Harnessing the Wisdom of Crowds: Video Event Detection based on Synchronous Comments

Xingtian Shi, Zhenglu Yang, Masashi Toyoda, Masaru Kitsuregawa
Institute of Industrial Science, the University of Tokyo
4-6-1 Komaba
Meguro-ku, Tokyo 153-8505, Japan
{xingtian, yangzl, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

ABSTRACT

With the recent explosive growth of the number of videos on the Web, it becomes more important to facilitate users' demand for locating their preferred event clips in the lengthy and voluminous programs. Although there has been a great deal of study on generic event detection in recent years, the performance of existing approaches is still far from satisfactory. In this paper, we propose an integrated framework for general event detection. The key idea is that we utilize the synchronous comments to segment the video into clips with semantic text analysis, while taking into account the relationship between the users who write the comments. By borrowing the power of "the wisdom of crowds", we experimentally demonstrate that our approach can effectively detect video events.

Categories and Subject Descriptors: H.3.3[Information search and retrieval]

General Terms: Algorithms, Experimentation.

Keywords: Video search, information extraction, tag analysis

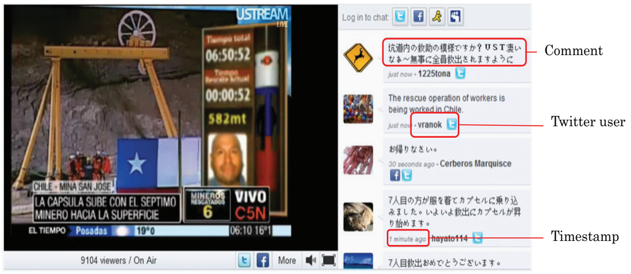
1. INTRODUCTION

Detecting video event effectively is an important issue. Suppose you are looking for some video clips on the event of "miner rescue" (e.g., Fig. 1). You can query the event labels on a web video search engine like Google or YouTube. Unfortunately, the returned videos, although related, are most likely lengthy and contain many uninteresting parts. This will waste your time not only to view, but also to download the data. A reasonable solution is that the programs are segmented into different event clips with semantic labels, then the most similar ones are returned.

Extensive efforts have been devoted to detecting the semantic events, in which most are based on visual features (e.g., color, shape, and motion). Some other features are incorporated later on, such as audio and texture. However, all of these efforts face a challenge that how to bridge the so-called "semantic gap" between the high level and low level features.

To address this issue, closed caption [3] and webcast text have been introduced [8]. While these high level features provide useful information for semantic understanding, there are still several limitations. First, only a small proportion of videos have close caption or webcast, which are mainly in news and sports domains. Second and most important, these features only reflect the perspective of the announcer or the participants in the video.

Inspired by the idea of "the wisdom of crowds" [6], we realize that if multi-viewers can annotate the videos just like what ESP game [7] did, the performance of event detection would be probably improved. Fortunately, there are several video sharing websites (e.g., Ustream [2]) that provide live stream broadcasting and enable viewers, most of whom are Twitter and Facebook users, to comment synchronously.



坑道内の救助の模様ですか？ＵＳＴ凄いなぁ～無事に全員救出されますように
just now · 1225tona

The rescue operation of workers is being worked in Chile.
just now · vranok

お帰りなさい。
30 seconds ago · Cerberos Marquisce

7人目の方が服を着てカプセルに乗り込みました。いよいよ救出にカプセルが昇り始めます。
1 minute ago hayato114

7人目救出おめでとうございます。

Figure 1: The snapshot of a Ustream live broadcast.

To the best of our knowledge, this is the first paper that proposes a general framework taking into account both the synchronous comments and the social network users, to facilitate video event detection. A novel approach is proposed to detect the events. The comments are first grouped into event clusters, taking into account the timestamp, the similarity between words, and the user friendship. Then the videos are segmented accordingly based on the aligned clusters. Preliminary experiments demonstrate that our method can detect semantic events effectively.

2. THE PROPOSED FRAMEWORK

The framework of our proposed strategy is illustrated in Fig. 2. There are three major parts involved with this system: 1) video, synchronous comment, and user information collection; 2) text event detection; and 3) event clip archive.

2.1 Video, Synchronous Comment, and User Information Collection

We collect raw videos with their synchronous comments from major video sharing websites (i.e., Ustream). Preprocessing on the comments is conducted and consists of several steps, such as stop word removal, word tagging, and stemming[1]. The friendship information of a user is obtained by traversing the graphs of major social networks (i.e., Twitter and Facebook), starting from the node of the user and walking n-hop (i.e., 1-hop) to find her friends.

2.2 Text Event Detection

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

[1] The task is implemented by the $Mecab$ toolkit [1].

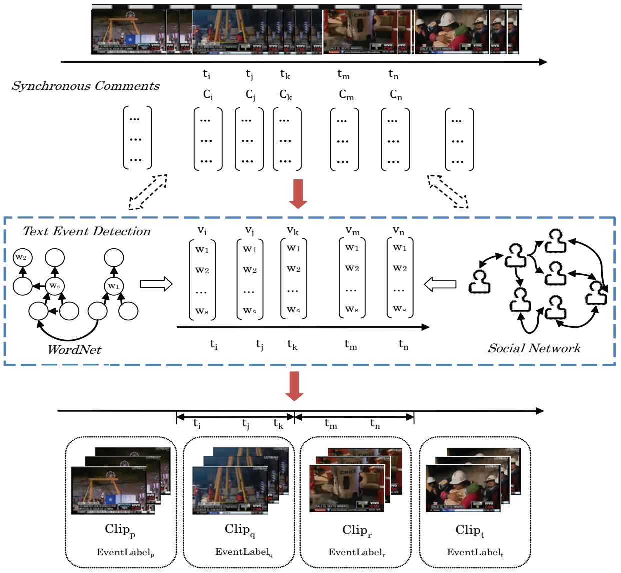


Figure 2: The general framework for event detection

In order to detect video events, we utilize the synchronous comments with regard to three aspects: time lap, similarity between words, and user friendship. Since the effect of time lap is straightforward, that two comments tend to be talking about the same event if their time lap is small, we focus on the remaining two issues: 1) measure of similarity between comments; and 2) how user friendship contributes to grouping comments into event clusters.

Intuitively, the solution for the first issue is to employ a standard vector space model with term weights (e.g., tf-idf). However, the drawback of traditional statistical model is that the semantic meaning of word is ignored. To address this problem, we can aggregate synonymous words defined by external sources (e.g., WordNet) to form a new vector space. Although more elaborate techniques (e.g., LSA [4]) could be applied, they will not be discussed here due to space limitation. We argue that our contribution on introducing "the wisdom of crowds" is stemmed from a new perspective and is orthogonal to existing ones in a complementary manner.

We realize that the user friendship is a typical feature in this scenario. Two users being friends in social network indicates that if they post comments on a single video, the comments tend to be talking about the same event.

We take the above three factors into consideration by ensembling them appropriately. In brief, the similarity of comments is defined as: $Sim(c_i, c_j){=}\alpha{\cdot}e^{-|t_i - t_j|}{+}\beta{\cdot}cos(v_i, v_j){+}\gamma{\cdot}isFriend(u_i, u_j)$, where α+β+γ=1. A text event is consequently detected with regard to the aligned boundary comments, whose similarity is greater than a specific threshold.

2.3 Event Clip Archive

Given the aligned text event, videos can be segmented based on the method in [5]. In addition, we need to archive these clips with index. The reason is that for real applications such as the example aforementioned, efficient label matching strategy is necessary. This task is similar to the traditional issue on document search, where $inverted\ list$ [9] is commonly used. A similar structure is constructed, with each element in the list storing not only the video ID, but also the starting and ending timestamps of the corresponding event clip. Exploration on efficient search algorithm is beyond the scope of this paper. Refer [9] for detail.

3. EXPERIMENT

A total of 30 live videos were crawled from Ustream [2] and 5,893 Twitter users posted 22,656 tweets on these videos synchronously. While there is little previous work on handling different types of videos, our dataset contains videos varying among different content categories, i.e., sports, entertainment, gaming, animals, news, etc.

	BA			SE			CR		
ID	Prec.	Rec.	F-1	Prec.	Rec.	F-1	Prec.	Rec.	F-1
1	0.476	0.567	0.518	0.307	0.620	0.410	0.540	0.605	0.571
2	0.600	0.366	0.454	0.554	0.463	0.504	0.726	0.488	0.583
3	0.916	0.260	0.405	0.933	0.333	0.491	1.000	0.396	0.568
4	0.853	0.924	0.887	0.850	0.964	0.903	0.910	0.929	0.919
5	0.316	0.085	0.134	0.466	0.135	0.209	0.220	0.142	0.172

Table 1: 1="ballot counting", 2="rescue 6th miner", 3 = "hair color", 4 = "Kojima"(*Japanese name*), 5 = "purple beautiful".

We conducted experiments to evaluate the performance of the proposed system on three metrics, i.e., $precision$, $recall$, and F-$measure$. Five event queries are randomly selected from different users interest domains and the results were manually testified. We implemented three methods to detect events: 1) BA, that applies simple vector space model with statistical measure, i.e., tf-idf; 2) SE, that utilizes WordNet to give weight to similar words; and 3) CR, that uses the friend relation from Twitter. In BA and SE, α and β are set to 0.3 and 0.7, respectively, in which way the two models have the best performance. We adjust γ to 0.1 in CR while keeping the ratio of α and β. The similarity threshold is set to 0.3.

The experimental result is shown in Table 1. We can see that compared with that of BA, recall of SE increases significantly yet meanwhile, precision drop-off occurs. This phenomenon reveals the trade-off between the two methods on precision and recall. Moreover, we can see that in most cases, CR overall outperforms the other two approaches (i.e., with higher F_1 score). This demonstrates the advantage on the usage of "the wisdom of crowds".

4. CONCLUSION

An integrated system on general video event detection is developed. As far as we know, our work is the first study on event detection by using synchronous comments and user information. A preliminary experimental evaluation was conducted to confirm the effectiveness of our strategy.

5. REFERENCES

[1] http://mecab.sourceforge.net/.

[2] http://www.ustream.tv/.

[3] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Trans. Multimedia*, 2002.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *American Soc. Inf. Sci.*, 1990.

[5] H. Miyamori, S. Nakamura, and K. Tanaka. Generation of views of tv content using tv viewers' perspectives expressed in live chats on the web. In *ACM MM*, 2005.

[6] J. Surowiecki. *The Wisdom of Crowds*. Doubleday, 2005.

[7] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM CHI*, 2004.

[8] H. Xu and T.-S. Chua. Fusion of av features and external information sources for event detection in team sports video. *ACM TOMCCAP*, 2006.

[9] J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Comput. Surv.*, 2006.