fusion to combine the three components, again for efficiency concern. Formally, the overall representativeness score of a segment is defined as:

$$\mathcal{O}(s_i) = \mathcal{E}(s_i) + \mathcal{M}(s_i) + \lambda \cdot \mathcal{Q}(s_i), \tag{4}$$

where λ controls the influence of the quality score. Equal weights are adopted for the semantic and the emotional clues as both are considered similarly important according to our user study. Based on this measure, the top-t segments from different segment groups in G are selected to form the summary.

5 Experiments

In this section, we evaluate our proposed approach using both subjective and objective measures. We start by introducing the experimental settings.

5.1 Experimental Settings

We randomly select 150 UGVs from the Columbia Consumer Video database [5] (different from the training videos used for the semantic recognition models). The Columbia Consumer Video database [5] also provides annotations of 20 semantic categories[1], which cover several popular events frequently seen in UGVs (e.g., "playing soccer"). 20 classifiers are trained and the prediction scores are used to form the semantic vectors. As the videos in this dataset contain annotations of the semantic categories, using them can also support the objective evaluation of our semantic representativeness measure. In order to evaluate the effectiveness of the emotional representativeness measure, we also select 50 UGVs from the dataset proposed by [4] (different from the emotion recognition training videos). In total, the dataset used in this work contains a total of 200 UGVs.

We empirically fix the weight parameter λ in Equation 4 to 0.5. We also fix both the number of clusters L and the number of selected segments t to 6, leading to a summary of fixed length (12 seconds) for all the UGVs. We will evaluate the effect of summary length later.

5.2 Alternative Methods for Comparison

We compare our approach with three alternative methods:

- k-means: This method simply groups the segments and selects the ones that are the closest to each cluster centroid. The bag-of-words representation based on the visual descriptors is used to represent each segment.
- Story-driven summarization [8]. In this method, three measures called story, importance and diversity scores were defined to summarize very long egocentric videos. A few minor modifications are made in our implementation for the UGVs to control the length of the summary.
- Image-based summarization: We also implemented a system using a similar idea as [6], which assumes that images contain rich and important information, and segments visually similar to a dominate group of images should be selected. For each of the 20 semantic categories, we retrieve 800 images from Google image search. Then the χ^2 distance is used to measure the feature similarity between the images and the video frames.

To ensure a fair comparison, the duration of the summaries generated by these alternative methods is controlled in the range of 8 to 16 seconds, which is similar to ours.

5.3 Results

5.3.1 Subjective Evaluation

Subjective evaluation is the typical way of validating the effectiveness of video summarization techniques. Three sets of subjective evaluations are conducted. First, we investigate the influence of the three clues separately, i.e., the summaries are generated by only considering one of the clues. After that, we evaluate all the possible combinations of the clues to validate the contribution of each clue. Last, we compare our method with the alternative methods.

Ten annotators (5 males and 5 females) are involved in this subjective study, whose ages range from 22–50 with different professions. For each video in our dataset, we show the entire original video to the annotators first, followed by the summaries by different approaches. To ensure a fair evaluation, the summaries are shown in random order with no indications of the used approaches. After viewing these summaries, the annotators are asked to rate the scores measured by integer values from 1 to 5, according to the following four subjective criteria. All the annotators are asked to watch all the 200 videos, and we report the average score from them. Notice that audio soundtrack is discarded in the summary, as discontinuous audio sounds are not meaningful (even noisy in most cases).

- Accuracy: most selected segments in the summary have strong connections to the dominate high-level semantics of the original video;
- Emotion: the summary contains similar emotional contents with the dominant emotion of the original video;
- Coverage: the summary contains sufficient information to understand the full story with little content redundancy;
- Quality: the quality of most selected segments in the summary is good.

As emotion is an important clue but has never been explored in similar context, we define the emotion criterion to validate its effectiveness, which is a new criterion in video summary evaluation. Comparing the accuracy criterion with coverage, the former focuses on the dominating semantics of each UGV while the latter aims at covering more details and minimizing redundancy.

Results of the first set of evaluations are shown in Table 1. Among the three clues, using only the semantic representativeness produces the best performance on the accuracy criterion. The quality clue demonstrates better performance on the quality criterion as expected, but is the worst in terms of other criteria due to its lack of semantic and emotional information. The emotion clue is slightly better than semantics on coverage and emotion. This indicates that emotion is at least equally important to semantics in video summarization, which is consistent with our observations from the user study. Overall, all the three clues have their own advantages, which is appealing as integrating them together may further improve the results.

Next, we report the results of integrating each two clues in Table 2. We see that the results of combining two clues are clearly better than only using one of them, confirming the fact that all the clues are complementary. The best result is from the combination of the semantic and emotional clues.

[1] More information at http://www.ee.columbia.edu/ln/dvmm/CCV/

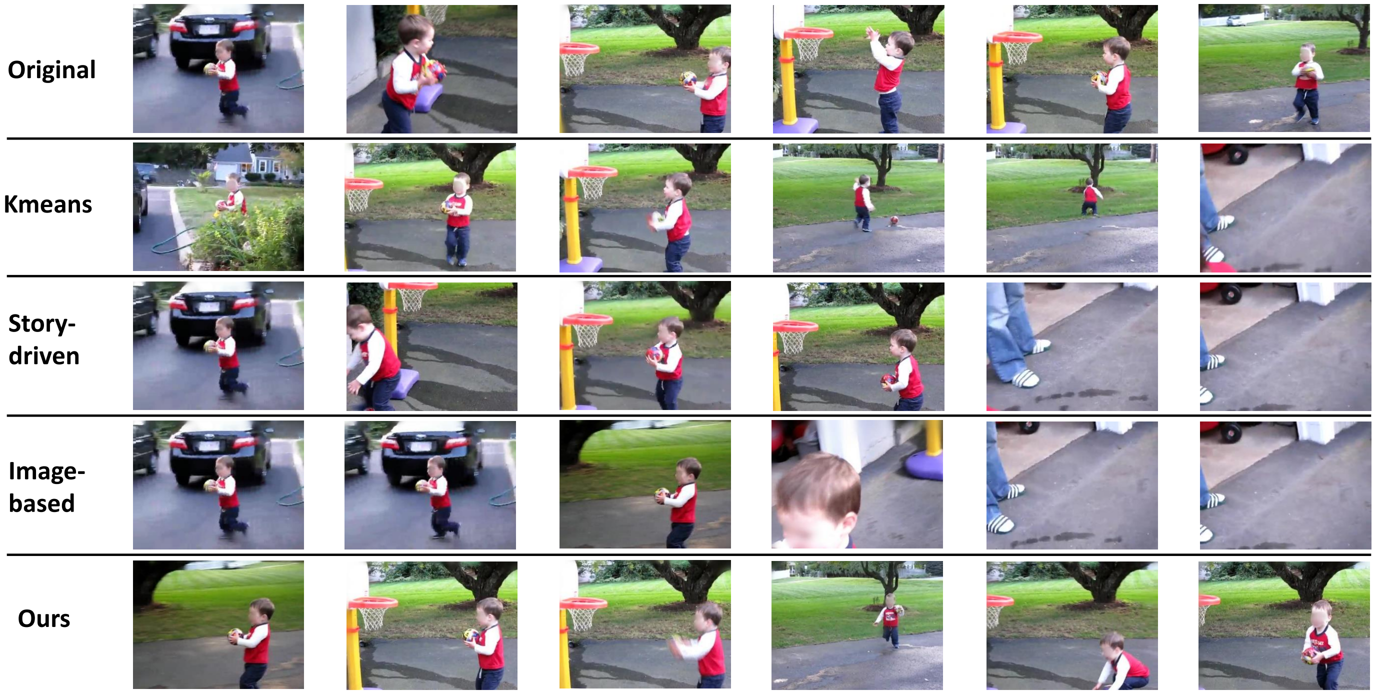

Fig. 3. Example frames of our summary and that of the three alternative methods. The frames are uniformly sampled. Our approach has a clear focus on the major semantics of the video ("kids playing basketball"). The joyful moment of the little boy is also captured. All the three compared alternative methods contain "noisy" segments.

TABLE 1
Subjective evaluation of using the three clues separately. The *overall* row shows the average score of the four criteria. We see that semantics and emotion are almost equally important. See texts for more discussions.

	Semantics	Emotion	Quality
Accuracy	**4.28**	4.23	3.91
Emotion	4.17	**4.32**	3.95
Coverage	3.85	**3.89**	3.44
Quality	3.98	4.01	**4.15**
Overall	*4.07*	***4.12***	*3.86*

TABLE 2
Subjective evaluation of integrating two clues. The *overall* row shows the average score of the four criteria.

	Semantics+ Emotion	Semantics+ Quality	Emotion+ Quality
Accuracy	**4.54**	4.33	4.25
Emotion	**4.41**	4.19	4.36
Coverage	**4.05**	3.94	3.9
Quality	3.96	**4.26**	4.18
Overall	***4.24***	*4.18*	*4.17*

We report the results of combining all the three clues and compare with the alternative methods in Table 3. As shown in the table, our approach produces very strong performance in terms of all the four criteria, outperforming all the three compared methods. Due to the use of the semantic and emotion recognition, our approach achieves very good scores particularly in terms of the accuracy criterion. The story-driven and the image-based approaches focus on different levels of semantics (i.e., objects) and different sources of semantics (i.e., from Web images), respectively. The image-based method is the worst, which may be partially due to the domain difference between the Web images and the UGVs. These results clearly suggest that, for UGV summarization, high-level semantics and emotions are very important. Because we partition of the videos into different segment groups and adopt a hard constraint of selecting maximally only one segment from each group, the coverage of our approach is also better than all the other methods. In addition, the scores of the quality criterion indicate that the simple method defined in Equation 3 works fairly well. The compared approaches do not have the function of measuring visual quality and thus produce lower scores. Figure 3 shows a few example frames of the summarization results.

We also perform *T-test* to validate the statistical significance of the results. The average score and the variance of all the methods are listed in Table 4. The P-values are computed by comparing our method with other baseline methods. We see that our method is superior to the baselines and the result is of statistical significance (P $<$0.001).

5.3.2 Effect of Summary Length

We now evaluate the effect of summary length, which is controlled by two parameters in our method: the number of clusters L and the number of selected segments t. We set both parameters to 4, 6 and 8 separately, leading to the following summary length: 8s, 12s and 16s. Other choices of the parameters are not considered as either very short or very long summaries are not desired in practice.

The subjective evaluation results are listed in Table 5. For the videos containing events with consistent visual contents like playing piano, singing, the short summaries can provide enough information. However, many other videos

TABLE 3
Subjective evaluation of integrating all the three clues (indicated by "Ours") and the alternative methods. The *overall* row shows the average score of the four criteria. Our approach significantly outperforms all the compared methods.

	k-means	Story-driven	Image-based	Ours
Accuracy	4.27	4.41	4.18	**4.68**
Emotion	4.04	3.90	3.79	**4.55**
Coverage	3.64	3.51	3.29	**4.10**
Quality	3.95	3.79	3.89	**4.46**
Overall	*3.98*	*3.90*	*3.79*	***4.45***

TABLE 4
T-test of the subjective evaluation results. The P-values are generated by the T-test of our method in comparison with the alternative baselines.

	k-means	Story-driven	Image-based	Ours
Average	3.98	3.90	3.79	**4.45**
Variance	0.41	0.54	0.70	**0.12**
P-Value	0.000	0.000	0.000	

with more story highlights like playing soccer may need longer summaries to cover the main contents. Therefore, the 8s summary has the worst accuracy, emotion and coverage scores. The result of 16s is similar to that of 12s, with better coverage score as expected but slightly worse accuracy and emotion scores. As shorter summaries are preferred in most practical applications, we recommend 12s as the suitable option for UGVs.

5.3.3 Objective Evaluation

We now conduct objective evaluations to measure the power of our approach in selecting semantically and emotionally representative segments.

For the semantic clue evaluation, we randomly choose 50 UGVs from the test set, manually annotate the dominate semantics (story) of each video, and then label whether each segment is related to the dominant semantics of its parent video. Based on the semantic representativeness scores computed by Equation 1, the selected segments are ranked and ROC curves are shown in Figure 4, in comparison with two methods. One is the image-based summarization method, which uses Web images of similar high-level semantic categories to locate semantically representative segments. As can be seen in the figure, this method generates almost random results, which is probably due to the fact that a significant data domain gap exists between the Web images and the UGVs. The other compared method is based on the dominant visual appearance of a segment, determined by its similarities to all the other segments. In other words, a segment is representative and important if it is visually similar to a large number of segments in the same video. This method outperforms the image-based approach but is still much worse than our solution. The story-driven summarization method is not compared in this experiment because it only considers object-level semantics.

For the emotional clue, we use the 50 test videos sampled from the dataset of [4]. Each video is labeled with one of the eight emotion categories as its dominate emotion. To support the objective evaluation, we also annotate each segment to determine whether it is related to the dominate emotion. The segments are selected based on Equation 2 and then ranked to plot the ROC curves shown in Figure 5. We compare with a baseline called direct emotion allocation, which directly uses the segment-level prediction scores of the emotion models for segment selection. For example, a segment is more representative in an "anger" video when its prediction score of the "anger" category is high. The curves clearly show that our method is better, confirming the effectiveness of our simple emotional representativeness scoring function. Other alternative methods are not compared here as they do not utilize the emotion clue.

TABLE 5
Subjective evaluation results of different summary length using our method. The *overall* row shows the average score of the four criteria.

	8s	12s	16s
Accuracy	4.34	**4.54**	4.45
Emotion	4.28	**4.43**	4.39
Coverage	3.82	4.07	**4.12**
Quality	**4.32**	4.28	4.25
Overall	*4.19*	***4.33***	*4.30*

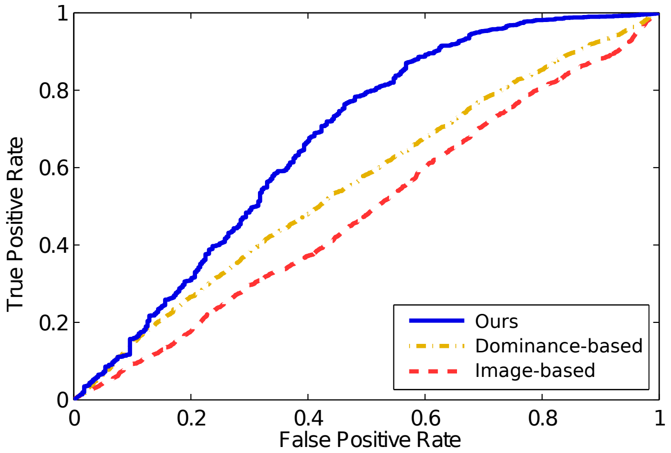

Fig. 4. Comparison on the capability of selecting semantically representative segments.

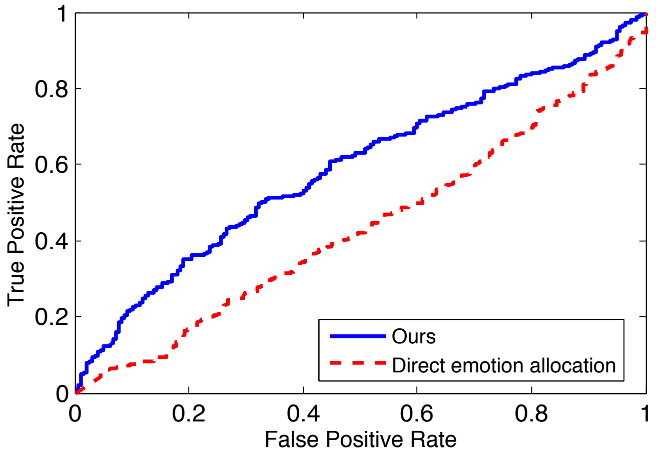

Fig. 5. Comparison on the capability of selecting emotionally representative segments.

TABLE 6
Time cost of the compared approaches to summarize a 2-minute UGV, evaluated on an Intel i7-4770k CPU.

	k-means	Story-driven	Image-based	Ours
Time Cost	11s	56s	123s	15s

5.4 Speed Efficiency

Finally, we discuss the speed efficiency of our approach. The semantic and emotion recognition components share the same set of efficient features. Recognizing both types of clues only requires around 12 seconds for a video of 2-minute duration, including both feature extraction and classification. Notice that the classification part in this recognition process is extremely efficient (a few milliseconds per sample [3]). Therefore, adding more semantic or emotion classes does not significantly increase the computational cost of our approach, which is important in real-world applications. The motion-based quality measure is computed on the motion vectors from MPEG-4, which are extremely fast to be extracted. Integrating all the components together, our approach requires just 15 seconds to summarize a 2-minute duration UGV, using a regular laptop computer.

We compare the speed of our method with the baseline approaches in Table 6. The k-means method is the most efficient because of its simplicity. Our method costs only 4 more seconds than the k-means and is a lot faster than the other two baseline methods. The speed of all these methods are roughly linear to the duration of the input videos. As the UGVs are normally short, our proposed approach is practically very appealing.

6 Conclusions

We have introduced an approach for UGV summarization by integrating multiple clues. High-level semantics and emotions are firstly recognized and simple scoring functions are proposed to select both semantically and emotionally representative segments to form a video summary. The quality of the video segments is also considered to avoid selecting segments containing severe camera motion and poor lighting conditions, which widely exist in the UGVs captured by amateurs with handheld devices.

Both subjective and objective evaluations are performed and results clearly validate the effectiveness of our approach. One important conclusion is that the high-level semantics and emotions are almost equally important in UGV summarization. We also show that the proposed approach is highly efficient, requiring just 1/8 of the video duration to produce a summary.

References

[1] J. Boreczky, A. Girgensohn, G. Golovchinsky, and S. Uchihashi. An interactive comic book presentation for exploring video. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2000.

[2] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi, and Y. Avrithis. Movie summarization based on audiovisual saliency detection. In *Proceedings of the IEEE International Conference on Image Processing*, 2008.

[3] Y.-G. Jiang. SUPER: towards real-time event recognition in internet videos. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2012.

[4] Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user-generated videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014.

[5] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2011.

[6] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[7] K. Lee and D. P. Ellis. Audio-based semantic concept classification for consumer video. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1406–1416, 2010.

[8] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[9] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *Proceedings of the European Conference on Computer Vision*, 2008.

[10] P. Over, A. F. Smeaton, and P. Kelly. The trecvid 2007 bbc rushes summarization evaluation pilot. In *Proceedings of NIST TRECVID Workshop*, 2007.

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

[12] J. Ren and J. Jiang. Hierarchical modeling and adaptive clustering for real-time summarization of rush videos. *IEEE Transactions on Multimedia*, 11(5):906–917, 2009.

[13] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham. Sports video summarization using highlights and play-breaks. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2003.

[14] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1):3, 2007.

[15] J. R. Uijlings, A. W. Smeulders, and R. J. Scha. Real-time visual concept classification. *IEEE Transactions on Multimedia*, 12(7):665–681, 2010.

[16] X. Wang, Y.-G. Jiang, Z. Chai, Z. Gu, X. Du, and D. Wang. Real-time summarization of user-generated videos based on semantic recognition. In *Proceedings of ACM Multimedia*, 2014.

[17] W.-Q. Yan and M. S. Kankanhalli. Detection and removal of lighting & shaking artifacts in home videos. In *Proceedings of the ACM International Conference on Multimedia*, 2002.

[18] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658, 1997.

Baohan Xu received the BS degree from Fudan University, Shanghai, China, in 2014. She is now pursuing the MS degree of Computer Science at Fudan University. Her research interests include multimedia and computer vision.

Xi Wang received the BS degree from Fudan University, Shanghai, China, in 2014. He is now pursuing the MS degree of Computer Science at Fudan University. His research interests include computer vision and deep learning.

Yu-Gang Jiang is an associate professor of computer science at Fudan University, China. His research interests include multimedia retrieval and computer vision. His work has led to many awards, including the 2015 ACM SIGMM Rising Star Award. He received a PhD in Computer Science at City University of Hong Kong in 2009 and spent three years working at Columbia University before joining Fudan.