

Automatic indexing and alignment of videos for medical care training

Hiromitsu Niwa¹, Takamori Kawamura¹, Satoshi Tamura², Satoru Hayamizu²

¹ Graduate School of Engineering, Gifu University
1-1, Yanagido, Gifu, JAPAN, 501-1193

² Faculty of Engineering, Gifu University
hiromitsu@hym.info.gifu-u.ac.jp

Abstract. This paper presents a framework of automatic indexing and alignment of audio and video materials used for medical care training, as an example of user-generated media contents. The video materials are obtained using single camera and do not have scene changes which are common in the broadcast video. Speech recognition is conducted to obtain audio metadata. Audio metadata consist of begin and end times, recognition results, and confidence scores. Image metadata, time information and scores computed using optical-flow analysis, are generated. The system integrates the results of audio and image scores according to the treatment procedures. The metadata of the novice trainee are aligned using DP matching with the correct procedures. This integration greatly improved the rate of correct time segmentation from 40% to 83%.

Keywords: Automatic Indexing, Alignment of Video, Medical Care Training, Speech Recognition, Information Integration

1 Introduction

Recent speeding up of internet communications and the spread of high-capacity recording media such as HDs, DVDs, and blue-ray discs encourage the widespread use of multimedia contents such as videos, images, and sound. It is difficult, however, to obtain information efficiently from these data. For that reason, indices for efficient data browsing have been devised. They can indicate what contents a video or a sound has in each part as metadata [Otsuki 03, Xiong 06, Kokaram 06].

Most studies on automatic indexing of video deal with broadcast video, such as baseball games or TV news. These video materials are made with multiple cameras and have scene changes. Previous works use these changes as key events for automatic indexing. Moreover, object detection is generally a difficult task for indexing only from image information of video. As an example of user-generated contents, we have worked on e-learning contents for the university lectures and medical care training [Hashimoto 04, 05, Tamura 06, Kawamura 07].

In medical care training, as shown in Fig. 1, trainees must provide treatment to a patient quickly and correctly. They should review treatment procedures that they

actually performed and determine whether they were correct in comparison to the actions of an instructor, and should learn points for improvement [Tsuyuki 06, Kawamura 07]. Furthermore, a higher understanding is obtainable by referring to treatment procedures used by many other trainees as well as their own. Emergency medical care training courses have been held all over Japan in recent years. However, it takes much time for trainees to visit an actual site and confirm a treatment procedure with an instructor.

To solve these problems, metadata for experts and trainees are assigned by speech recognition and image processing to video about medical care training. Videos about basic life support (BLS) and advanced cardiovascular life support (ACLS) were recorded for the purpose of medical care training in practice. This system is intended to reduce the time for viewing and listening to the contents that users need, and for discovering points to be improved.



Fig.1 View of medical care training instruction.

2 System Configuration

Figure 2 shows the system configuration. Metadata are prepared with information mainly on speech recognition result, and image processing results are used as auxiliary. Metadata include the title, beginning time, and ending time of scenes in a video that are defined based on the scenario of medical courses, such as artificial respiration, and cardiac compression. They enable the efficient browsing of media data. Each process is described below.

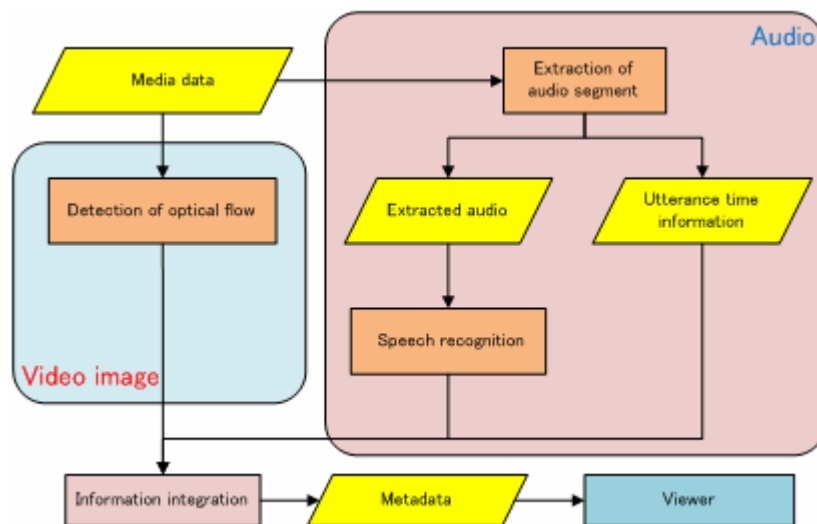


Fig.2 System Configuration

2.1 Speech Recognition

The audio segments with utterances of a video are extracted first. This is for preventing the drop of recognition rate by noises other than utterances, wherein speech is detected with low output power. The speech recognition engine Julius [Kawahara 05] is applied to the extracted audio for speech recognition. The acoustic models used for this study are speaker independent triphone models using the Corpus of Spontaneous Japanese (CSJ). A word N-gram is constructed as a language model, based on texts transcribed from audio in videos from ACLS and BLS courses.

Many professional expressions and abbreviations are contained in the medical field. Therefore, speech models constructed with a common corpus have difficulty in recognizing such words and phrases that are related to a high level of expertise. A word dictionary is prepared in advance. To suppress problems related to the difference in expressions such as "AED" and "defibrillator," they are judged by the system to be similar words in the dictionary for recognition. Speech recognition results provide the utterance time of words and contents in each extracted audio data segments. Consequently, an utterance time chart for all media data is obtained.

The reliability of each word and other information is also recorded simultaneously as XML data, as presented in Fig. 3. Word reliability extends within 0.0 - 1.0. Its value tending to 1.0 implies that few candidates have similar scores and compete with the word concerned, while that approaching 0.0 indicates that many candidate words with similar scores appear. The cmscore, shown in Fig. 3, represents word reliability.

```

<?xml version="1.0" encoding="utf-8" ?>
- <root>
- <SEGMENT begin="1.95" end="7.58" name="bls2\\bls2.0000.wav">
  <WORD id="1" begin="2.24" end="2.62" cmscore="0.96" n-score="-25.01">周囲</WORD>
  <WORD id="2" begin="2.63" end="2.88" cmscore="0.48" n-score="-27.33">が</WORD>
  <WORD id="3" begin="2.89" end="3.21" cmscore="0.53" n-score="-25.42">安全</WORD>
  <WORD id="4" begin="3.22" end="3.51" cmscore="0.28" n-score="-23.85">です</WORD>

```

Fig.3 Speech recognition results.

2.2 Detection of Optical Flow

Optical flow analysis [Horn 81] is adopted as an image processing of video. The shading pattern on each video image is correlated with that on another time frame of video image. The optical flow is a vector expression of its relative shift. The following two factors are considered to affect the result when using optical flow:

- Camera work of the crew
- Gestures of subjects

Because a skilled camera crew operates cameras in TV programs, feature points can be extracted to be available to detect a certain scene, using the feature of camera work. On the other hand, in medical care training videos, often people without skills operate cameras. It is difficult on this occasion to extract feature points from their camera work. Such camerawork also sometimes obscures subtle motions of a subject, thereby making it difficult to use such motion as a feature point.

A video image is divided into three parts along the horizontal axis: in this study, the information of its central part is used to acquire the sum of shifts along the vertical axis. Figure 4 shows that this value changes periodically on the scene of cardiac compression. A distinctive result is obtainable using this value for a video by a camera crew with common movie knowledge that videos are taken with important information positioned at the screen center; for that reason, little effect is imparted by differences in camera work. Moreover, a large part of scenes in this video portray cardiac compression, as shown in Fig. 5. Image feature score $G(t)$ at time t is accordingly defined as

$$G(t) = \frac{50 \times A \times T}{W \times H} \quad (1)$$

where the swing span A , cycle T , the width of the detection part is W , and the height is H .

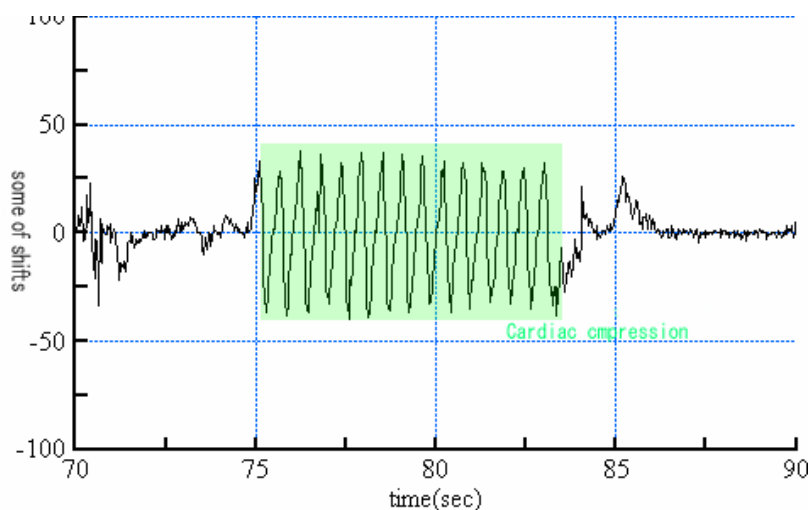


Fig.4 Sum of optical flow along the vertical axis.

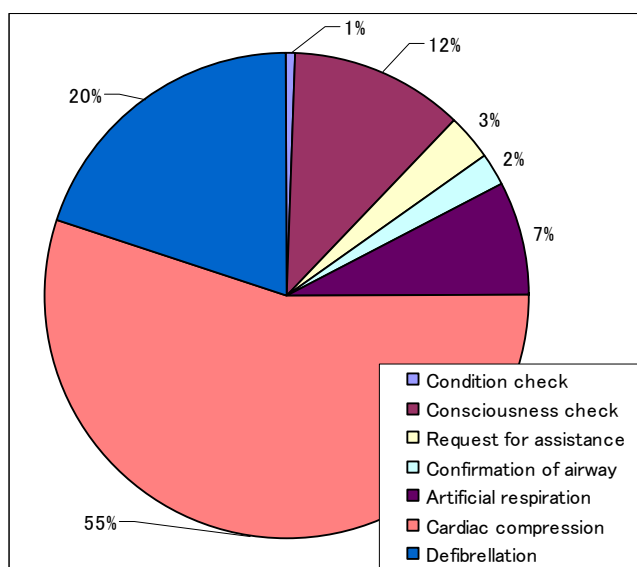


Fig.5 Fraction of treatment procedures in video.

2.3 Information Integration

Metadata corresponding to visual data are constructed based on data obtained through image processing. Keywords are defined for each scene, as listed in Table 1, out of words obtained using the speech recognition result. They are chosen manually from transcribed texts prepared at the construction of the speech model and word dictionary. The score after integration $P(t)$ is computed as

$$P(t) = \begin{cases} (C(t) + G(t)) / 2 & \dots \text{if keyword} = \text{Cardiac compression} \\ C(t) - G(t) & \dots \text{otherwise} \end{cases} \quad (2)$$

as shown in Fig. 6, where image feature score $G(t)$, and the score of a word in speech recognition (cmscore) $C(t)$ at time t . The system integrates the results of audio and image scores according to the treatment procedures. When the speech contents are related to cardiac compression, $P(t)$ is taken as a mean of the image and speech scores; otherwise, $P(t)$ is determined as the speech score subtracted by the image score. When $P(t)$ obtained using this is above the threshold, a scene at time t is determined. This study adopts a threshold of 0.5.

Errors are corrected by performing DP matching between the correct answer data of treatment procedures produced from the scenario of medical care training; scene information is finally obtained.

Table 1 Exemplary Keywords.

Scenes	Keywords
Condition check	周囲 安全
Consciousness check	もしもし 大丈夫
Request for assistance	百十九番 ナースコール
Confirmation of airway	気道 見て
Artificial respiration	フェイスシールド 循環サイン

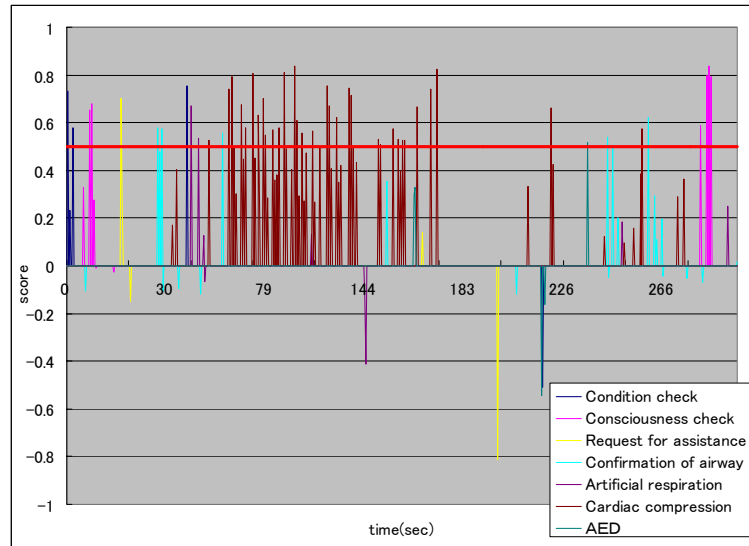


Fig.6 Example of score after integration P(t).

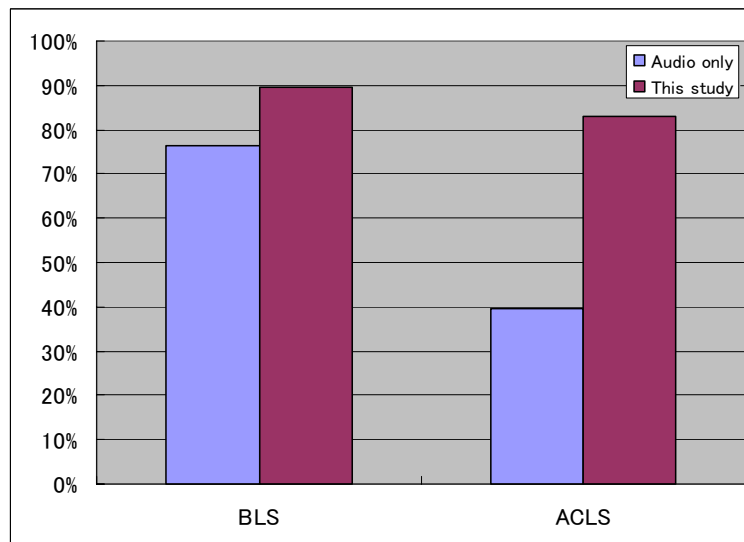


Fig.7 Scene recognition result.

3 Evaluation Experiments

It was evaluated whether the scene segmentation of obtained data were carried out correctly. A recorded video was used for evaluation in which a major speaker is equipped with a pin microphone at ACLS and a BLS courses. The data used are about

30 min long. A video with manual scene segmentation was prepared as a correct answer for reference. Then, data produced only from audio was compared with data in this study. It is difficult to perform scene segmentation using only image information because image feature scores apply only for particular scenes. The correct time of scene segmentation is determined as the time fraction of scenes correctly segmented to the entire data. When only audio metadata was used, the respective correct answer rates were 39.7% in ACLS and 76.6% in BLS, vs. 89.5% and 82.9% when image metadata was combined. Therefore, improvement of the latter is considerable, as depicted in Fig. 7.

The results of automatic indexing and alignment have the information of time and content of educational events for scenes. These events are aligned with correct answer data of treatment procedures. Thus both expert's video with metadata and trainee's video can be viewed simultaneously by the metadata. A browsing and viewer system was implemented. It was tested for the usability and effectiveness for education and the evaluation result was positive.

4 Conclusion

This study conducted automatic indexing and alignment of a medical care training video using audio and image metadata. The use of audio information alone was compared with the combined use of image information. Results show that the combination of image information improved the correct answer rate in scene segmentation.

Future subjects include using more image features, improving integration procedures, and achieving high-accuracy scene segmentation results for each scene. A limited domain of medical care training provides a more constrained set of data for speech recognition.

Today, trainees must attend a course of medical care training and then be evaluated by the instructor. However, future online automatic indexing will allow students to receive lectures from a remote location.

Acknowledgment

The authors would like to thank those with Graduate School of Medicine, Gifu University and Gifu Prefectural Research Institute of Information Technology for invaluable assistance through recording of medical care training video and system construction.

This study was partially supported by Gifu / Ogaki Knowledge Cluster Initiative by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

- [Horn 81] B. K. P. Horn and B. G. Schunk: Determining Optical Flow, *Artificial Intelligence*, Vol. 17, pp.185-204 (1981).
- [Otsuki 03] Katsutoshi Ohtsuki, Katsuji Bessho, Osamu Mizuno, Yoshihiro Matsuo, Shoichi Matsunaga and Yoshihiko Hayashi: Automatic Multimedia Content Indexing with Continuous Speech Recognition, *Tech. Reports, 2003-SLP-047*, Vol. 2003, No.75 (2003. 7), Information Processing Society of Japan.
- [Hashimoto 04] Koji Hashimoto and Satoru Hayamizu: Retrieval of e-Learning Media Contents using Speech Recognition, *Proc. of VSM 2004*, pp.1043-1049 (2004-11).
- [Hashimoto 05] Koji Hashimoto and Satoru Hayamizu: e-Learning Using Speech Recognition - System Evaluation, *Proc. 19th Ann. Conf. of JSAI, (2005.6)*, The Japanese Society for Artificial Intelligence.
- [Kawahara 05] Tatsuya Kawahara and Akinobu Lee: Julius, Continuous Speech Recognition Software, *J. JSAI*, Vol. 20, No. 1, (2005), The Japanese Society for Artificial Intelligence.
- [Xiong 06] Ziyou Xiong, Xiang Sean Zhou, Qi Tian, Yong Rui, Huangm TS: Semantic Retrieval of Video - Review of Research on Video Retrieval in Meetings, Movies and Broadcast News, and Sports, *IEEE Signal Processing Magazine*, Vol.23, No.2, pp.18-27 (2006-03).
- [Kokaram 06] Kokaram, A., Rea, N., Dahyot, R., Tekalp, M., Bouthemy, P., Gros, P., Sezan, I.: Browsing Sports Video: Trends in Sports-related Indexing and Retrieval Work, *IEEE Signal Processing Magazine*, Vol.23, No.2, pp.47-58 (2006-03).
- [Tsuyuki 06] Toshikatsu Tsuyuki, Masayuki Nishioka, Shioko Mitsuhashi, Yumiko Iwase and Taiyo Suganuma: Practice of e-Learning in Medical Care Training, *Proc. 26th JCMI, (2006)*, Japan Association for Medical Informatics.
- [Tamura 06] Satoshi Tamura, Koji Hashimoto, Jiong Zhu, Satoru Hayamizu, Hirotsugu Asai, Hideki Tanahashi and Makoto Kanagawa: Automatic Metadata Generation and Video Editing based on Speech and Image Recognition for Medical Education Contents, *Proc. INTERSPEECH2006, Pittsburgh, Thu2WeO-4 (2006-9)*.
- [Kawamura 07] Takamori Kawamura: Study on Educational Support Using Tagged Information, Master's thesis, Graduate School of Engineering, Gifu University, (2007).