iVAS: Web-based Video Annotation System and its Applications

Daisuke Yamamoto
Graduate School of Information Science
Nagoya University
Furo-cho, Chikusa-ku,
Nagoya 464-8603, Japan
yamamoto@nagao.nuie.nagoya-u.ac.jp

Katashi Nagao
EcoTopia Science Institute
Nagoya University
Furo-cho, Chikusa-ku,
Nagoya 464-8603, Japan
nagao@nuie.nagoya-u.ac.jp

Abstract

We present a Web-based video annotation system named iVAS (intelligent Video Annotation Server) that allows users to associate Internet video content with annotations. The system analyzes video content to acquire cut/shot information and color histograms. Then it automatically generates a Web document that allows the users to edit the annotations.

We also present two application systems based on annotations: video retrieval and video simplification. An automatic evaluation method of annotation reliability is also implemented.

1 Introduction

In recent years cost reductions of hard disk drives and the popularization of video editing tools have increased the dissemination of such digital video content as personal recorded video content and video content on the Web. The demand for such applications as video summarization and video retrieval is also greatly increasing. To semantically retrieve or summarize video content, it must be annotated with meta-information. MPEG-7 is one of the hottest annotation methods for multimedia content. Even though much video annotation research has been performed (Davis, 1993; Lin et al., 2002; Nagao et al., 2002), the human cost is still very high since annotation is quite time-consuming. So we believe that a Web-based video annotation system works better when ordinary Web audiences can easily annotate video content with conventional Web browsers.

In our system, even if the quantity of annotations per capita is small, we can still acquire a lot of advanced annotations by merging them.

We also developed applications such as video retrieval and video simplification.

2 iVAS : intelligent Video Annotation Server

In this paper, we present a Web-based video annotation system named iVAS (intelligent Video Annotation Server) whose users can associate any video content on the Internet with various annotations.

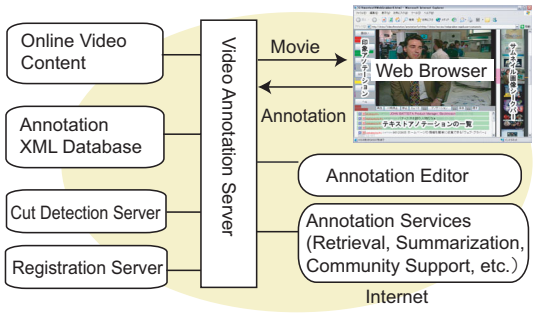

Figure 1: System Configuration

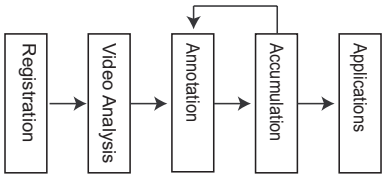

Figure 2: Flow of processing

2.1 System Configuration

Figure 1 shows the system's configuration. Individuals can annotate any video content on the Internet using a video annotation server. Currently video content to be annotated must be registered by a registration server, but in the future such registration will probably be automatically managed by a content collecting server. When a user registers video content, it is analyzed by a "cut detection server" that acquires cut/shot and color histogram information. Then users can associate this content with annotations by using a Web page that edits annotations.

Annotations are stored in the Annotation XML database on the Internet.

2.2 Digital Video Content

The following digital video content accessible by a computer, can be associated with annotations:

- online video content on the Internet
- TV video content on a hard disk video recorder
- DVD video content

To distinguish the content uniquely, we believe that DVD media uses its own ID, Web video content uses

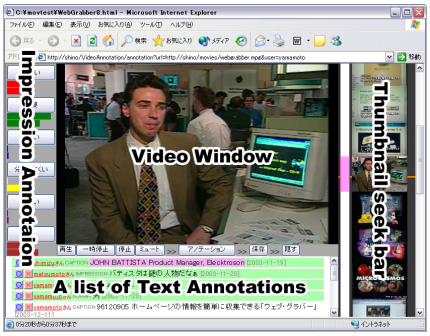

Figure 3: Annotation editing page

Universal Resource Identifiers (URI), and TV contents use its Electronic Program Guide (EPG) information as a key. This system doesn't change the original content and only deals with meta-information. Copyright problems can probably be avoided.

2.3 Video Analysis

When a user annotates video content with a Web browser, it is difficult to analyze video content interactively because of the processing speed. So the system needs to analyze it previously. Therefore, we developed a cut detection server that can get cut times and thumbnail images from the video content.

Since our target is content that features many cuts, it is inapplicable to content with few cuts, such as sporting broadcasts, a home video, or a lecture video. Such contents are divided at fixed time intervals and perform annotation by considering formal cuts.

Operating as a socket server program, this system can handle multiple requesting and communicate with Java or C applications, etc. This server can also store color histogram information in the XML database.

3 Online video annotation by audiences

Audiences can annotate the following services by using iVAS: Text Annotation, Impression Annotation, and Evaluation Annotation.

3.1 iVAS Annotation Editing Page

Figure 3 shows the annotation editing page. The impression annotation interfaces are on the left side of the browser, the video window is in the middle at the top, a list of text annotations is at the bottom in the center, and a "seek bar" that uses thumbnail images is on the left. Since this seek bar can be sought seamlessly by the scroll button of a mouse, we can rapidly find video content. The list of text annotations shows the information relevant to the present shot efficiently sorted by importance.

3.2 Text Annotation

Text Annotation is the annotation mechanism through which text comments are input. When an object in the video window is clicked, the text annotation window

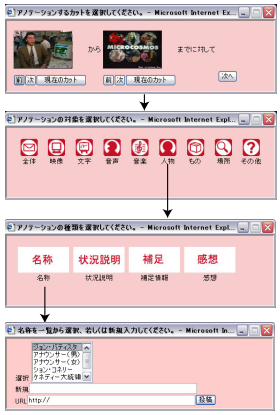

Figure 4: Example of text annotation: a user is annotating: "This person's name is John Batista."

Table 1: Available annotation types

annotation	means	condition
annotation ID	system	automatic
object position	mouse click	implied
time range	selecting shots	essential
object for comments	selecting items	essential
type for comments	selecting items	essential
comment	text-entry	essential
name	text-entry	optional
URL	text-entry	optional
evaluation	O-X buttons	optional

is opened, which simplifies machine processing; consumers can select video shots previously detected by the cut detection server. Furthermore, consumers can choose from the following comments: ALL, MOVIE, CAPTION, VOICE, MUSIC, HUMAN, OBJECT, PLACE, etc. Users can also choose a type of comment for each annotation dialogically: NAME, SITUATION, DESCRIPTION, COMMENT, etc. Consumers evaluate each text annotation by pushing the O-X button. User names, e-mail addresses, annotation Ids, and time information are automatically stored in the XML database.

Table 1 shows the list of annotations that individuals can annotate.

3.3 Impression Annotation

Impression annotation is the mechanism that can associate video content with subjective impressions for content by clicking on a mouse. The number of continuous hits expresses the strength of an impression.

Each impression annotation is set to I_1,I_2 ... I_n. Suppose that it gives the impression information by a normal distribution: $N(\mu, \sigma^2)$ focuses on the clicked time, and each impression I_k is expressed with the following formula.

$$I_k(t) = \sum_{i=ImpressionAnnotation I_k} N(t_i, m)$$

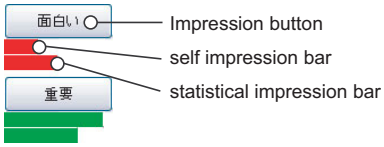

Figure 5: Impression Annotation

where t_i is the media time that carried out the i-th impression annotation about I_k, and m is a constant.

Not only this annotation result but also the result of the entire visitor's annotation is displayed by the bar graph (Figure 3 left, Figure 5). Six is the maximum number of buttons that can be specified by the registration server. Since verification of which button is effective and how many buttons are required is dependent on the kind of content, it is a subject for future research.

4 Annotation Reliability

If the general public posts annotations, they may contain much unreliable information. Therefore, we need to sort out the information by reliability for each annotation.

The reliability calculation method is based on the following principle: "The information from a person who input much reliable information is reliable." First, we calculate a simple evaluation e_k for the annotation A_k. The number of people who evaluated O (good) is set to g_k, and the number of the people who evaluated X (bad) is set to b_k. The score of automatic evaluation, which is decided by whether the description is correct in Japanese or there are any inconsistencies in the selection item, is set to c_k where $-1 < c_k < 1$. If it is decided as good annotation, the value of c_k is large. The better annotations have the larger value of c_k.

Thereby, the simple evaluation e_k becomes:

$$e_k = s \cdot \frac{g_k - a \cdot b_k}{g_k + a \cdot b_k} + t \cdot c_k \quad (1)$$

where s is the evaluation rate by consumers, t is the evaluation rate by the system, and $s + t = 1$. t depends on the accuracy of the mechanical evaluation. a is a coefficient that rectifies the rate of O evaluation and X evaluation, and a becomes the following formula.

$$a = \frac{g_{all}}{b_{all}} \quad (2)$$

g_{all} is the number of O-buttons (good) for all annotations that all the annotators performed, and b_{all} is the number of X-buttons (bad). e_k takes the value of $-1 < e_k < 1$. Although formula 1 is intuitive and combines automatic and human evaluations, the reliability of the human annotator is not considered. Next we calculated the annotator reliability p. G is the number of good evaluations for all annotations that the annotator has annotated, and B is the number of bad evaluations for all annotations the annotator has annotated.

$$p = d(G + B)\frac{G - a \cdot B}{G + a \cdot B} \quad (3)$$

When there are only a few samples, $d(x)$ is a function that holds evaluation values low and is expressed with the following formulas.

$$d(x) = 1 - \exp(-\tau \cdot x) \quad (4)$$

where τ is a constant that decides how much to hold evaluation values down, and $\tau > 0$.

Since the annotated information list is constantly changing, consumers may mistake O-X button evaluations. Therefore, when insufficient consumer evaluations have been gathered, we need to emphasize annotator reliability. On the other hand, when enough consumer evaluations have been gathered, we need to emphasize the consumer reliability. The annotation reliability is expressed as:

$$r_k = (1 - d(g_k + b_k)) \cdot p + d(g_k + b_k) \cdot e_k \quad (5)$$

Annotations with a large value of r_k are relatively reliable, and $-1 < r_k < 1$. In anonymous writing, annotator reliability is the minimum, $p = -1$.

The motivation of calculating annotation reliability is caused by the difficulty of automatic evaluation of the annotation. We assume that the reliability of the information must be undecidable when user evaluations have not been sufficiently collected.

5 Application Using the Annotations

For examples that show the availability of the annotation information, we developed video retrieval and the video simplification servers. Although difficult to handle by automatic analysis, we developed these systems by using the annotations obtained by this system easily.

5.1 Web Video Retrieval based on the Annotations

We developed a content-based video retrieval system based on the following annotations: color histogram information for the shot, text annotation information, impression annotation information, and the evaluation annotation.

First, we used Chasen (Matsumoto et al., 2000) and decomposed retrieval keywords into morphemes: verbs, adjectives, nouns, and unknown words. Unknown words are those not registered in Chasen's dictionary and comprise mainly nouns, English words, or other foreign terms. Next, we calculated a cosine distance score between the retrieval sentence and the text annotation information based on the basic form of each word. We added this score to each shot to which the text annotation applies. Furthermore, we used the impression annotation information and the annotator reliability information. When adding the text annotation score to the shot, the annotation reliability score was also added. We

Figure 6: Results of Web video retrieval

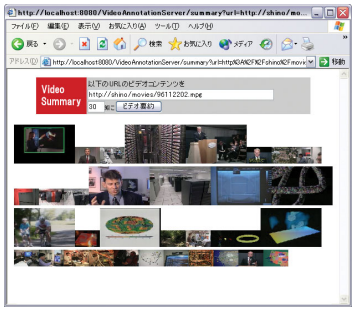

Figure 7: Results of video simplification.

added the highest score to the shot with the larger value of impression annotation. Based on the added score, a user could get a retrieval result sorted by score.

Figure 6 shows an example of the results.

5.2 Web Video Simplification based on the Annotations

We developed a Web video simplification system based on impression annotations. This system simplifies video content based on an easy rule: "the scene that is rising is important." As the score of importance to each scene, we calculate the accumulated value of the impression annotation and the total amount of the text annotation to each scene. After choosing the high importance shots within the specified media time, video content is simplified. Here we are unconcerned with the summary of the story of contents.

Figure 7 shows the results of the video simplification.

6 Experiment and Evaluation

To evaluate the iVAS system's usability and data collection, we ran experiments with 30 college students. We used four video contents edited into 5-minute segments from the video database for image processing evaluation (Babaguchi et al., 2002). Content included news, drama, variety, and cooking programs.

Just by having more users use this system, it is a meaningful system. Then, we conducted a survey on the usability of the system. Although the mother group was college students, many people gave good evaluations for the "easy-to-use" items, showing the ease of iVAS. So

Table 2: Questionnaire results.

	1	2	3	4	5
Text Annotation	0	3	7	12	2
Impression Annotation	0	3	8	11	8
Accuracy for annotaion	0	2	11	16	1
Easy-to-use	0	2	7	10	11
Do you want to use iVAS?	1	1	11	11	6

we concluded that it is easy to use the interface of this system. High evaluation scores to the question "Do you want to use iVAS?" suggest a possibility that many users will use this system.

7 Conclusion

In this paper, we developed a Web-based video annotation system. Consumers can associate any video content on the Internet with annotations. We also developed two application systems based on annotations: video retrieval and video simplification.

Additionally, since our annotation system is open to the public, we must consider the reliability or accuracy of annotation data. We also developed an automatic evaluation, annotation reliability method that utilizes users' feedback. In the future, such fundamental technologies will contribute to the formation of new communities centered around video content.

This system is exhibited at: http://www.nagao.nuie.nagoya-u.ac.jp/ivas/

References

Noboru Babaguchi, Minoru Etoh, Shinichi Sato, Jun Adachi, Akihito Akutsu, Yasuo Ariki, Tomio Echigo, Masahiro Shibata, Heitou Zen, Yuichi Nakamura, Michihiko Minoh, and Takashi Matsuyama. 2002. Video database for evaluating video processing. In *The Institute of Electronics, Information and Communication Engineers, PRMU 2002-30 June*.

M. Davis. 1993. An iconic visual language for video annotation. In *Proceedings of IEEE Symposium on Visual Language*, pages 196–202.

Ching-Yung Lin, Belle L. Tseng, and John R. Smith, 2002. *VideoAnnEx Annotation Tool*. http://www.research.ibm.com/VideoAnnEx/.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, and Masayuki Asahara Kazuma Takaoka. 2000. Japanese morphological analysis system chasen. http://chasen.aist-nara.ac.jp/.

Katashi Nagao, Shigeki Ohira, and Mitsuhiro Yoneoka. 2002. Annotation-based multimedia summarization and translation. In *Proceedings of the Nineteenth International Conference on Computational Linguistics(COLING-2002)*, pages 702–708.