Video Scene Annotation Based on Web Social Activities

Daisuke Yamamoto, Tomoki Masuda, Shigeki Ohira, and Katashi Nagao
Nagoya University

This article describes a mechanism to acquire the semantics of video content from the activities of Web communities that use a bulletin-board system and weblog tools to discuss video scenes.

The amount of multimedia content on the Web has been increasing in recent years as more and more video and music clips are posted and shared not only by commercial entities but also by ordinary users. At the same time, the influence of users and Web communities has increased as a result of Web communication tools, such as weblogs, social-networking services, and wikis. Indeed, there are some existing video-sharing systems capable of video-editing functionality, such as Motionbox, Jumpcut, and Kaltura, but while these systems let users edit and share videos for communications, they lack functionality for recognizing the content of videos. There is therefore a need for platforms and tools that help users post, manage, and search for video clips.

Applications summarizing videos or retrieving scenes from them must acquire metainformation about the contents of those videos. Here, we call this information *annotation*, and it can occur by extracting keywords related to the content.[1] Previous studies either used an automatic method[2] that extracted metadata by using image-recognition and voice-recognition technologies or they used a semiautomatic method[3-5] where expert annotators associated video content with high-quality annotation data by using annotation tools for describing MPEG-7. Accuracy of automatic recognition of the content of nonexpert-created videos is still not high because this content might include noise, indistinct voices, and blurred or out-of-focus images.[6] Automatic recognition technologies are therefore not necessarily effective with nonexpert-created videos. Another problem with audiovisual content created by nonexpert users is that it's not cost-effective to have experts using semiautomatic annotation tools associate all of it with detailed annotations.

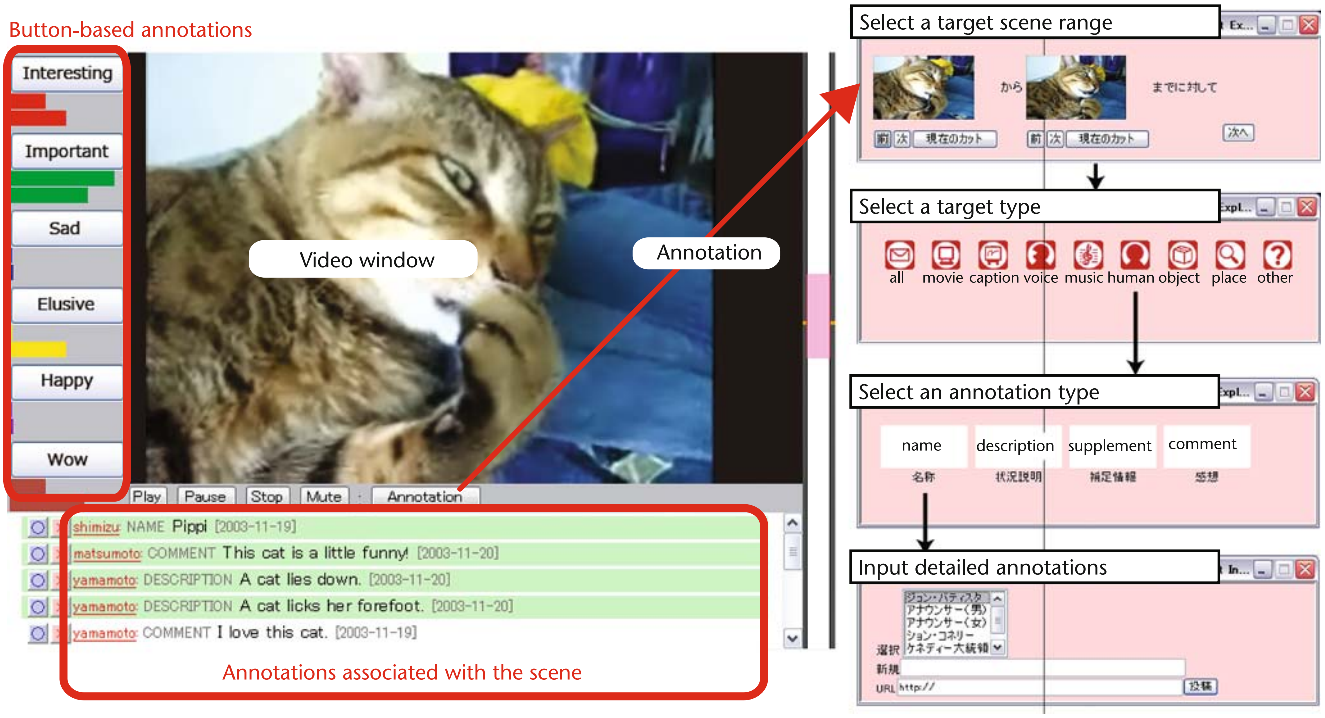
We propose a new solution to these problems. Our approach is based on social activities, especially user comments and weblog authoring, associated with the content of video clips on the Web. We developed a mechanism that helps users of online bulletin-board-type communications associate video scenes with user comments and another mechanism that helps users of weblog-type communications generate entries that quote video scenes. We also developed a system that can extract deep-content-related information about video contents as annotations automatically. We believe that the cost of the proposed system is low and that the system is robust with respect to quality of the automatic pattern recognition used.

System architecture

A user-friendly, video-annotation system that would enable any user to refer to any fragment of an online video is sorely needed. Parker states that advanced applications such as video-blog search and video-blog feeds could be developed by applying certain weblog mechanisms, such as the trackback and permalink functions, to video content.[7] We believe that a new weblog mechanism with permalinks to video scenes and user comments on them would contribute to advanced video applications, such as video scene retrieval. In our previous work, we developed an online video-annotation system called iVAS, shown in Figure 1.[8] Although iVAS originally was a Web-annotation tool that let users associate detailed content descriptions with scenes while watching a video, it mainly was a communication tool that allowed users to post and share comments and impressions. Although these comments were short and unclear, they had some valuable keywords that corresponded to scenes, so we decided to extend iVAS.

Objectives for our new video-sharing system, called Synvie, include acquisition of

1070-986X/08/$25.00 © 2008 IEEE



shimizu: NAME Pippi [2003-11-19]

matsumoto: COMMENT This cat is a little funny! [2003-11-20]

yamamoto: DESCRIPTION A cat lies down. [2003-11-20]

yamamoto: DESCRIPTION A cat licks her forefoot. [2003-11-20]

yamamoto: COMMENT I love this cat. [2003-11-19]

Figure 1. Interface of the iVAS online video annotation system.

annotation data from Web users' social activities and development of annotation-based video applications. Figure 2 (next page) shows the system's architecture. Annotation methods allow users to view any video, submit and view comments about any scene, and edit a weblog entry to quote scenes using an ordinary Web browser. These user comments and the links between comments and video scenes are stored in annotation databases. An annotation analysis block produces tags from the accumulated annotations, while an application block has a tag-based, scene-retrieval system.

Video is represented in Synvie by a set of shots. A *shot* is defined as a sequence of frames that don't differ from each other drastically.[9] An interruption between two shots is called a *shot transition* or a *cut*. Because some of the video shots posted in blogs can be long, our system divides a long shot into 2-second subshots to capture and summarize the shot's semantics more effectively. Thus, our system combines the temporal sampling of the video with fixed sampling to select the most representative keyframes from the video.

Annotation methods

Explicit annotations (such as MPEG-7 description) are users' descriptions of the semantics, attributes, and structures of video content. Tags and content descriptions also are explicit annotations. Implicit annotations, on the other hand, are extracted from user activities such as submitting comments and writing weblog articles.[10] Although implicit annotation involves some analysis errors, it spares users the trouble of making annotations. By offering tools for communication and authoring weblog entries, Synvie supports the creation of implicit annotations. We define a *content comment* as a user's comment about an entire video and a *scene comment* as a user comment on a specific video scene. We define a *scene quotation* as a mechanism that allows users to quote any video scene in a weblog entry.

Scene comment

Scene-comment-type annotation allows users to associate video scenes with text messages. Because we needed an interface that helps users submit comments on any video scene easily, we simplified the iVAS annotation interface and optimized it for user communication.

To comment on a scene while watching a video, the user presses the scene comment button, which pauses the video and lets the user inspect the thumbnail image corresponding to the scene, as shown in Figure 3. If the image is not suitable, the user presses the fast-forward or rewind button, changing the thumbnail image accordingly. When the user is satisfied with the beginning image, he or she

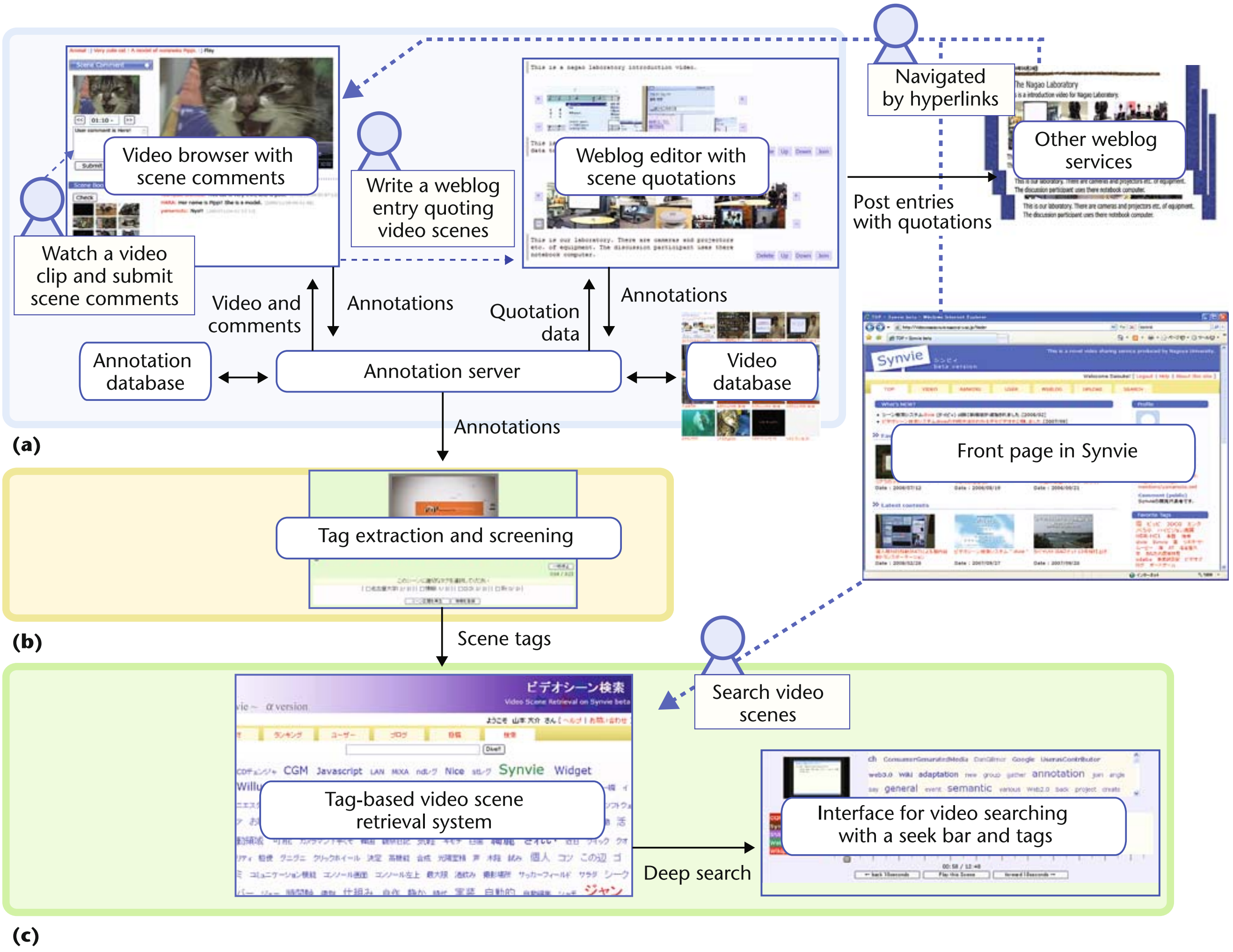

Video browser with scene comments
Write a weblog entry quoting video scenes
Weblog editor with scene quotations
Navigated by hyperlinks
Other weblog services
Post entries with quotations
Watch a video clip and submit scene comments
Video and comments
Annotations
Quotation data
Annotations
Annotation database
Annotation server
Video database
Annotations
(a)
Front page in Synvie
Tag extraction and screening
(b)
Scene tags
Search video scenes
Tag-based video scene retrieval system
Interface for video searching with a seek bar and tags
Deep search
(c)

Figure 2. Architecture of the Synvie annotation and quotation system.

writes a messages in the text-input area just below the adjustment buttons and presses the submit button. If a user wants to be able to retrieve a scene later but not write a comment on it, he or she presses the check button to bookmark the scene. Thumbnails that were either commented on or checked are displayed at the lower left of the interface.

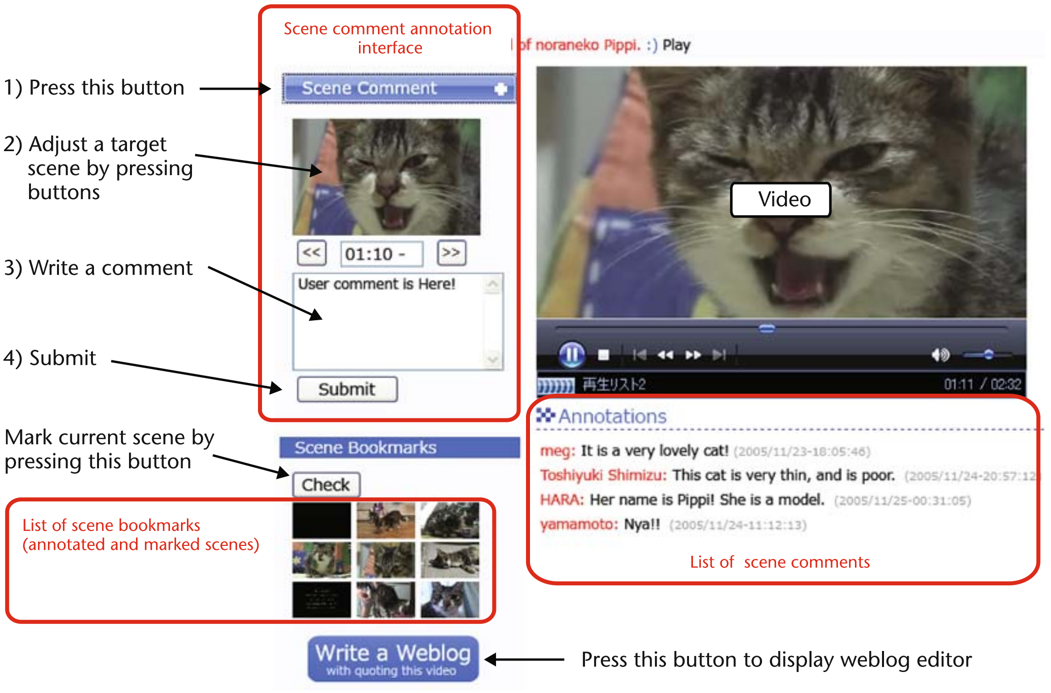
Each scene comment is displayed when the corresponding video scene is shown, so several messages associated with the same scene can be displayed simultaneously. This interface lets users asynchronously exchange messages about video scenes in a communication style similar to that in which the users of online bulletin-board systems share impressions and exchange information.

Scene quotation

By supporting user quotation of video scenes in a weblog entry, we accumulate a detailed user-editing history and acquire annotations that relate the sentence structure of the weblog entry to the scene structure of the video content.

We call a weblog entry that quotes video scenes a *video blog entry*. We expect many video blog entries discussing a video clip, and here we consider two types. In one entry type, the user who submits video content edits an entry that introduces that content. Users can easily indicate specific parts of the content by quoting them. This kind of entry is useful for advertising the video clip. In the other entry type, a user who watched a video clip and liked it edits an entry introducing the clip to other users. We expect that the more popular a video clip is, the more entries there will be.

A video blog entry is composed of multiple paragraphs quoting video scenes, and each of those paragraphs contains thumbnail images of quoted video scenes, user comments, and


1) Press this button
2) Adjust a target scene by pressing buttons
3) Write a comment
4) Submit
Mark current scene by pressing this button
Scene comment annotation interface
Scene Comment
<<
01:10 -
>>
User comment is Here!
Submit
Scene Bookmarks
Check
List of scene bookmarks (annotated and marked scenes)
Write a Weblog
with quoting this video
Press this button to display weblog editor
of noraneko Pippi. :) Play
Video
再生リスト2
01:11 / 02:32
Annotations
List of scene comments

meg: It is a very lovely cat! (2005/11/23-18:05:46)

Toshiyuki Shimizu: This cat is very thin, and is poor. (2005/11/24-20:57:12

HARA: Her name is Pippi! She is a model. (2005/11/25-00:31:05)

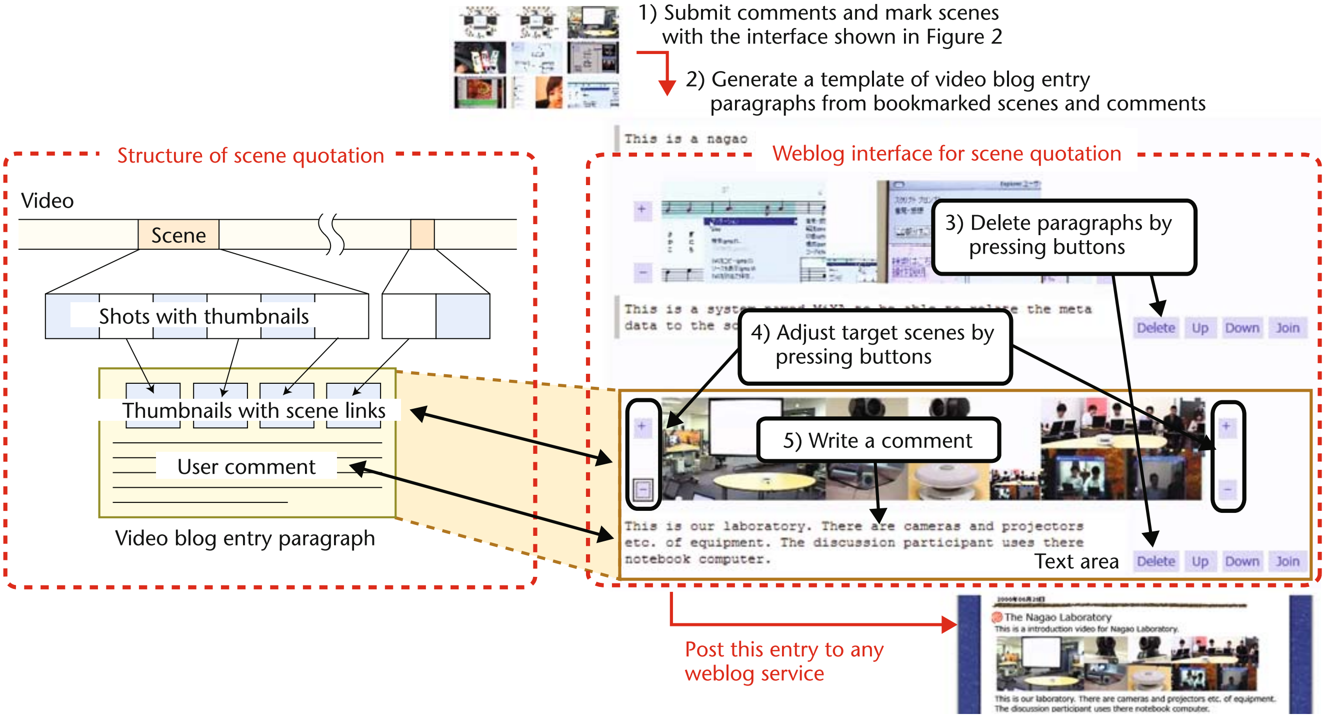
yamamoto: Nya!! (2005/11/24-11:12:13)

Figure 3. Interface for a scene-comment-type annotation.

links to scenes (see Figure 4). We treat a comment as a scene quotation, that is, an annotation to the scenes.

A user must select scenes that he or she wants to quote, and for this selection we use the same mechanism that we use for making scene annotations. We assume that a user submits scene comments because those scenes are interesting to the user. They will therefore be candidate references for quotation in video blog entries; a user can edit a video blog entry by quoting these nominated scenes.

We think it's best if users can edit video blog entries with a Web browser in the same way as editing typical weblog entries. Our interface for quoting video scenes is suitable


1) Submit comments and mark scenes with the interface shown in Figure 2
2) Generate a template of video blog entry paragraphs from bookmarked scenes and comments
Structure of scene quotation
Video
Scene
Shots with thumbnails
Thumbnails with scene links
User comment
Video blog entry paragraph
Weblog interface for scene quotation
3) Delete paragraphs by pressing buttons
4) Adjust target scenes by pressing buttons
5) Write a comment
Text area
Delete
Up
Down
Join
+
−
Post this entry to any weblog service
The Nagao Laboratory

This is a nagao

This is a system named MiXY so be able to relate the meta data to the sc

This is our laboratory. There are cameras and projectors etc. of equipment. The discussion participant uses there notebook computer.

This is a introduction video for Nagao Laboratory.

This is our laboratory. There are cameras and projectors etc. of equipment. The discussion participant uses there notebook computer.

Figure 4. Structure and a weblog interface for scene quotation.

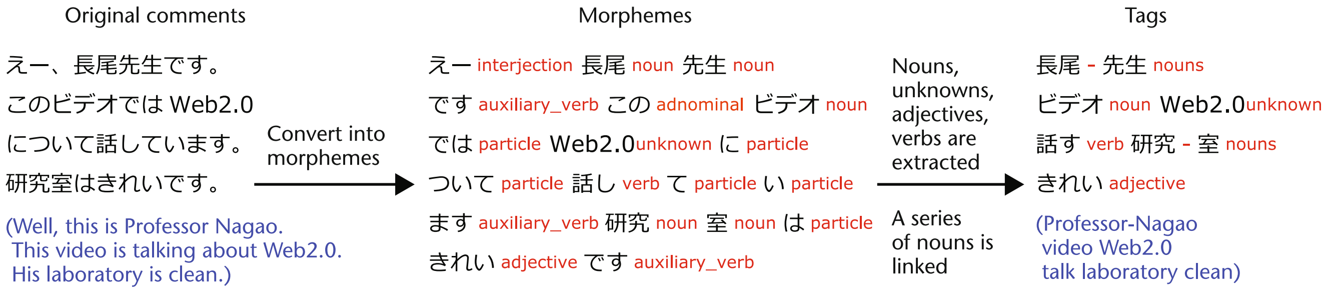

Original comments
えー、長尾先生です。
このビデオでは Web2.0
について話しています。
研究室はきれいです。
(Well, this is Professor Nagao.
This video is talking about Web2.0.
His laboratory is clean.)
Convert into morphemes
Morphemes
えー interjection 長尾 noun 先生 noun
です auxiliary_verb この adnominal ビデオ noun
では particle Web2.0unknown に particle
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 clean)

Figure 5. Tag extraction from comments.

for editing a video blog entry that quotes continuous scenes, as Figure 4 shows. This interface lets users select scenes, adjust their beginning and end times, and write messages longer than scene comments by going through the following steps (which are shown in Figure 3):

1. Create comments about or mark video scenes while watching video clips.

2. Press the write a weblog button that opens another interface for weblog authoring. This interface shows a video blog template that is automatically generated by retrieving scene bookmarks.

3. Browse thumbnail images of all marked scene paragraphs and delete unnecessary paragraphs by pressing the delete button.

4. Adjust time intervals of scenes by pressing the buttons labeled + and −.

5. Write messages in text areas just below thumbnail images and then submit the generated content in HTML to an existing weblog service.

Because this interface lets us modify the beginning and end time of a quoted continuous scene by expanding and contracting video shots on the media time axis, we can select video scenes more exactly than in scene comments. This interface is suitable for editing an entry that stresses the importance of the description of a video story. Weblog entries are stored into annotation databases when they are being posted to a weblog site.

We treat a scene quotation as an annotation. A scene quotation has two different annotation types. One is a more correct and informative scene comment. In addition, we consider paragraphs in video blog entries as annotations of quoted scenes; their sentences will have better wording and fewer misspellings than scene comments created while watching videos because video blogs can be written more carefully. The other type is semantic relationships between video scenes. Simultaneous quotation of continuous scenes of a video clip can specify that a video shot series has a semantic chunk. Simultaneous quotation of scenes of different video clips might clarify the semantic relationships between quoted scenes and videos. For example, if a user quoted scenes of different video clips in a video blog, the system might also find some semantic relationships between these scenes and clips. We will be able to calculate the semantic similarity of content on the basis of these relations in the future.

Annotation analysis

Our system accumulates annotations, such as scene comments and scene quotations, without degrading any of the information in them. Because these annotations are only sets of user comments, it's not necessary for a machine to be able to understand them. We think that these annotations include some semantics corresponding to video scenes. To develop applications using these annotations, we must analyze annotations and convert them into machine-understandable data.

Tag extraction

We extract from scene comments keywords that express scene semantics. We call a keyword associated with content a *tag* and, in particular, a keyword associated with scenes a *scene tag*. We extract scene tags (see Figure 5) by automatically analyzing each comment and converting it into morphemes. Then we extract nouns other than dependent nouns, verbs other than dependent verbs, adjectives, and unknown words from these morphemes by using the Japanese morphological analysis system, ChaSen.[11] Unknown words are treated as proper nouns. The basic form of each morpheme becomes a tag.

Because this method can't generate tags that consist of compound words, we extended it to convert a series of nouns into a tag. Our system assumes that comments are mainly written in Japanese, so if a series of alphanumeric characters such as "Web 2.0" appear in Japanese text, the algorithm generates a tag corresponding to the series.

Tag screening

Automatically extracted scene tags include ineffective scene tags unrelated to scenes. Because it's difficult to screen these tags automatically, we use manual screening. We developed a tag-selection system that enables Web users to select appropriate scene tags from automatically extracted tags.

Scene tags synchronized with video scenes are displayed for users. This system shows automatically extracted scene tags to users who are watching the scenes corresponding to those tags. Each tag is shown accompanied with a check box, and the user can screen tags by selecting ones related to the present scene, as shown in Figure 6. The time taken to select tags is the cost of screening the tags. By getting many users to participate, we expect to make the per capita cost of screening the tags small. Because this is a routine work, we need to develop a mechanism in which tags can be screened more efficiently.

Experimental results

To evaluate our annotation system and accumulate a lot of annotation data about video content, we ran a public experimental service based on our system (see http://video.nagao.nuie.nagoya-u.ac.jp/, a service available only in Japanese). The service started on 1 July 2006. We evaluated data accumulated from 1 July to 22 October 2006. We gathered 97 registered users, 94 submitted video clips, 4,769 annotations, and 7,318 accesses by users. Nonexpert users submitted video clips related to education, travel, entertainment, vehicles, animals, and so on, with their average length being 320 seconds. We compared the experimental service with existing video-sharing services such as YouTube. YouTube-type comments correspond to our content comments. In addition, we provided scene-comment type and scene-quotation-type annotations and confirmed the usefulness of these annotations by comparing their quantity and quality with those of the annotations generated using the other services.

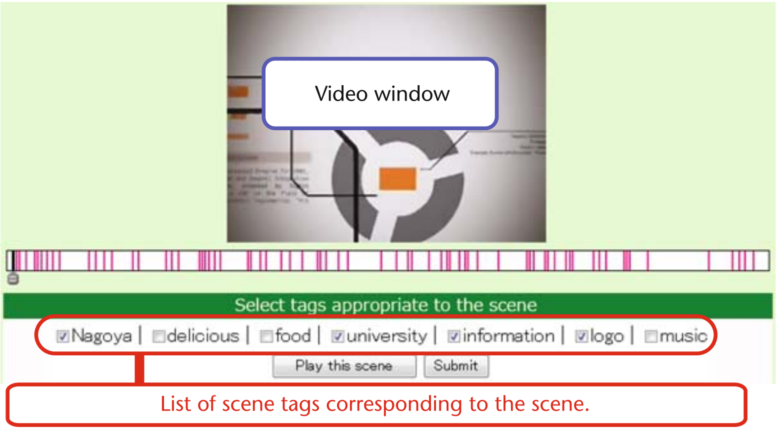

Video window
Select tags appropriate to the scene
☑Nagoya | □delicious | □food | ☑university | ☑information | ☑logo | □music
Play this scene
Submit
List of scene tags corresponding to the scene.

Figure 6. Tag screening system.

Quality of annotations

Annotation data acquired by our experimental system consisted mainly of text comments created by ordinary Web users. We used them in our video-scene retrieval system, which we describe in the next section. For this purpose, each message must clearly describe the content of its corresponding video scene. We therefore had to evaluate characteristics of the comments. We did so by manually classifying all of the accumulated annotations into the following A through D classes by considering relevance of the message's description to the content of the corresponding video scene:

- A. Comment that mainly explains video scene content.
- B. Comment that consists mainly of opinions of video scenes and includes keywords related to the scene.
- C. Comment that discusses topics derived from the scene content.
- D. Incomprehensible comment with text that consists of exclamations or adjectives or text about topics irrelevant to content, such as the video-streaming quality or video-capturing method.

We also categorized the text of categories A, B, and C into subcategories based on the text quality:

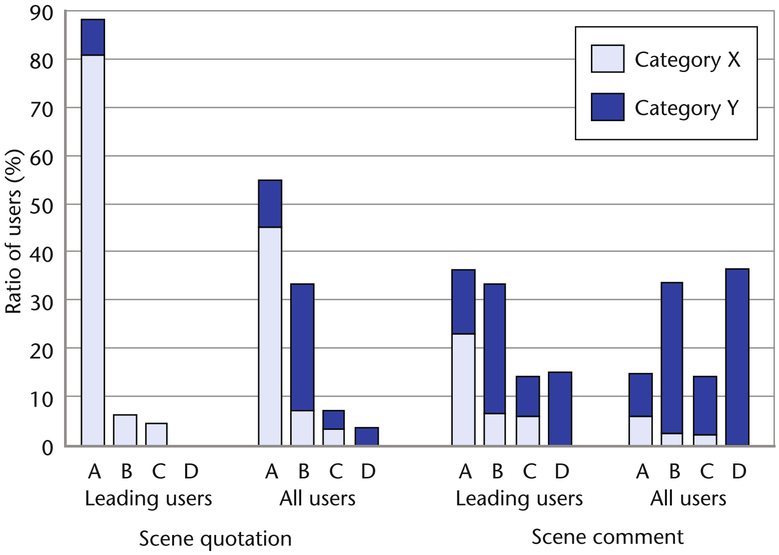


Figure 7. Annotation quality for each method.

- X. Comments that express enough content, such as text that consists of subject, predicate, and objects.

- Y. Comments that don't express enough content.

Two evaluators also simultaneously categorized all annotations. When they disagreed, they reached an agreement by discussion.

An example of category A–X is a scene quotation about a morning glory exhibition in a photographer's weblog entry: "It's a morning glory of the Yashina type cut in the Nagoya style called *Bon*. It enables bonsai tailoring without extending the vine, and it is unique. It has a history of 100 years." These sentences adequately express the scene's content; we can extract more semantics by analyzing the language. An example of category A–Y is a scene comment about an image uploaded in a Web application: "Upload image." This phrase doesn't adequately express the content but does include some keywords related to the scene. An example of category B–X is the scene quotation, "I wonder how this cat stayed alive under the bench. It seems so cold…," which expresses an opinion about a video scene. In the example of category B–Y we can extract some keywords from the scene comment, "You eat too much junk food." An example of category C–X is a scene comment about the scene caption's URL: "It is a recycling toner specialty store, and it produces free CG movies and music." An example of category C–Y is the scene comment, "Prof. Nagao mainly researches annotation." Although these scene comments don't directly express the content of the corresponding video scene, we can use them as supplementary information. Examples of category D include scene comments that alone can't express meaning, such as "Great!," "Beautiful!," or a scene comment about video image quality, such as "This image is unclear. Please use Window Media Encoder to make this image clearer."

Discussion of quality

Most of the accumulated annotations (795) were scene comments, less than half as many (334) were scene quotations, and only 40 were content comments. We think that the amount of annotations accumulated is related to the ease of annotation. Because most of the annotations were for scene comments, we infer that scene comments are the easiest to use of these three communication tools. You might think that there would be more content comments than scene comments because scene comments are more difficult to make, but users who watch the same scenes share context and can therefore easily submit brief comments.

Although a strict definition of annotation quality is application-dependent, we consider a higher-quality annotation to be one that consists of grammatically correct sentences, describes the details of the scene semantics, and includes keywords. We consider the quality of the four annotations classes to be ranked in the order A > B > C > D and the quality of subcategory X to be higher than that of subcategory Y. So we consider A–X, A–Y, and B–X annotations to be effective ones.

We can see in Figure 7 that the scene quotations were better than the scene comments. 59 percent of the scene quotations posted by all users were effective, while only 11 percent of the scene comments posted were effective. Only 4.8 percent of the scene quotations posted by all users were of the lowest quality (category D), while 36 percent of the scene comments posted by all users were category D annotations. This breakdown was largely due to scene quotations containing fewer irrelevant and spam-like comments.

These results show that scene quotations, which users who don't necessarily share context use for weblog entries, tend to be written more politely than scene comments used for ad hoc communication among users watching a video and sharing a context. This finding reflects the fact that the quality of weblog text is generally higher than that on a bulletin board. Annotation quality therefore depends on the type of annotation.

We define as leading users the 30 percent submitting most of the annotations categorized as A. Figure 7 shows that 95 percent of the scene quotations created by leading users were effective, that only 62 percent of those posted by all users were effective, and that 43 percent of the scene comments posted by leading users were effective. Annotation quality thus depends on the characteristics of the users creating the annotations.

We can conclude that our video-scene-oriented annotation methods will gather more and higher quality information than other methods do and that this result will depend on whether they are scene comments (bulletin-board type) or scene quotations (weblog type). We found that when there are a lot of annotations, we should use only the scene quotations, and that when there are few annotations we must also use the scene comments.

Discussion of scene tags

To evaluate tags generated by using the technique described in the previous section, we manually classified them as either effective or ineffective according to whether they were or were not closely related to the content of the corresponding scenes.

More than three times as many effective tags were extracted from scene quotations: the average number of effective tags extracted from a scene quotation was 5.96, and the average number of effective tags extracted from a scene comment was 1.51. Both kinds of annotations are comments discussing scenes, but scene quotations are written in more detail and tend to be better sources of effective tags.

Evaluating the effectiveness of the tags generated automatically by the tag-screening technique described in ''Tag screening'' subsection of this article, we found that 59 percent of the tags extracted from the scene comments submitted by all users were effective. This might not seem to be a high percentage, but all annotations are more or less related to scenes in that the users write them while watching the content. Even those tags classified as ineffective in this study might therefore be effective for some applications.

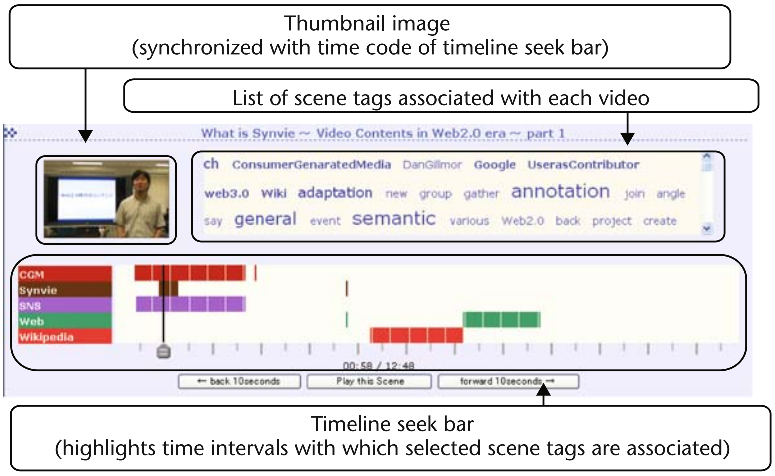

Thumbnail image
(synchronized with time code of timeline seek bar)
List of scene tags associated with each video
What is Synvie ~ Video Contents in Web2.0 era ~ part 1
ch ConsumerGenaratedMedia DanGillmor Google UserasContributor
web3.0 Wiki adaptation new group gather annotation join angle
say general event semantic various Web2.0 back project create
CGM
Synvie
SNS
Web
Wikipedia
00:58 / 12:48
← back 10seconds
Play this Scene
forward 10seconds →
Timeline seek bar
(highlights time intervals with which selected scene tags are associated)

Figure 8. Interface for video searching with a seek bar and tags.

Annotation-based application

To confirm the usefulness of annotations acquired in Synvie, we developed a tag-based, scene-retrieval system that is based on tag clouds.[12] When scene tags are generated automatically from annotations, appropriate tags aren't necessarily given to all scenes. And when there are not enough annotations for each video, it's difficult to use conventional search techniques like the exact-matching methods. To solve these problems, we used the tag cloud mechanism for our scene-retrieval system. A user selects some tags from a tag cloud consisting of all scene tags for all videos, and the system displays a list of videos that include these tags. Each video has a seek bar associated with scene tags and has thumbnail images arranged along the time axis (see Figure 8). When the user drags the seek bar, the system displays thumbnail images and scene tags synchronized with the seek bar. By browsing these tags and thumbnail images, the user can understand the content of the video without actually watching it. Moreover, when the user clicks an interesting-looking tag, the tag's temporal location is displayed on the seek bar. Because users can search for target scenes by using scene tags and the temporal distribution of tags and thumbnail images, they can browse and search for scenes interactively.

Evaluation of scene retrieval

We prepared α, β, and γ as data sets for a retrieval experiment. α was a data set of scene

Table 1. Experimental results of video scene retrieval.

Tag set	Creation cost (sec.)	Average number of scene tags	Average retrieval time (sec.)	Cost effectiveness
α	0	153	169	n/a
β	314	55	145	3.48
γ	1,480	53	118	7.18

tags extracted by the automatic tag-extraction method described in the previous "Tag extraction" subsection. β was a data set of scene tags screened from α by using the manual tag screening method described in the "Tag screening" subsection. γ was a data set of scene tags generated by using a manual annotation tool we made for scene tagging for comparative experiments.

Table 1 shows the creation cost (total creation time per content) and the average number of scene tags for α, β, and γ. These values would of course depend on the interface used for creating or screening tags, but we used the same type of interface for each set.

Our targets for the experiment were 27 videos contributed to Synvie. The average length of these videos was 349 seconds. The retrieval question consisted of a blurred thumbnail image and sentences describing the content of the answer scene. Because there was a possibility that the answer could be found easily when the retrieval question included feature keywords in the answer scene, we tried to prevent the inclusion of these keywords. We assumed the situation in which the user searching for a target scene had an uncertain memory.

Examples of question sentences include the following: "A scene in which a parent-child pair of certain animal is stopping in the middle of the road" and "A scene in which a certain person is snow surfing before he slides off the track." The subjects (nine university students) retrieved the scene corresponding to each question, and we measured these retrieval times. They retrieved nine scenes each.

Discussion

In this experiment, all users were able to find the correct target scenes. Table 1 shows the average retrieval time for each data set. The difference between the retrieval times for sets α and β shows that using the tag-screening technique reduces retrieval time significantly.

To evaluate the cost effectiveness of automatic tag extraction, we took the tag-creation time into account when comparing the retrieval times for the β and γ data sets (see Table 1 for the average creation times). Defining cost effectiveness as how much the retrieval time was reduced per 100 seconds of creation time, we found the cost effectiveness of set γ to be 3.48 and that of set β to be 7.18. Thus, we can say a tag-screening mechanism is significant from the viewpoint of cost effectiveness.

Through these experiments, we showed that tags extracted from annotations acquired in Synvie are useful for video-scene retrieval. Because experimental results depend greatly on the amount of annotation, however, more detailed evaluation will be a future project. We showed that the tag-screening technique is useful for improving retrieval efficiency, so we must use tags appropriately and consider factors such as the target video content and retrieval frequency.

Conclusion

The main purpose of Synvie is to provide a novel approach to integrating video sharing, user communication, and content annotation. By associating user communication about videos with the videos themselves, we can uncover more information about scenes and use that information to strengthen user applications. Because a communication space should not be restricted within a single video-sharing site, our scene-quotation system can extend to the entire Web. And because users typically struggle with tag creation and selection, we believe that communication is a motivation toward collaborative tagging.

There are some problems that were not in the research discussed in this article. Currently, we extract only tags. In the future, we intend to extract more semantic information by applying the concept of ontology to scene tags or using more advanced language analysis. In addition, we can construct semantic hypermedia networks based on quotations from video scenes. In these networks comprising weblog entries and video content, the granularity of hyperlinks can be refined from units of videos and weblog entries to units of scenes and paragraphs, the scale of network links can be extended from communities within a single site to communities on the whole Web, and links between contents can be expanded from hyperlinks for navigation to semantic links based on the meanings of quotations. We therefore think that this general method for extracting knowledge about multimedia content from the activities of communities can provide data for other applications. **MM**

References

1. K. Nagao, Y. Shirai, and K. Squire, ''Semantic Annotation and Transcoding: Making Web Content More Accessible,'' *IEEE MultiMedia*, vol. 8, no. 2, 2001, pp. 69-81.
2. D. Howard et al., ''Intelligent Access to Digital Video: Informedia Project,'' *Computer*, vol. 29, no. 5, 1996, pp. 140-151.
3. M. Davis, ''Media Streams: An Iconic Visual Language for Video Annotation,'' *Proc. IEEE Symp. Visual Language*, 1993, pp. 196-202. http://www.w3.org/People/howcome/p/telektronikk-4-93/Davis_M.html.
4. J.R. Smith and B. Lugeon, ''A Visual Annotation Tool for Multimedia Content Description,'' *Proc. SPIE Photonics East, Internet Multimedia Management Systems*, SPIE, 2000, pp. 49-59.
5. K. Nagao, S. Ohira, and M. Yoneoka, ''Annotation-Based Multimedia Summarization and Translation,'' *Proc. 19th Int' l Conf. Computational Linguistics*, Morgan Kaufmann Pub., 2002, pp. 702-708.
6. M. Ben-Ezra and S.K. Nayar, ''Motion-Based Motion Deblurring,'' *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, 2004, pp. 689-698.
7. C. Parker and S. Pfeiffer, ''Video Blogging: Content to the Max,'' *IEEE MultiMedia*, vol. 12, no. 2, 2005, pp. 4-8.
8. D. Yamamoto and K. Nagao, ''iVAS: Web-Based Video Annotation System and Its Applications,'' *Proc. 3rd Int' l Semantic Web Conf., Demonstration Session*, 2004; http://www.nagao.nuie.nagoya-u.ac.jp/papers/pdfs/yamamoto_iswc04.pdf.
9. N.V. Patel and I.K. Sethi, ''Video Shot Detection and Characterization for Video Databases,'' *Pattern Recognition*, vol. 30, no. 4, 1997, pp. 583-592.
10. P.A. Dmitriev et al., ''Using Annotations in Enterprise Search,'' *Proc. 15th Int' l World Wide Web Conf.*, ACM Press, 2006, pp. 811-817.
11. Y. Matsumoto et al., *Japanese Morphological Analysis System ChaSen Version 2.2.1*, 2000; http://chasen.aist-nara.ac.jp/.
12. A.W. Rivadeneira et al., ''Getting Our Head in the Clouds: Toward Evaluation Studies of Tagclouds,'' *Proc. SIGCHI Conf. Human Factors in Computing Systems*, ACM Press, 2007, pp. 995-998.

Related Work

There is a plethora of video-sharing and video-editing services on the Web. Some video-sharing services, such as YouTube, already offer functions for embedding video content in weblogs and commenting on them. Bulletin-board-type communications that comment on content and embed them in weblog entries are used every day. Some Web services, such as Motionbox, Jumpcut, and Kaltura, allow users to edit and share videos easily. Because the aim of these systems is to support communication by sharing videos, they don't provide a mechanism for video annotation. We regard the content of these communications as annotations associated with the video content, but because these annotations are related to the whole video, they can't be used to retrieve certain video scenes. In addition, there are several Web-annotation systems, such as the IBM Efficient Video Annotation (WebEVA) and Google Image Labeler, that can recognize the content of videos or images. WebEVA is a Web-based system that facilitates collaborative annotation on large collections of images and video.[1]

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 might be contradictory evaluations. The ESP Game[2] and Google Image Labeler provides a mechanism that lets users add tags to an image. The approach lets users label random images to help improve the quality of image search results. It's a clever mechanism that requires minimal effort and provides some entertainment value to users. Tags acquired from this mechanism are used for technical improvement of content-based image retrieval.

Because the aim of these annotation systems is to create annotations and applications, they don't support communications. We are proposing another level of content-annotations in video by combining the best from video-editing systems and annotations systems on the Web.

References

1. 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,'' *Proc. 13th ACM Int'l Conf. Multimedia*, ACM Press, 2005, pp. 892-901.
2. L. von Ahn and L. Dabbish, ''Labeling Images with a Computer Game,'' *Proc. CHI Conf. Human Factors in Computing Systems*, ACM Press, 2004, pp. 24-29.

Daisuke Yamamoto is an assistant professor in the Information Technology Center at Nagoya Institute of Technology, Japan. His research interests include Web services, content technologies, and multimedia systems. Yamamoto has a PhD in information science from Nagoya University. Contact him at yamamoto.daisuke@nitech.ac.jp.

Tomoki Masuda is a student in the Graduate School of Information Science at Nagoya University, Japan. His research interests include Web technologies, information retrieval for multimedia content, and video-retrieval systems based on online video annotation. Masuda has a BS in engineering from Nagoya University. Contact him at masuda@nagao.nuie.nagoya-u.ac.jp.

Shigeki Ohira is an assistant professor at EcoTopia Science Institute, Nagoya University. His research interests include multimedia content processing and real-world computing. Ohira has an MIS from the Graduate School of Science and Engineering at Waseda University. Contact him at ohira@nagao.nuie.nagoya-u.ac.jp.

Katashi Nagao is a professor in the Center for Information Media Studies and Graduate School of Information Science at Nagoya University. His research interests include digital content technology, such as authoring, retrieval, transformation, and distribution of digital content, and intelligent agent technology. Nagao has a PhD in computer science from the Tokyo Institute of Technology. Contact him at nagao@nuie.nagoya-u.ac.jp.