Video Scene Retrieval Using Online Video Annotation

Tomoki Masuda[1], Daisuke Yamamoto[1], Shigeki Ohira[2], and Katashi Nagao[3]

[1] Graduate School of Information Science, Nagoya University
[2] EcoTopia Science Institute, Nagoya University
[3] Center for Information Media Studies, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
{masuda,yamamoto,ohira,nagao}@nagao.nuie.nagoya-u.ac.jp

Abstract. In this paper, we propose an efficient method for extracting scene tags from online video annotation (e.g., comments about video scenes). To evaluate this method by applying extracted information to video scene retrieval, we have developed a video scene retrieval system based on scene tags (i.e., tags associated with video scenes). We have also developed a tag selection system that enables online users to select appropriate scene tags from data created automatically from online video annotation. Furthermore, we performed experiments on tag selection and video scene retrieval. We found that scene tags extracted by using our tag selection system had better cost performance than ones created using a conventional client-side video annotation tool.

1 Introduction

In recent years, a lot of video contents have been delivered and shared on the Web through the development of internet technology and the spread of broadband access lines. With the appearance of video sharing services, such as YouTube[1], the amount of video contents, which includes not only commercial but also user-generated contents, has been increasing explosively. Therefore, the demand for applications such as video scene retrieval and video summarization is rising and expected to rise even more in the future.

To make these applications, we must acquire meta information corresponding to the content of a video, which we call annotation [1]. We have developed an online video annotation system called Synvie in recent years [2]. We call data extracted from users' natural knowledgeable activities online video annotation. And we opened a public experimental service of Synvie and are accumulating annotation data now[2]. Online video annotation data have both advantages and disadvantages. They may contain video information from various people's viewpoints. And because they are accumulated from users' natural activities, the annotation costs are very small. However, they contain useless information, so we must screen them for use in practical applications.

[1] YouTube: http://www.youtube.com/
[2] Synvie Beta: http://video.nagao.nuie.nagoya-u.ac.jp/

K. Satoh et al. (Eds.): JSAI 2007, LNAI 4914, pp. 54–62, 2008.

In this paper, we propose an efficient screening method and an application based on online video annotation. Specifically, we propose an efficient method for extracting scene tags from online video annotation. To evaluate the method by applying extracted information to video scene retrieval, we developed a video scene retrieval system based on scene tags.(i.e., tags associated with video scenes). Moreover, we performed experiments on screening tags and video scene retrieval. Through these experiments, we verified the usefulness of online video annotation and our screening method and application of it.

2 Creating Scene Tags (Tagging Video Scenes)

Creating scene tags means relating keywords to an arbitrary time code in a video. Scene tags contain nouns, verbs, and adjectives, but do not contain particles or auxiliary verbs. Unknown words are treated as nouns. We created scene tags by using three methods for 27 videos that were registered in Synvie. The length of the used videos was about 349 seconds on average: the longest video was 768 seconds and the shortest was 76 seconds. We used various kind of videos, e.g., educational videos, stories, and entertainment videos. In next chapter, we compare the usefulness of tags created by each method.

2.1 Tagging Using an Annotation Tool

One annotator added scene tags by using a tool that enables a user to add tags at an arbitrary time in a video. The annotator who was not a creator and did not have any special knowledge about the videos added objective information acquired from images and sounds to video scenes as scene tags in detail and exhaustively. This method is a kind of conventional client-side video annotation [3]. We defined the human cost for creating scene tags as the time that an annotator spent adding them. This was 1480 seconds on average: the longest time was 3692 seconds and the shortest was 582 seconds.

2.2 Automatic Extraction of Scene Tags from Online Video Annotation

Synvie is a video sharing system that lets users comment on video scenes and quote them in a weblog. We have been running a public experimental service since July 1, 2006 and analyzing data accumulated from July 1 to October 30, 2006. We gathered 97 registered users and 94 videos. From the accumulated annotation data, we could acquire text data related to time data. Through some processing, we created scene tags automatically from annotation data. The tag creation process is shown below.

1. Using morphological analysis (using a Japanese morphological analyzer "Cabocha"[4]).
2. Removing stop words.

3. Extracting nouns, verbs, adjectives, and unknown words.
4. Relating words to time data and saving them in a database.

These processes can be performed automatically and annotation data can be accumulated through natural communication by humans on the web, so it can be said that the human costs for creating scene tags is extremely small. 153 scene tags were created on average for 27 videos by this method.

2.3 Scene Tag Extraction Using an Online Tag Screening System

We can easily predict that annotation data may include useless data such as data having no relation to the video. In comments or weblogs, users do not necessarily refer to the contents of the video. Therefore, scene tags created from the online video annotation automatically, as described in section 2-2, have a high probability of including useless tags. Indeed, we found scene tags that were obviously unsuitable for scenes being viewed. They included tags that were not meaningless but were unsuitable for the scenes and tags that lost their meaning as a result of the morphological analysis processes. For these reasons, the quality of tags created automatically from online video annotation will not be high. So we must screen tags in order to use them in practical applications. If we succeed in screening them appropriately, we will obtain higher-quality tags.

Because it would be ideal for this screening to be achieved successfully through automatic processing, we tried to do it by various different methods. First, we used the well-known technique TF-IDF (Term Frequency-Inverse Document Frequency) [5]. However, this technique needs a large quantity of documents to be successful in finding appropriate words. Second, by using Google Web API[3], we tried to score words by co-occurrence relations to tags that had been added when the video was registered. But in this method, the scores of words that appear in general documents were higher than those of words that were strongly related to the scene. These results show that it is very difficult to perform appropriate screening of scene tags by automatic processing and that manual processing by humans is necessary for it to be successful. So we developed a tag screening system that enables online users to select appropriate scene tags from data created automatically from online video annotation (Figure 1). We can guess that the quality of tags selected by humans is high. But the more human cost we include in this process, the more we lose the advantages of using online video annotation. Mechanisms that enable users to select tags efficiently are required in order to reduce human costs. This system is used by one or more users. Users watch a video, and when a time code to which tags have been added comes, the video stops temporarily and the users select tags that are appropriate to the scene. We performed experiments on screening tags using this system. The number of subjects for each video was two or three. We defined the human cost for creating tags by using this system as the time that each user spent in selecting. We calculated this value from the automatically measured time. The average time was 314 seconds, which is 1/5 of the time spent for creating tags using the

[3] Google Code: http://code.google.com/

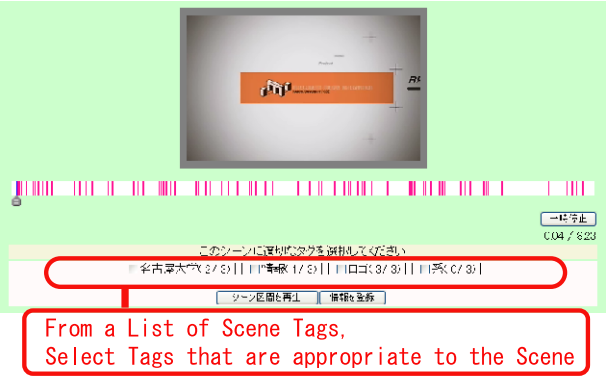

Fig. 1. Online tag screening system

Table 1. Comparison of human costs for creating tags

Tag Creation Method	Human Costs (Second)
Tagging Offline	1480
Extracting Automatically	0
Screening Online	314

annotation tool described in section 2.1. In this experiment, 55 scene tags were created on average for 27 videos. The results show that 36.2 percentages of tags that were created automatically from online video annotation were judged to be appropriate to the scene. A comparison of the three methods in terms of human cost for creating tags is given in Table 1.

3 Video Scene Retrieval

We have developed a tag-based video scene retrieval system based on a new concept. And we performed experiments on video scene retrieval using the scene tags created as described in the previous section.

3.1 Tag-Based Video Scene Retrieval System

Our video scene retrieval system is based on the mechanism of tag-cloud [6] and makes the most of the characteristics of scene tags. Scene tags generated from annotations have an essential problem in that appropriate tags are not necessarily given to all scenes, and their completeness is small. When there are not enough annotations for each video, it is hard to apply usual search techniques such as exact matching using tags. But tags also have a strong point in that a large number of tags can be displayed in a small space on a browser and this can be helpful for

efficient retrieval. We developed this system considering these characteristics of scene tags. The process of retrieving video scenes is shown below.

1. Select an arbitrary number of tags and submit them as a query.
2. A list of videos is returned corresponding to the query. And a timeline seek bar that highlights the time ranges of tags used as the query has been added and a list of all scene tags that have been added to the video are displayed with each video.
3. Select tags from a list to correspond to the timeline seek bar.
4. Move the seek bar to view thumbnail images for an arbitrary time code.
5. Play the video from the arbitrary time code.

Fig. 2. Tag-cloud for video scene retrieval

The top page of this video scene retrieval system is shown in Figure 2. A tag-cloud composed of scene tags and tags that were added when the video was registered are displayed. Tags are classified into nouns (including unknown words), verbs, and adjectives in ABC order and A-I-U-E-O order (Japanese order). When a tag is clicked, the word of the tag is added to the text field for searching, so it is not necessary to input text using a keyboard. And users can use an incremental search for tags. Incremental searching is a search that progressively finds a match for the search string as each character is typed. In this system, when a letter is typed, only tags that start with that letter are displayed and the others are hidden. These functions help users to find tags for making queries from a large number of tags. The output of the search is a list of videos that include these tags (Figure 3).

Fig. 3. Results of search

Each video has a seek bar associated with scene tags and thumbnail images arranged along the time axis. The timeline seek bar helps users to view video scenes on a web browser without accessing the video itself. When the user drags the seek bar to an arbitrary time code, the system displays thumbnail images synchronized with the time code of the seek bar. This function helps users to view images of a time code to which tags have not been added. Because the time ranges to which tags have been added are highlighted on the seek bar, the user can understand the content of the video by browsing these tags and thumbnail images without actually watching it. Moreover, when the user clicks an interesting-looking tag, the temporal location of the tag is displayed on the seek bar. These actions are repeated and video scenes that users want to see are found. An example of an image for which video scene retrieval was performed is shown in Figure 4. We are continuing with the development to make retrieval more efficient. And we have been running a public experimental service from February 27, 2007.[4]

3.2 Experiment on Video Scene Retrieval

We performed experiments on our video scene retrieval system. We chose nine scenes as retrieval targets. The questions asked for a "scene where a certain animal was reflected by the parent and child", "scene before the person who was snowboarding crashed into the edge of the course", etc. and did not necessarily include words and phrases given as scene tags, but subjects could guess the scene by getting hints from the scene tags. Moreover, to ensure that the answer to each question was the only time range, a thumbnail image was also given as well as the

[4] Divie: http://video.nagao.nuie.nagoya-u.ac.jp/search/top

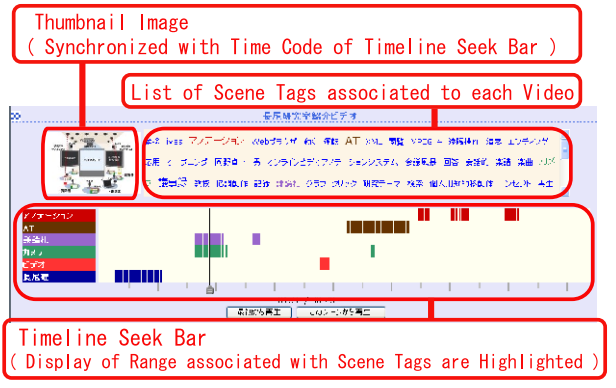

Fig. 4. Interface for video searching with a seek bar and tags

text. Subjects retrieved the answer scene to each question, and the time spent on the answer was measured automatically. The number of subjects was nine. Subjects retrieved scenes using tags created by the three methods described in the previous section. Each subject retrieved three scenes by using tags created by each method (9 scenes in total). Therefore, each scene tag creation method could be compared impartially. We prepared an experimental top page that did not reveal which method was used to create the tags that subjects used for each retrieval. The data that we acquired in this experiment are shown below.

- Scenes decided as answers.
- Time spent on retrieval.
- Queries used for retrieval.
- Viewed scenes.

Because all subjects were able to discover a correct scene for all questions in this experiment, we could not compare the scene tag creation methods from this viewpoint. Therefore, we compared them according to the time taken for the retrieval. The experimental results are shown in Table 2. It can be thought that the difference did not go out by some influences other than unlike tags at the retrieval time because the number of queries submitted increased with the average retrieval time. The retrieval time was longest for tags extracted automatically from online video annotation and shortest for tags created using an annotation tool. This result shows that retrieval time was shorter in the two methods that involved human cost for creating scene tags. Though automatic tag creation from online video annotation was the best method in terms of tag creation cost, its retrieval costs were very high. And because it is essential to shorten the retrieval time for video contents, which will grow in volume in the future, it is necessary to put the human cost for tag creation.

Therefore, we compared tags created using an annotation tool with ones created using the online tag screening system in terms of their cost performance

Table 2. Results of experiments on video scene retrieval

Tag Creation Method	Time (Second)	Number of Queries
Tagging Offline	118.1	132
Extracting Automatically	169.6	202
Screening Online	145.4	156

(cost-effectiveness), that is, the ratio of retrieval cost to creation cost. Cost performance C of the tag was calculated by the following equation, where n is equal to the time spent for retrieval when automatically created tags were used, RT is the time spent for retrieval when the tags created by each method were used, and CT is time spent creating the tags.

$$C = \frac{n - RT}{CT} \times 100 \tag{1}$$

From this equation, we can calculate how much the retrieval time was reduced by spending 100 seconds creating tags. The results are given in Table 3. These results indicate that, when comparing methods from the viewpoint of cost performance of each tag, the method of creating tags using the online tag screening system was best in this experiment.

Table 3. Cost performance of each tag

Tag Creation Method	Cost Performance of Each Tag
Tagging Offline	3.48
Screening Online	7.71

4 Conclusion and Future Work

4.1 Conclusion

In this paper, by using and screening online video annotation, we could create useful scene tags without high human costs. The results of experiments showed the usefulness of online video annotation and of the screening method. We have developed a tag-based scene retrieval system based on a new concept and proposed an application of online video annotation.

4.2 Future Work

We must verify our method using a large quantity of data accumulated on a public experimental service. Moreover, we must improve the interface and the algorithm of the video scene retrieval system for more efficient retrieval.

References

1. Nagao, K.: Digital Content Annotation and Transcoding. Artech House Publishers (2003)
2. Yamamoto, D., Ohira, S., Nagao, K.: Weblog-style Video Annotation and Syndication. In: Proceedings of the 1st International Conference on the Automated Production of Cross Media Content for Multi-channel Distribution (AXMEDIS) (2005)
3. Nagao, K., Ohira, S., Yoneoka, M.: Annotation-Based Multimedia Summerization and Translation. In: COLING 2002. Proceedings of the Nineteenth International Conference on Computational Linguistics (2002)
4. Kudo, T., Matsumoto, Y.: Fast Methods for Kernel-Based Text Analysis. In: ACL 2003. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 24–31 (2003)
5. Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Communications of the ACM 18(11), 613–620 (1975)
6. Rivadeneira, A.W., et al.: Getting our head in the clouds: Toward evaluation studies of tagclouds. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 995–998 (2007)