Real-Time Summarization of User-Generated Videos Based on Semantic Recognition

Xi Wang[1], Yu-Gang Jiang[1], Zhenhua Chai[2], Zichen Gu[2], Xinyu Du[2], Dong Wang[2]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[2]Huawei Technologies, Beijing, China
{xwang10, ygj}@fudan.edu.cn
{chaizhenhua, guzichen, duxinyu, dave.wangdong}@huawei.com

ABSTRACT

User-generated contents play an important role in the Internet video-sharing activities. Techniques for summarizing the user-generated videos (UGVs) into short representative clips are useful in many applications. This paper introduces an approach for UGV summarization based on semantic recognition. Different from other types of videos like movies or broadcasting news, where the semantic contents may vary greatly across different shots, most UGVs have only a single long shot with relatively consistent high-level semantics. Therefore, a few semantically representative segments are generally sufficient for a UGV summary, which can be selected based on the distribution of semantic recognition scores. 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. Experiments on over 100 UGVs with both subjective and objective evaluations show that our approach clearly outperforms several alternative methods and is highly efficient. Using a regular laptop, it can produce a summary for a 2-minute video in just 10 seconds.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding - Video analysis

Keywords

Video Summarization, Semantic Recognition, User-Generated Videos

1. INTRODUCTION

As the amount of user-generated videos (UGVs) grows rapidly on the Internet, techniques for efficiently managing them are becoming increasingly important. For example, summarizing a UGV into a short clip is needed in many applications. One straightforward use-case is efficient content browsing. By viewing the summaries, people can understand the major contents of the UGVs very quickly.

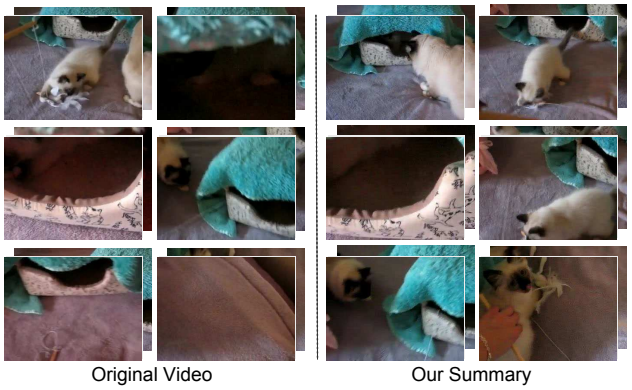

Figure 1: Sampled frames from (left) an original UGV at 1/24 fps, and (right) the summary generated by our approach at 1/2 fps. The frames are ordered from left to right and top to bottom. Key contents related to the major semantics in this video (cat playing) are well preserved in the summary.

Typically, there are two categories of video summarization approaches, divided based on the formats of the outputs. The first one is key-frame based summarization, also known as static storyboard summarization, which extracts a collection of static key-frames to form a summary [4, 11]. In contrast, the second category, dynamic skimming summary, produces a short clip consisting of several video segments selected from the original video [8]. In this paper, we propose a semantics-driven UGV summarization approach, which belongs to the latter category.

The major contribution of this work is an approach that harnesses semantic recognition results to select representative segments. Starting from efficiently recognizing video semantics, we define a simple scheme to pick the segments that have relatively more reliable recognition results, which are also consistent with the dominate semantics of the entire video. In addition, because the UGVs are mostly captured by amateur users with handheld devices, the quality of the videos vary and camera shaking may happen frequently. Therefore, our approach also integrates simple factors that judge the quality of the segments. On a set of over 100 UGVs downloaded from YouTube, we conduct both subjective and

MM'14, November 03 - 07 2014, Orlando, FL, USA

http://dx.doi.org/10.1145/2647868.2655013.

objective evaluations to validate the effectiveness of our approach, and compare with a few recent alternative methods.

In the literature, research on video summarization was mostly done on professionally produced videos, which are very long and consist of multiple shots. Many works relied on low-level information like frame differences [11] or motion cues [8], which can hardly generate satisfying results on the UGVs, partly due to the poor shooting quality of these videos. A few recent studies started to use high-level features such as important objects [5], object tracks [4] or tag localization [9] to address this problem. In [5], Lu and Grauman modeled the influence between sub-events of egocentric videos and designed a method to find a chain of video segments that conveys a fluid story. Wang et al. [9] integrated video tag localization into event-driven video summarization, and Mei et al. [7] proposed a method to compress long videos while maintaining the maximum amount of information. Very recently, Khosla et al. [3] focused on summarizing the UGVs 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. The key difference of our approach is that we use semantic recognition instead of matching to the auxiliary Web image collections. As will be discussed in the experiments, our approach is very efficient and outperforms the image matching based pipeline with clear margins.

In the following, we first describe a user survey to discover important clues for selecting representative segments in UGV summarization. After that we elaborate the proposed approach and discuss experimental results.

2. A SURVEY

We conducted a user survey to determine what kinds of video segments should be included in a UGV summary. Observations from this study motivated the approach presented in this paper.

We randomly selected 200 UGVs from Youtube with durations ranging from 2 to 10 minutes. Ten human subjects with different backgrounds and gender were selected. After viewing each video, the subjects were asked to mark a few short segments that *are representative and should be included in a summary*. A short justification was also needed to explain why each segment was selected.

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

- The chosen segments should have strong connection to the dominate semantics (story) of the original video.
- 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 should be avoided in the summary.

Besides, we also found that the audio clue was important for a number of segment selections. For example, some selected segments are a part of a longer sub-clip that contains beautiful music. However, the desirable length of the UGV summarization is short. Chopping the soundtracks into segments of a few seconds and then merging some selected discontinuous ones will not be understandable by the users. This is the main reason that audio was not considered in most video summarization approaches, including ours.

3. OUR APPROACH

According to the survey, we design an approach integrating both semantic recognition and quality estimation for UGV summarization. First, the videos need to be partitioned into short segments, among which representative ones will be selected to form the summaries. As most UGVs contain only a single shot, shot boundary detection used in the traditional summarization approaches cannot be deployed here. Therefore we use simple uniform segmentation to partition each video into multiple segments (2 seconds each). Semantic recognition is then performed on each segment, using the fast algorithm proposed in [1]. We use the algorithm because it can recognize tens of categories with high accuracy within only a few seconds.

More formally, a UGV is represented as a collection of N segments $\{s_1, s_2, \ldots, s_N\}$, where a segment s_j is described by a K-dimensional semantic distribution vector e_j, containing prediction scores of K semantic classes. k-means clustering is performed to group these segments into L clusters $C = \{C_1, C_2, \ldots, C_L\}$. The similarity of the segments is measured by the χ^2 distance, which was found to be better than the traditional Euclidean distance. 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 may select temporally discontinuous segments but not the temporally adjacent ones within the same cluster, we further split the 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.

Next, we define a criterion that helps us pick at most one segment from each group. The first factor to be considered is representativeness based on the semantic recognition results. As indicated from the survey, a good summary should highlight the major semantics in the original video. To this end, we define a semantic score of a segment s_i as:

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

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., $e_i^T \cdot \mu$, ensures a high score of a segment if it is semantically consistent with the overall video. The remaining part prefers semantic prediction scores with a larger variance, which indicates that the classifiers are more confident as the scores are either very high or very low (far from the decision boundary).

The second factor considered in our approach is video quality, which is defined as follows:

$$\mathcal{Q}(s_i) = exp(-P_{s_i}/P) + exp(-\Delta L_{s_i}/L_{max}). \qquad (2)$$

The first part evaluates motion stability, based on P batches of frames extracted from a segment s_i, each of which contains 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 [6]. Since most UGVs are in the H.264/MPEG-4 format, we can directly use the MPEG motion vector. For those not in this format, we may either convert the format first, or use alternative

Figure 2: Visual comparison of our summary with three methods. The original video is sampled at 1/12 fps and the summaries are sampled at 1 fps. Our approach has a clear focus on the major semantics of the video ("playing baseball"), while the others contain several "noisy" scenes.

methods such as [10] to estimate global motion. The second part of Equation 2 evaluates lighting imbalance, where ΔL_{s_i} measures 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 [10].

Integrating the discussions above, the overall score of a segment is defined as:

$$\mathcal{O}(s_i) = \mathcal{E}(s_i) + \lambda \cdot \mathcal{Q}(s_i), \quad (3)$$

where λ balances the influences of the semantic score and the quality score. Based on this measure, the top-t segments from different segment groups in G are selected to form the summary.

4. EXPERIMENTS

In this section, we evaluate our proposed approach and compare with alternative methods. We randomly select 150 UGVs from the Columbia Consumer Video database [2], which also provides annotations of 20 semantic categories, covering several popular events frequently seen in UGVs (e.g., "playing baseball"). Following [1], 20 efficient classifiers are trained using dense SURF (Speeded Up Robust Features) and MFCC (Mel-Frequency Cepstral Coefficients) features, and we use the prediction scores of these classifiers to form the semantic distribution vector e for each segment (i.e., $K = 20$). A part of the Columbia Consumer Video database [2] is adopted in this training process. For the balance parameter λ, we empirically set it to 0.5. We set both the number of clusters L and the number of selected segments t to 6, leading to a summary of fixed length (12 seconds), which was observed to be suitable in applications like fast video browsing on mobile platforms.

4.1 Compared Approaches

Three approaches are adopted for comparison:

- k-means: This method simply clusters the segments and selects those closing to each cluster centroid. The bag-of-words representation based on the SIFT features is used to represent each segment.
- Story-driven summarization [5]. Three measures called story, importance and diversity scores were defined to summarize very long egocentric videos. A few minor modifications were made in our implementation to control the length of the summary.
- Image-based summarization: We also implemented a system using a similar idea from [3], which, as mentioned in Section 1, 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 retrieved 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.

The video summaries generated by these approaches range from 8 to 16 seconds, which are similar to that generated by our approach.

Table 1: Subjective evaluation results. The *overall* row shows average scores of the three criteria. Our approach outperforms all the compared methods.

Approach	k-means	Story-driven	Image-based	Ours
Accuracy	3.868	3.960	4.061	**4.308**
Coverage	3.556	3.627	3.773	**3.946**
Quality	3.595	3.596	3.703	**3.911**
Overall	*3.673*	*3.728*	*3.846*	***4.055***

4.2 Results and Discussions

4.2.1 Subjective Evaluation

Like most previous works on video summarization, we first adopt subjective evaluation to measure the quality of the video summaries. 10 subjects were involved in this study, whose ages range from 22–50 and have different backgrounds. For each video in our dataset, we show the entire original video to the subjects first, followed by the four summaries (one by our approach and three from the compared approaches). To ensure a fair evaluation, the summaries are shown in random order with no indications of the methods used. After viewing these summaries, the subjects were asked to provide scores (integer values from 1 to 5) according to three criteria: (1) Accuracy: most segments of the summary have strong connections to the dominate high-level semantics of the original video; (2) Coverage: the summary contain sufficient information to understand the original story with little content redundancy; and (3) Quality: the quality of most segments in the summary is good.

The results are summarized in Table 1. As shown in the table, our approach significantly outperforms all the three compared methods. Due to the use of the semantic recognition, our approach achieves very good scores particularly in terms of the "accuracy" criterion. The k-means based method is the worst as it does not consider any semantic information like the other three methods. The story-driven and the image-based approaches, however, focus on different levels of semantics (i.e., objects) and different sources of semantics (i.e., from Web images), respectively. The results

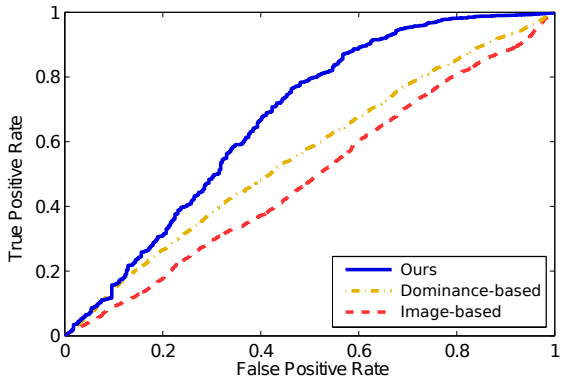

Figure 3: Comparison on the capability of selecting semantically representative segments.

clearly suggest that, for UGV summarization, higher level semantics related to the overall story of the videos are more important. In addition, the scores of the "quality" criterion indicate that the simple method defined in Equation 2 works fairly well. The compared approaches do not have the function of evaluating visual quality and thus received lower scores. Finally, due to the partitioning of the videos into different segment groups and the limitation of selecting maximally only one segment from each group, the semantic coverage of our approach is also better than all the other methods.

4.2.2 *Picking Semantically Representative Segments*

In this subsection, we adopt objective evaluation to measure the power of our approach in selecting semantically representative segments. We randomly selected 50 videos from the test set, determined the dominant semantics (story) of each video, and manually annotated whether a segment is related to the dominant semantics. Equation 1 is then used to measure the scores of the segments, which are ranked and ROC curves are plotted in Figure 3, in comparison with two methods. One of them is the image-based summarization method, which uses Web images of similar high-level semantic categories to locate representative segments. As shown in the figure, this method generates almost random results, which is 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 appearances of a segment, determined by its similarities to all the other segments. In other words, a segment is more representative and important if it is visually similar to a large number of segments in the video. As can be seen from the figure, it outperforms the image-based method but is much worse than our approach. The story-driven summarization method is not compared here because it only exploits lower level object semantics.

4.2.3 *Speed Efficiency*

We conclude this section by discussing the speed efficiency of our approach. The semantic recognition module is based on the work in [1], which only needs around 8 seconds to process of 2-minute video, including both feature extraction and classification. Notice that the classification part in this recognition process is very efficient (a few milliseconds per sample), therefore adding more semantic classes does not significantly increase the computational cost. The quality measure is based on the motion vectors from MPEG-4, which are extremely fast to compute. Therefore, our approach is very efficient. On average, it only requires 10 seconds to summarize a 2-minute UGV, using a regular laptop.

5. CONCLUSIONS

We have introduced an efficient approach for UGV summarization based on semantic recognition. High-level semantics were firstly recognized on segment-level, and a simple criterion was defined to utilize the recognition outputs for selecting semantically representative segments. The quality of the video segments was also considered to avoid selecting segments containing severe camera motion, which widely exists in the UGVs captured by handheld devices.

We conducted both subjective evaluations of the summaries and objective evaluations on the power of selecting semantically representative segments. The results have clearly validated the effectiveness of our approach, and have corroborated a fact that high-level semantics are important in UGV summarization.

Acknowledgement

This work was supported in part by Huawei Technologies, a National 863 Program (#2014AA015101), a Key Technologies Research and Development Program (#2013BAH09F01), and three grants from the Science and Technology Commission of Shanghai Municipality (#13PJ1400400, #13511504503, #12511501602).

6. REFERENCES

[1] Y.-G. Jiang. SUPER: towards real-time event recognition in internet videos. In *Proceedings of ICMR*, 2012.

[2] 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 ICMR*, 2011.

[3] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *Proceedings of CVPR*, 2013.

[4] D. Liu, G. Hua, and T. Chen. A hierarchical visual model for video object summarization. *IEEE TPAMI*, 32(12):2178–2190, 2010.

[5] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proceedings of CVPR*, 2013.

[6] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *Proceedings of ECCV*. 2008.

[7] T. Mei, L.-X. Tang, J. Tang, and X.-S. Hua. Near-lossless semantic video summarization and its applications to video analysis. *ACM TOMCCAP*, 9(3):1–23, 2013.

[8] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. Video summarization and scene detection by graph modeling. *IEEE TCSVT*, 15(2):296–305, 2005.

[9] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua. Event driven web video summarization by tag localization and key-shot identification. *IEEE TMM*, 14(4):975–985, 2012.

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

[11] 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.