Video Summarization via Crowdsourcing

Shao-Yu Wu

Institute of Information Science

Academia Sinica

128 Academia Road, Section 2

Nankang, Taipei 115, Taiwan

derecw@iis.sinica.edu.tw

Ruck Thawonmas

Department of Human and

Computer Intelligence

Ritsumeikan University

1-1-1, Noji-higashi, Kusatsu,

Shiga, 525-8577, Japan

ruck@ci.ritsumei.ac.jp

Kuan-Ta Chen

Institute of Information Science

Academia Sinica

128 Academia Road, Section 2

Nankang, Taipei 115, Taiwan

ktchen@iis.sinica.edu.tw

Abstract

Although video summarization has been studied extensively, existing schemes are neither lightweight nor generalizable to all types of video content. To generate accurate abstractions of all types of video, we propose a framework called Click2SMRY, which leverages the wisdom of the crowd to generate video summaries with a low workload for workers. The framework is lightweight because workers only need to click a dedicated key when they feel that the video being played is reaching a highlight. One unique feature of the framework is that it can generate different abstraction levels of video summaries according to viewers' preferences in real time. The results of experiments conducted to evaluate the framework demonstrate that it can generate satisfactory summaries for different types of video clips.

Keywords

Video summarization, video skimming, crowdsourcing, human computation

ACM Classification Keywords

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems.

Introduction
As online video services are becoming increasingly popular, people are able to watch and share video clips from all over the world with ease. On popular platforms, such as YouTube, billions of videos are watched and hundreds of thousands of videos are uploaded everyday [2]. As the number of videos uploaded each day continues to increase rapidly, users often do not have the time or patience to watch every video clip that looks interesting according to its title and description. However, currently, users have to watch a video completely before they can determine if it is interesting. Therefore, a video summary, like a movie trailer (i.e. a short summary), of each video clip would help potential viewers decide whether a video clip is worth watching.

Existing video summarization methods provide either static story board summaries or dynamic video skimming summaries. A static story board summary is a collection of key frames extracted from the original video. Although this kind of summary can be compiled efficiently, it does not help viewers understand the video content fully since the audio information is missing [4]. Therefore, most recent focus on dynamic video skimming, which yields a short version of the original video, i.e., a selection of video segments.

Although automatic video summarization techniques have been studied extensively, there are no lightweight methods that can summarize any type of video clip effectively. The reason is that it is extremely difficult to develop a program that can understand the semantics of video segments in general. This also explains why several earlier techniques can be only applied to certain types of video clips, such as news, presentation, and sports videos [1].

In this paper, we propose *a simple yet effective general approach that performs video summarization by crowdsourcing*. Because of the above-mentioned difficulty of video understanding, we consider that *human subjects are the best decision makers to judge which parts of a video clip are important*. By including human decision logic, the proposed crowdsourcing approach can be generalized to all types of video clips.

In the remainder of this paper, we present our crowdsourcing design, Click2SMRY, which can capture users' preferences about every segment of a clip. In addition to its generalizability, the scheme is unique because it can generate different abstractions of a video summary based on viewers' preferences. Via experiments, we show that our approach can generate dynamic video skims that yield acceptable levels of user satisfaction.

Design
We have developed a framework called Click2SMRY, which leverages the wisdom of the crowd to generate video abstractions. In our implementation, workers receive a single instruction before they watch a video: "*Click the SPACE key whenever you feel that a particular part of the video should be included in a trailer*." This kind of *fire-and-forget* user feedback mechanism was used in [5,6] to evaluate the QoE (Quality of Experience) of audio material, videos, and games. Since workers only need to perform an intuitive click action to select video highlights, they do not need
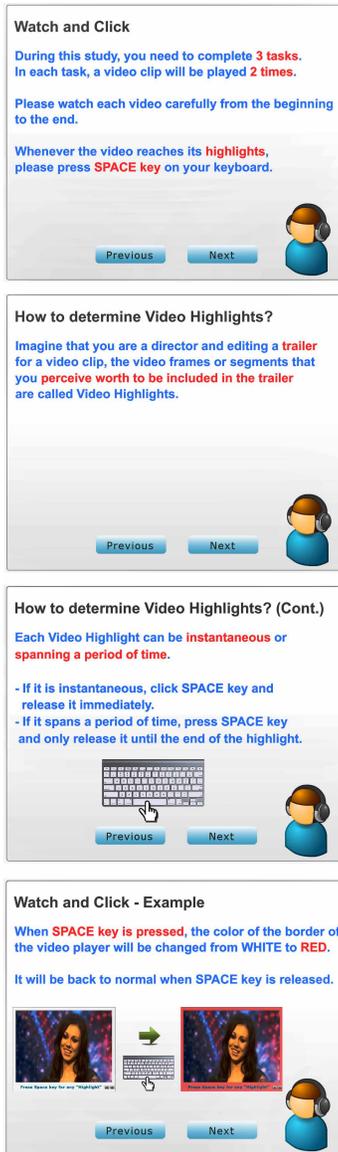
Figure 1. The instructions shown to the participants in the first part of the experiment

to be well-trained or have any experience in video editing.

After collecting a reasonable number of samples (i.e., timestamps when workers press the SPACE key) for a video clip, the Click2SMRY framework can generate a video summary of any specified length based on the collected samples. Assuming the video clip is comprised of n non-overlapping video shots, the k samples collected are denoted by $t_1, t_2, ..., t_k$; and the length of the video summary specified by the viewer is denoted by T. We generate a video summary based on the following algorithm:

1: compute the histogram, H, of $t_1, t_2, ..., t_k$
2: $summary = \varnothing$
3: **while** video length $(summary) < T$ **do**
4:     $t_{add} = \arg \max_t H(t), \forall t$ not contained in $summary$
5:     $shot_{add} =$ the video shot containing t_{add}
6:     $summary = summary \cup shot_{add}$
7: **end while**

The output of the algorithm, *summary*, is a set of video shots that comprise the video summarization result. Since a video shot can be arbitrarily long, before running the algorithm, we divide a video shot into successive shots, each of which runs for a maximum of 5 seconds.

Experiment

To evaluate the effectiveness of the Click2SMRY framework, we design and put its Flash-based implementation online to conduct Internet experiments.

Subjects
We recruited 101 participants (63 males and 38 females) from Amazon Mechanical Turk. The age range of the participants was 18 to 64 years old, and their occupations were as follows: IT (21.8%), student (17.8%), homemaker (11.9%), self-employed (9.9%), business (7.9%), service (6.9%), unemployed (5.9%), education (4.0%), management (3.0%), and others (7.9%).

Materials
We selected three video clips as summarization sources from YouTube. The clips covered three genres, namely, news, sport, and commercial; and the length of each video clip was 1.5–2minutes.

Procedure
The 101 MTurk workers who accepted our HIT (Human Intelligence Task) were directed to our Click2SMRY system. Each experiment comprise three steps: 1) read the instructions, 2) watch the video and click the SPACE key whenever a highlight is identified, and 3) complete a demographic survey and obtain a verification code for MTurk submission. When a step is finished, the next step is shown automatically, so participants only need to follow the instructions to complete the whole experiment.

After reading the instructions (see Figure 1) in the step one, participants were led to the main experiment (step two), which involved three tasks. In each task, the participants watched one of the three videos twice in turn with a randomized order; and they were asked to press the SPACE key whenever they identified a highlight. The video highlight, as defined in the instructions, can be momentary or it can span a period

of time. For momentary highlights, participants were asked to press the SPACE key and release it immediately; otherwise, they pressed the SPACE key and only released it at the end of the highlight. To ensure that each participant focused on the experiment, the video player did not have fast forward or backward functions, and the video was paused if the system lost input focus; for example, if the participants opened other web pages or switched to other applications. At the end of the experiment, 661 raw traces had been logged for further analysis.

Evaluation

To evaluate the quality of the video summaries derived by crowdsourcing, we used two summarization methods for comparison: manual extraction by experts and an automatic video summarization scheme. The three videos used as the summarization sources in the experiment were also used in the evaluation. We set the level of compaction of the skims at 15%, so the length of each video summary was between 12 and 18 seconds.
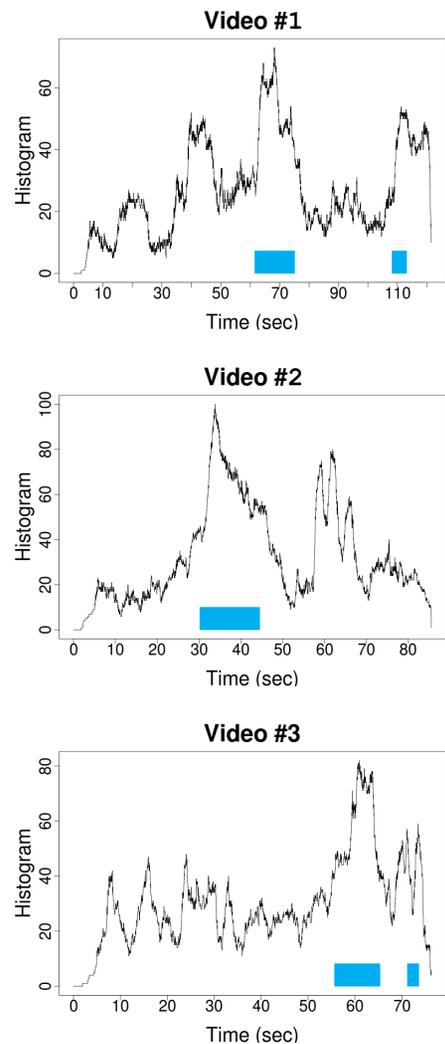
Procedure
Video summaries via crowdsourcing
We used the procedure described in the Design section to generate video summaries for each video clip. Figure 2 shows the summary segments generated from the collected samples (i.e., timestamps when the workers pressed the SPACE key) by using our algorithm.

Video summaries by experts
Three members of a multimedia lab were asked to judge the three videos and manually extract the best segments for summarization. Each judge was familiar with at least one video editing tool. Two of the judges extracted the best segments with their preferred video editing tools. Then, the third judge made the final video summaries by combining the overlapping segments selected by the other two judges, and ensured that the length of each summary was consistent with the required compaction rate.

Baseline video summaries
As our evaluation baseline, we used subsampling, the most basic automatic video skimming scheme, to select segments from each source video [3]. Subsampling selects frames at fixed intervals (e.g., a 5-second shot is selected for every 50 seconds of the original video), and the selected segments are then concatenated to form the video summary at the original frame rate. As the compaction rate used in the study is 15%, the baseline video summary was comprised of the segments between seconds 1–5 of the source video, then between seconds 33.3–38.3, between seconds 66.6–71.6, and so on.

Table 1. The statistics of the three videos used as video summarization sources in the experiment and the evaluation, and the corresponding summary segments derived by crowd-sourcing, experts and subsampling respectively

Video No.	Genre	# of Scenes	Video Length (sec)	SMRY length (sec)	Summary Segments		
					Crowd sourcing	Expert	Baseline
1	news	43	121	18	62–75, 108–113	61–73, 99–105	0–5, 33.3–38.3, 66.6–71.6, 99.9–102.9
2	sport	20	85	15	30–45	30–35, 56–66	0–5, 33.3–38.3, 66.6–71.6
3	commercial	42	76	12	56–65, 71–74	12–16, 59–63, 69–73	0–5, 33.3–38.3, 66.6–68.6

Figure 2. The summary segments highlighted in blue were generated from the collected samples (i.e., the timestamps when the participants pressed the SPACE key) by using our algorithm

Rating

To evaluate the quality of the video summaries generated by Click2SMRY, we conducted a within-subjects study (*N*=81) of the participants recruited from Amazon Mechanical Turk (MTurk). To ensure that each MTurk worker only participated in the study once, we recorded his/her IP address. We also set up a HTTP cookie on each browser so that we could determine if the worker had already visited our system. Moreover, to ensure that the rating result was reliable, we disqualified workers who submitted a wrong unique verification code (the MD5 value of their IP and their browser version information). Also, we discarded data from workers who could not answer the following question correctly: "How many videos did you watch in this study?"

The age range of the final qualified 81 participants (53 males and 28 females) was 18 to 58 years, and their occupations were as follows: student (25.0%), IT (16.7%), self-employed (9.5%), homemaker (9.5%), management (8.3%), business (8.3%), unemployed (4.8%), education (3.6%), service (3.6%), and others (4.8%).

There were also three steps in this phase: 1) read the instructions, 2) watch and rate each video and the corresponding summaries, and 3) complete the demographic survey and obtain a verification code for MTurk submission.

The rating procedure involved three tasks. In each task, the participants watched the three videos in random order, and then watched the three corresponding summaries of each video in random order. The versions of the summaries were "expert," "crowdsourcing," and

"baseline." No fast forward or backward functions were provided on the video player, and the title of each summary only stated "Summary Version #."

After completing each summary, the participants were asked to rate the summary segments on a scale of 1–7 (where 1=very bad, 2=bad, 3=poor, 4=neutral, 5=reasonable, 6=good, and 7=very good) with the following instruction: "Please rate the summary segments that you just watched in terms of how much they helped you make sense of the full video."

Results and Discussion

Figure 3 shows the analysis of the rating results. We observe that the mean score of each video summary derived by crowdsourcing is significantly higher than those generated by the baseline (subsampling) scheme ($p<.05$ using the paired t-test).

The mean scores of the crowdsourcing-generated summaries were either significantly higher ($p<.01$) than those extracted by the experts or there was no significant difference between the mean scores of the two mechanisms for video #1 and video #2. In contrast, the mean score of the crowdsourcing summary in video #3 (commercial) was significantly ($p<.01$) lower than that of the expert version. By analyzing the video summaries, we found that the commercial summary in the expert version consisted of the following segments: prologue, highlight, and brand mark; however, compared to the crowdsourcing version, we found that only the longer highlighted segments were included. This is reasonable because we only asked participants to mark the "video highlights" in the first experiment. Therefore, it was expected that the

Figure 3. Summary of the video summary rating results
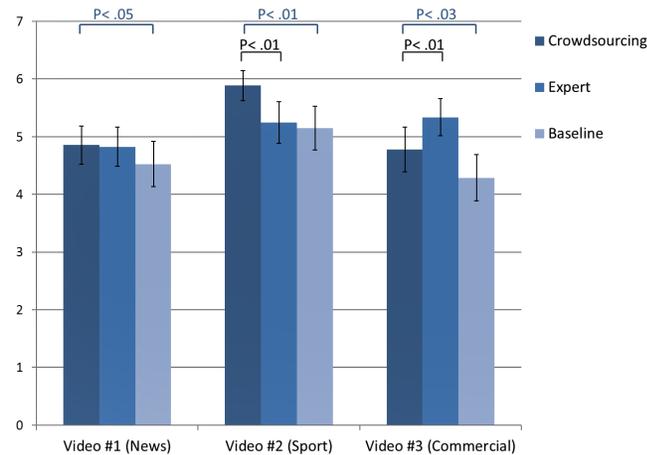
expert version would get a higher score with regard to "video summary."

Conclusion and Future Work

In this paper, we propose the Click2SMRY framework, which leverages the wisdom of the crowd to generate video summaries with different levels of abstraction. Our experiment results show that, though users' feedback is that simple and intuitive, i.e., a click action for an identified video highlight at any time, the resulting video summary is reasonably satisfactory.

Although we paid the workers who helped with video summarization in the experiments, we expect that crowdsourcing could be exploited on a voluntary basis. We believe certain incentive mechanisms, such as a ranking scoreboard of users' contributions, would encourage some users to participate in video summarization voluntarily while they are watching videos, and they would appreciate intangible rewards in the form of community recognition. Then, popular videos on the Internet could be summarized quickly after a certain number of viewings without anyone being paid for the work. In the future, we will investigate how to encourage users to participate in video summarization for non-monetary rewards.

References

[1] Truong, B.T. and Venkatesh, S. Video Abstraction: A Systematic Review and Classification. ACM Transactions on Multimedia Computing, Communications, and Applications, 3(1). (2007), 1–37.

[2] YouTube Fact Sheet. http://www.youtube.com/t/fact_sheet.

[3] Christel, M., Smith, M., Taylor, C., and Winkler, D. Evolving Video Skims into Useful Multimedia Abstractions. In *Proc. CHI '98*, 171–178.

[4] Gao, Y., Wang, W.B., Yong, J.H., and Gu, H.J. Dynamic Video Summarization Using Two-level Redundancy Detection. Multimedia Tools and Applications 42(2), (2009), 233–250.

[5] Chen, K.-T., Tu, C.-C., Xiao, and W.-C. OneClick: A Framework for Measuring Network Quality of Experience. In *Proc. INFOCOM 2009*.

[6] Chen, K.-T., Wu, C.-C., Chang, Y.-C., and Lei, C.-L. Quantifying QoS Requirements of Network Services: A Cheat-Proof Framework. In *Proc. MMSys 2011*.