Collaborative Video Scene Annotation Based on Tag Cloud

Daisuke Yamamoto†‡, Tomoki Masuda†, Shigeki Ohira†, and Katashi Nagao†

daisuke@nitech.ac.jp, {masuda,ohira}@nagao.nuie.nagoya-u.ac.jp,
nagao@nuie.nagoya-u.ac.jp
†Nagoya University. Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan.
‡Nagoya Institute of Technology. Gokiso-cho, Showa-ku, Nagoya, 466-8655, Japan.

Abstract. In this paper, we propose a video scene annotation method based on tag clouds. First, user comments associated with a video are collected from existing video sharing services. Next, a tag cloud is generated from these user comments. The tag cloud is displayed on the video window of the Web browser. When users click on a tag included in the tag cloud while watching the video, the tag gets associated with the time point of the video. Users can share the information on the tags that have already been clicked. We confirmed that the coverage of annotations generated by this method is higher than that of the existing methods, and users are motivated to add tags by using tag-sharing and tag-cloud methods. This method assists in establishing highly accurate advanced video applications.

Key words: Collaborative Tagging, Video Annotation, Tag-cloud, Web Service

1 Introduction

The amount of multimedia content available on the Web has been explosively increasing in recent years as an increasing number of video and music clips are being posted and shared in existing video sharing services such as YouTube [1]. Since these video contents include very interesting or valuable information, many users desire to search these contents. The existing video retrieval method in video sharing services is realized by searching metatexts posted by the owner, such as title, description, and tags. Because these metatexts are not associated with video scenes, this method cannot perform video scene retrieval.

Applications summarizing videos or retrieving scenes from them must acquire meta-information about the contents of those videos. Here, we call this meta-information *annotation* [6], and this information can be obtained by extracting keywords related to the video contents. Previous studies either used an automatic method [12] in which metadata were extracted by using image-recognition and voice-recognition technologies or a semi-automatic method [2, 5, 9] in which expert annotators associated video contents with high-quality annotation data by using annotation tools for describing MPEG-7. The accuracy in automatic recognition of the content of amateur videos contents is still not high because they may include noise, indistinct voices, and blurred or out-of-focus images

[1] http://www.youtube.com/

[1]. Automatic recognition technologies are therefore not necessarily effective for amateur videos. Another problem with the audiovisual content created by amateur users is that it is not cost-effective to have experts using semiautomatic annotation tools for associating all the video contents with detailed annotations.

We then developed a video scene annotation system Synvie [13] that allows Web users to voluntarily associate user comments with the video content. Synvie has two annotation methods. One method allows users to submit comments to particular scenes. The other method allows users to write a weblog entry quoted video scenes. Both these methods associate user comments with video scenes. However, a drawback is that the coverage of annotation to video scenes is low.

In this paper, in order to overcome the disadvantages of Synvie, we proposed a tag-cloud-based annotation method that allows Web users to easily increase the coverage of annotations.

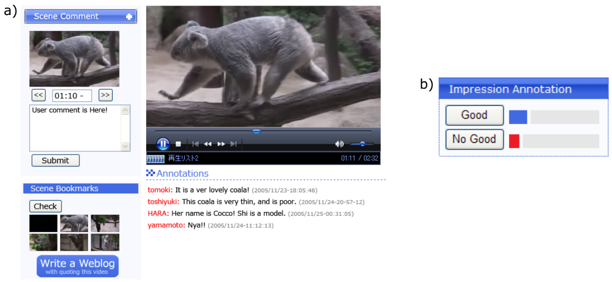
2 Synvie

Let us define "user comment" as a text written by a user when the user watches a video, such as impression, opinion, and evaluation of the video. We assume that this comment is written in the body of a videoblog [7] entry or a bulletin board system. We define a "scene tag" or "tag" as a keyword or a short token associated with a scene. A scene tag consists of a word, a word class, a level of importance, and pointer to scene. A "tag cloud" is the display style of a set of tags, and the font size of each tag is changed corresponding to the level of importance of the tag in order to allow users to easily find objective tags. We define "media time" as the internal time of a video; the starting time of a video is considered as zero media time.

Synvie is a video sharing service that can acquire annotations associated with video scenes. Synvie allows Web users to annotate and share comments while watching a video by using an interface, as shown in Fig.1. At the lower right of Fig. 1 (a), a list of users' comments is displayed, and these users' comments change with the video scenes. Moreover, Synvie allows users to write a Weblog entry for specific video scenes. These user comments include some keywords that express the contents of a scene. We accumulate these keywords as annotations to realize video applications. Users may desire to submit their own opinions, share comments, and write blogs that introduce the video to readers. The difference between Synvie and YouTube is that the former includes a function that allows users to associate comments with "scenes" of a video.

This type of annotation has some problems. First, the coverage of annotations associated with scenes is low, as shown in Fig.2. In the graph of Fig. 2, the vertical axis shows the number of scene tags generated from the user comments associated with the scene, and the horizontal axis shows the scene (divided by 2 s, which we call a scene). The highest number of tags possible is 20 and the lowest number of tags is 0. In short, although impressive scenes have many tags, non-impressive scenes cannot acquire tags. In order to realize advanced video applications, all scenes must have many tags. Therefore, we need to develop an annotation system that allows users to associate annotations flatly.

Synvie allows users to vote for scenes by clicking buttons, which we term as "button annotation," as shown in Fig.1 (b). There are "Good" and "No Good" buttons in order to evaluate the video scenes. Web users can click on these



tomoki: It is a ver lovely coala! (2005/11/23-18:05:46)
toshiyuki: This coala is very thin, and is poor. (2005/11/24-20-57-12)
HARA: Her name is Cocco! Shi is a model. (2005/11/25-00:31:05)
yamamoto: Nya!! (2005/11/24-11:12:13)

Fig. 1. Synvie. a)Comment-based annotation. b)Button-based annotation.

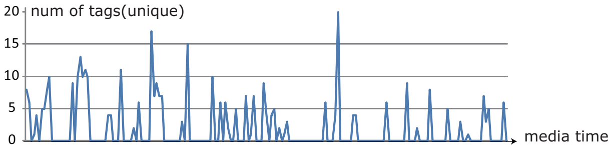


Fig. 2. Distribution of user comments.

buttons repeatedly. Moreover, users can share the results of votes by watching a level meter near each button. In the result, favorite scenes are clicked more frequently. These results may be usual for advanced video applications, such as video summarization. In short, this is the easiest annotation method. In the public demonstration of Synvie, the number of comment-based annotations is 3534 and the number of button annotations is 18854. We believe that it is better to acquire more semantical annotations by using the button annotation method.

At the same time, we developed a video scene retrieval system Divie [3] based on the annotation acquired from Synvie. Divie allows users to select tags from a tag cloud generated from annotations acquired from Synvie in order to input a query.

3 Annotation System

We propose an annotation system that allows users to easily add annotations while watching a video on a common browser. We also combined the button annotation method with the tag-cloud approach in order to acquire annotations effectively. The major advantage of this system is that since Web users add annotations extensively and voluntarily, it is cost-effective.

3.1 System Archtechture

Fig.3 shows the system archtecture.

First, Web users watch videos using video sharing services such as Synvie and YouTube. Then, they submit user comments or blog entries concerning the videos they are interested in. The public demonstration of Synvie [2] involved

[2] http://video.nagao.nuie.nagoya-u.ac.jp/ (Japanese only)

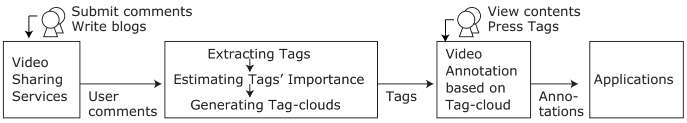


Fig. 3. Structure of our system

324 registered users, 126 videos, and 3534 user comments associated with the videos. Moreover, there are numerous comments and blog entries written by YouTube users on the Web. These comments can generate tag sets by using the tag extraction method. Moreover, a level of importance of each tag is estimated automatically. A tag cloud is generated corresponding to the level of importance of the tags. Web users can click any tag in the tag cloud at any time.

3.2 Annotation Method

Fig.4 shows the interface of the proposed annotation method. There is a video window on the left-hand side of the window, and a tag cloud on the right-hand side of the window. Web users can associate a tag with the media time of a video by clicking the tag in the tag cloud. When a user clicks a tag, the tag is highlighted. The highlight disappears completely in 2 s. Users can click the same tag repeatedly. Moreover, the tags clicked by other users are also highlighted. Therefore, a user can know other users' impressions by sharing the highlighted tags. Expected tags are adjectives (such as "cool" and "cute"), nouns (such as "boy" and "panda"), and verbs (such as "run" and "sleep").

Web users can click any tag at any time. Because a user clicks a tag while watching a video scene, there is some relation between the meaning of the tag and the scene. By associating a clicked tag with the scene, we use this information as the video scene annotation. By accumulating these annotations, we develop advanced video applications such as a video scene retrieval system.

The advantage of this method is that users can add annotations while the video is being played. A user can thus share other users' impressions and opinions. By sharing these impressions, we expect to motivate the users to add more annotations.

The effect of this method is that semantical annotations are associated with video scenes by associating tags with the media time of a video. Moreover, even if an annotation added by a user is too subjective, a majority of annotations added by a majority of users may be objective. By sharing the clicked tags between users, the collaborative annotation [10] is expected. For example, a user may click tags not clicked by other users.

4 Generation of Tag Cloud

Next, we discuss a tag generation method. In general, a tag cloud is statistically generated from tags posted by many users. Because a tag cloud cannot be generated from a single tag, it is necessary for many users to submit tags, which is troublesome. Because there are many user comments associated with videos in video sharing services such as Synvie and YouTube, it is effective when tags are generated from these user comments.

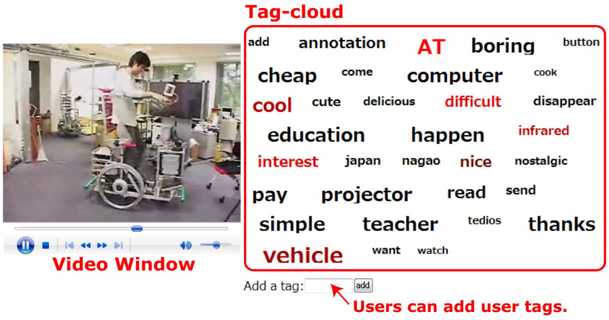


Fig. 4. Tag-cloud based annotation interface. Clicked tags are highlighted.

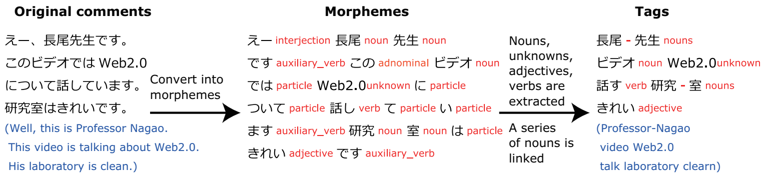

Original comments
えー、長尾先生です。
このビデオでは Web2.0
について話しています。
研究室はきれいです。
(Well, this is Professor Nagao.
This video is talking about Web2.0.
His laboratory is clean.)
Convert into morphemes
Morphemes
えー interjection 長尾 noun 先生 noun
です auxiliary_verb この adnominal ビデオ noun
では particle Web2.0unknown に particle
ついて particle 話し verb て particle い particle
ます auxiliary_verb 研究 noun 室 noun は particle
きれい adjective です auxiliary_verb
Nouns, unknowns, adjectives, verbs are extracted
A series of nouns is linked
Tags
長尾 - 先生 nouns
ビデオ noun Web2.0unknown
話す verb 研究 - 室 nouns
きれい adjective
(Professor-Nagao
video Web2.0
talk laboratory clearn)

Fig. 5. Tag extraction from user comments.

We propose a method to generate a tag cloud from user comments acquired from Synvie. First, tags are generated from user comments. Next, the level of importance of the tags is estimated. Finally, tag clouds are generated from the tags and by considering level of importance. A tag consists of a word, word class, and pointer to video. Since we focus on not only Synvie but also YouTube whose comments are not associated with the media time of a video, we do not use the media time of user comments acquired from Synvie.

4.1 Tag Extraction

We discuss a tag extraction method. Tags are automatically generated from user comments acquired from Synvie.

They are extracted as follows (Fig. 5). Each user comment is first automatically analyzed and converted into morphemes, and subsequently, nouns other than dependent nouns, verbs other than dependent verbs, adjectives, and unknown words are extracted from these morphemes (unknown words are treated as proper nouns) by using the Japanese morphological analysis system ChaSen [4]. The basic form of each morpheme becomes a tag.

Because this method cannot generate tags that consist of compound words, we modified it to convert a series of nouns into a tag. Our system also assumes that comments are mainly written in Japanese; therefore, if a series of alphanumeric characters such as "Web 2.0" appears in Japanese text, the algorithm generates a tag corresponding to the series.

4.2 Estimation of Tag Importance

In general, the font size of a tag in a tag cloud depends on the number of users who submitted the tag. In other words, the more popular a tag, the more important it is. Since our method uses tags generated automatically from user comments, the level of importance of general words such as “man,” “data,” and “thing” is high when we use the common algorithm. We then employ the tf-idf algorithm. Tf-idf is an algorithm that extracts important words from a document. The expansion of tf-idf is given as follows: tf represents term frequency and idf represents inverse document frequency.

The term frequency $tf_{t,c}$ of tag t that belongs to video c is the following.

$$tf_{t,c} = \frac{n_{t,c}}{\sum_k n_{k,c}} \tag{1}$$

where $n_{t,c}$ is the count of tag t that belongs to video c.

The inverse document frequency idf_t of the tag is the following.

$$idf_t = \log \frac{|D|}{|\{d : d \ni t\}|} \tag{2}$$

where $|D|$ is the total number of videos, and $|d : d \ni t|$ is the number of videos associated with the tag t.

Then, $tfidf_{t,c}$ of the tag t that belongs to video c is the following.

$$tfidf_{t,c} = tf_{t,c} idf_t \tag{3}$$

The level of importance of a tag corresponds to the value of tf-idf, provided that the level of importance of an adjective tag corresponds to the value of tf. Although some adjective tags are often used, they are also important. The tag size is changed on the basis of these levels of importance.

4.3 Generation of Tag Cloud

A tag cloud and a video are displayed in the same window. Because it is difficult to find tags from a tag cloud consisting of many tags, we have to limit the number of tags. In this study, we limit the tags to up to 30 words. These tags are selected randomly each time. We expect that all the tags are selected if many users watch a video. We expect the collaborative effect by multiple users. The font sizes of tags are changed in accordance with the level of importance. These tags are sorted in an alphabetical order.

5 Experimental Result

We perform experiments to evaluate the proposed system. Sixteen university students are chosen as the subjects. The subjects had already learned how to use the system. The target videos are A, B, C, D, E, and F, as shown in Table 1. The contents were posted by users of Synvie.

5.1 Evaluation of Interface

The proposed interface has the following two features: (1) a tag cloud and (2) tag sharing that enables users to share the tag information. A tag cloud consists of tags whose sizes are modified by their importance. To evaluate these features, we perform the following three experiments.

Table 1. Experimental contents

video	category	length	num of scenes
A	laboratory intoroduction	8 min 23 s	60
B	entertainment	12 min 47 s	4
C	animal	2 min 32 s	16
D	animation	5 min 39 s	41
E	art	1 min 41 s	1
F	cooking	1 min 16 s	18

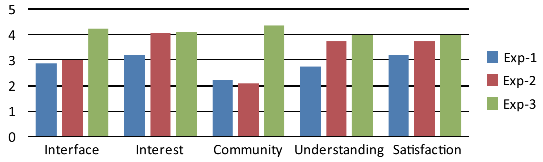


Fig. 6. Evaluations through the questionnaire. “5” signifies very good and “1” indicates poor.

- **Exp-1** is based only on a list of tags whose sizes are constant.
- **Exp-2** is based on a feature of a tag cloud.
- **Exp-3** is based on both tag cloud and tag sharing features.

The target contents are A and B. Two groups are considered in each experiment. No person watches a video twice. That is, we assume a situation in which a user watches a video only once.

The following questionnaire was given to the subjects.

- **Interface:** Usability of the interface.
- **Interest:** Do you like this system?
- **Community:** Do you have a sense of belonging to the community of users?
- **Understanding:** Do you understand the video content?
- **Satisfaction:** Are you satisfied with this system?

Subjects vote for each questionnaire by performing a five-stage evaluation. The averages of these votes are shown in Fig.6.

Let us compare Exp-1 with Exp-2. From the viewpoint of **Interest**, the average of Exp-2 (4.1) is higher than that of Exp-1 (3.2). This result suggests that the tag cloud is an important factor that enables users to develop an interest in movies. In the case of **Understanding**, the average of Exp-2 (3.7) is higher than that of Exp-1 (2.8). This result suggests that users can watch a movie attentively because the tag cloud interface allows them to find tags more easily.

Next, let us compare Exp-2 with Exp-3. In the case of **Community**, the average of Exp-3(4.4) is higher than that of Exp-2(2.1). Although movies in both Exp-2 and Exp-3 are shared on the Web, the average of Exp-2 is low. This result suggests that the users have a sense of belonging to the community of users when sharing tags. From the viewpoint of **Satisfaction**, users accept the proposed interface because the average of the averages of Exp-3(4.0) is higher than Exp-1(3.2) and Exp-2(3.8).

5.2 Evaluation of Annotation

Let us evaluate the annotations acquired by the proposed approach. The target contents were posted in Synvie. Because these movies had scene-related-annotations that had already been acquired from Synvie, we compare the Synvie approach with the proposed approach by analyzing the tags corresponding to scenes. Tag clouds are generated from annotations acquired from Synvie without using scene-related information. These annotations include more keywords related to scenes than that of YouTube.

Evaluation Method We proposed a method for evaluating the annotations. In the Synvie approach, although a particular scene has numerous annotations, the other scenes have little annotation, as shown in Fig.2. That is, the coverage of the Synvie approach is low. To evaluate these characteristics, we proposed the following evaluation functions. In this study, the scenes in a video are automatically divided every 2 s.

First, $T_{average}(x)$ is the average number of tags in the scenes that include one or more annotations.

$$T_{average}(x) = \frac{\sum_{s \in S'_x} T(s)}{|S'_x|} \tag{4}$$

where S'_x is a set of scenes (belonging to content x) that include one or more annotations, and $T(s)$ is the number of tags that are associated with Scene s.

Next, $T_{coverage}(x)$ is the ratio of the scenes that include one or more annotations in content x.

$$T_{coverage}(x) = \frac{\sum_{s \in S_x} K(T(s))}{|S_x|} \tag{5}$$

where S_x is a set of scenes belonging to content x, and $K(u)$ is the following step function.

$$K(u) = \begin{cases} 1 \; u \geq 0 \\ 0 \; u < 0 \end{cases} \tag{6}$$

To minimize the difference between the video contents, we defined the evaluated value as the average value for all the videos.

Experimental Result The experimental results are shown in Fig.7. In this system, we feel that the quality and quantity of annotation varies with the number of users. The results are expressed in the graph; the horizontal axis represents the number of users and the vertical axis represents the evaluated value. Fig.7 (a) shows the evaluated values of $T_{average}(x)$, and Fig.7 (b) shows the evaluated values of $T_{coverage}(x)$.

We consider the average number of tags per scene, as shown in Fig.7 (a). In the Synvie method, this number is approximately five times higher than that in the tag-cloud method. Although numerous tags can be generated when a user writes a long comment in a scene by using the Synvie method, the tag-cloud method does not allow users to click many tags in a scene. However, numerous

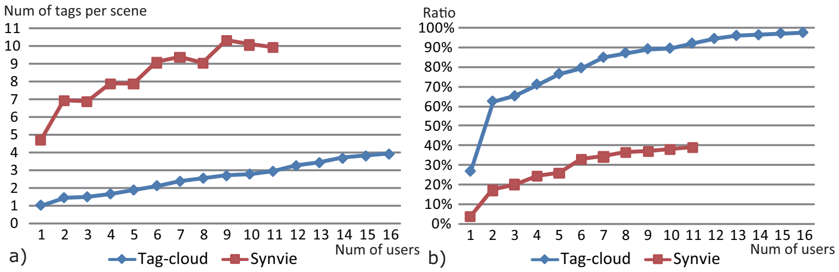


Fig. 7. Results. a) The number of tags per scene, b) The coverage of tags per content.

users will participate in adding annotations in the latter method because it is easier than the former method. Because the number of tags increases monotonically with the number of users, numerous tags may be acquired when many users participate.

Next, we consider the coverage of tags, as shown in Fig.7 (b). The coverage in the Synvie method has increased to approximately 40%, while the peak in the tag-cloud method was nearly 100%. These results show that the users clicked tags in almost every scene. Because the users can share the tag-clicked information, there may be a psychological compulsion on users to click tags that have never been clicked. That is, the tag-cloud method is the best from the perspective of the coverage of tags.

These results suggest that the tag-cloud method generates tags more widely and easily than the Synvie method. In the case of the tag average, the Synvie method is superior to the tag-cloud method. Further, the tag-cloud method consists of tags generated from Synvie. It may be useful to combine both the tag-cloud and Synvie methods. This approach will be investigated in a future study.

6 Related Work

In order to recognize content of videos or images, there exists some Web annotation systems such as WebEVA and Google Image Labeler. IBM Efficient Video Annotation (WebEVA) System [10] is a Web-based system that has a mechanism to annotate some concepts on large collections of image and video contents collaboratively in order to make a TRECVID ground truth. Many users evaluate the relation between a concept and a video by associating the following tags with each relation: positive, negative, ignore, and skip. When many users evaluate the same content at the same time, there may of course be contradictory evaluations.

The ESP Game [11] and Google Image Labeler[3] are mechanisms that enable users to add tags to an image in a manner that users consider to be a form of entertainment. It is an online contest that allows you to label random images and help improve the quality of image search results. It is a very clever mechanism that requires minimal efforts and provides entertainment. Tags acquired from

[3] http://images.google.com/imagelabeler/

this mechanism are used for the technical improvement of content-based retrieval of images.

7 Conclusions

In this paper, we proposed video scene annotation method based on tag clouds. This method has two features. One is a tag-cloud based annotation interface. The other is a tag-sharing function with other users. We confirm that these functions motivate the users to add annotations. Further, the proposed method resolved the problem of low coverage of tags faced in previous studies. The combination of this method with the Synvie method will enable the acquisition of superior annotations.

In a future study, by considering this method with the Synvie method, we will put this system to use. Further, we will study some applications based on annotations acquired in both the Synvie method and tag-cloud method.

Acknowledgments The authors would like to thank members of Takahashi-Katayama Laboratory at Nagoya Institute of Technology for their contribution to experiments of this system. .

References

1. M. Ben-Ezra and S.K. Nayar. Motion-based motion deblurring. *IEEE Trans. on Pattern analysis and machine intelligence*, Vol. 26, No. 6, pp. 689–698, 2004.
2. M. Davis. Media Streams: an iconic visual language for video annotation. In *Proc. of the IEEE Symp. on Visual Language*, pp. 196–202, 1993.
3. T. Masuda, D. Yamamoto, S. Ohira, and K. Nagao. Video scene retrieval using online video annotation. In *Lecture Notes on Artificial Intelligence (LNAI 4914: JSAI 2007)*, pp. 255–268, 2008.
4. Y. Matsumoto et al., *Japanese Morphological Analysis System ChaSen version 2.2.1.* http://chasen.aist-nara.ac.jp/, 2000.
5. K. Nagao, S. Ohira, and M. Yoneoka. Annotation-based multimedia summarization and translation. In *Proc. of the 19th Int'l Conf. on Computational Linguistics*, pp. 702–708, 2002.
6. K. Nagao, Y. Shirai, and K. Squire. Semantic annotation and transcoding: Making Web content more accessible. *IEEE MultiMedia*, Vol. 8, No. 2, pp. 69–81, 2001.
7. C. Parker and S. Pfeiffer. Video blogging: Content to the max. *IEEE MultiMedia*, Vol. 12, No. 2, pp. 4–8, 2005.
8. A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proc. of the ACM CHI Conference on Human factors in Computing Systems*, pp. 995–998, 2007.
9. J. R. Smith and B. Lugeon. A visual annotation tool for multimedia content description. In *Proc. of the SPIE Photonics East*, 2000.
10. T. Volkmer, J. R. Smith, and A. P. Natsev. A web-based system for collaborative annotation of large image and video collections: An evaluation and user study. In *Proc. of the 13th ACM Int'l Conf. on Multimedia*, pp. 892–901, 2005.
11. L. Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. of the ACM CHI Conf. on Human Factors in Computing Systems*, pp. 24–29, 2004.
12. H. D. Wactlar et al., Intelligent access to digital video: Informedia project. *IEEE Computer*, Vol. 29, No. 5, pp. 140–151, 1996.
13. D. Yamamoto, T. Masuda, S. Ohira, and K. Nagao. Video scene annotation based on web social activities. *IEEE MultiMedia*, Vol. 15, No. 3, (in press), 2008.