Cross-lingual retrieval of identical news events by near-duplicate video segment detection

Akira Ogawa[1], Tomokazu Takahashi[1], Ichiro Ide[1,2], and Hiroshi Murase[1]

[1] Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464–8601, Japan
{aogawa, ttakahashi, ide, murase}@murase.m.is.nagoya-u.ac.jp
http://www.murase.m.is.nagoya-u.ac.jp/
[2] National Institute of Informatics,
2–1–2, Hitotsubashi, Chiyoda–ku, Tokyo, 101–8430, Japan
ide@nii.ac.jp

Abstract. Recently, for reusing large quantities of accumulated news video, technology for news topic searching and tracking has become necessary. Moreover, since we need to understand a certain topic from various viewpoints, we focus on identical event detection in various news programs from different countries. Currently, text information is generally used to retrieve news video. However, cross-lingual retrieval is complicated by machine translation performance and different viewpoints and cultures. In this paper, we propose a cross-lingual retrieval method for detecting identical news events that exploits image information together with text information. In an experiment, we verified the effectiveness of making use of the existence of near-duplicate video segments and the possibility of improving retrieval performance.

1 INTRODUCTION

1.1 BACKGROUND

Recently, the importance of archiving news videos has increased, and its utility value as a social property has also been focused on. Therefore, for reusing large quantities of accumulated news video, technology for news topic searching and tracking is necessary. When news programs from all over the world become available, understanding a certain topic from various viewpoints increases. So, we focus on identical news event detection in various news programs from different countries.

Currently, the retrieval of news videos is generally performed using only the text information in a similar way to the retrieval of newspaper stories [1, 2]. The retrieval of news video in the same language can be achieved rather easily and with high performance. However, news video broadcasts not only exist domestically but also in other countries as information resources. Utilizing such video resources effectively is our aim. When existing methods, which exploit text information, retrieve video resources in other languages, performance falls due

to low machine translation performance and the necessity for advanced natural language understanding for handling different viewpoints and cultures. Besides, since transcript text called *closed-caption* (CC) does not always accompany every news video, outputs from automatic speech recognition may have to be used, which further complicates the task. In this paper, we circumvent such difficulty in cross-lingual retrieval for detecting identical news events by image information in news video together with text information.

When reporting a certain news event over a long period of time, news programs tend to repeat the same footage. Moreover, video sources of events in other countries are frequently provided by local broadcasters to distribute and broadcast throughout the world. From these properties, when identical video footage is contained in different news videos, a certain relation should exist between them; perhaps they are identical events.

From the above assumption, we detect identical news events by detecting near-duplicate video segments from news video broadcasts in two or more countries in different languages.

1.2 DEFINITION OF TERMINOLOGY

Before describing the details of the process, we must define the terminology of news videos. News video is composed of images and audio, and sometimes closed-caption text. It is structured as follows:

- **Frame:** a still image, which is the minimal unit of a video stream.
- **Shot:** a sequence of frames that are continuous when seen as an image.
- **Cut:** a boundary between shots.
- **Story:** a semantic unit, which contains one event in a news video.
- **Event:** a real-world phenomenon, which occurred at a certain place and time.

2 RELATED WORKS

Many works relevant to the retrieval and tracking of news video using text information have been reported, including bidirectional retrieval by matching TV news videos and newspaper articles [1], topic segmentation, and tracking and thread extraction [3]. However, most use text information from the same language. When extending these methods to consider cross-lingual retrieval between news texts of two or more languages, performance generally falls due to low machine translation and speech recognition performances as well as different viewpoints or cultures. The problem of declining retrieval performance cannot be solved easily.

In our work, we retrieve news video using the existence of near-duplicate video segments together with text information. In this case, the detection of identical video segments in different video sources is important. Regarding this technology,

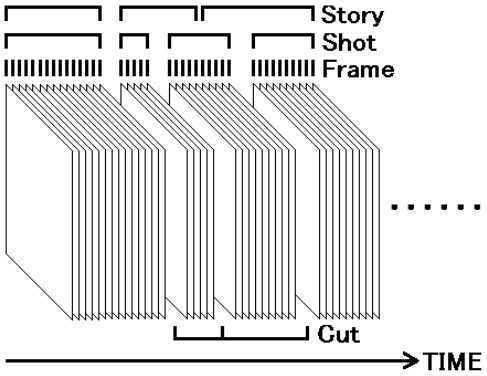

Fig. 1. Composition of broadcast news video

some methods exist, including time-series active search [4], which finds identical video segments from a long video stream. In addition, some methods detect arbitrary pairs of identical video segments using the features of short video fragments, such as a method for the fast retrieval of identical video segments by feature dimension compression using *Principal Component Analysis* (PCA) [5].

3 RETRIEVAL OF IDENTICAL NEWS EVENTS BY NEAR-DUPLICATE VIDEO SEGMENT DETECTION

An overview of the retrieval process of identical news events by near-duplicate video segment detection is shown in Fig. 2. In this experiment, since the input videos are two news programs in different languages, each channel is expressed as C1 and C2, and each language is expressed as L1 and L2, respectively.

3.1 PREPARATION

As preparation for the detection of identical news events, each news video is segmented into stories. Identical events are detected in each story. As story segmentation in news video, various methods have been proposed, such as those that utilize the appearance of anchor-person shots and others that apply NLP methods to CC text [3, 10]. Because we can exploit such segmentation methods, in this paper we assume that stories have already been extracted. In case CC text is not available, output of automatic speech recognition or existing story segmentation methods that refers to the existence of anchor-shots may be used.

3.2 EXTRACTION OF MATCHING REGION

As the first part of the main processing, the matching region, the region in each frame used when matching between videos, is extracted. Here telops and logos are

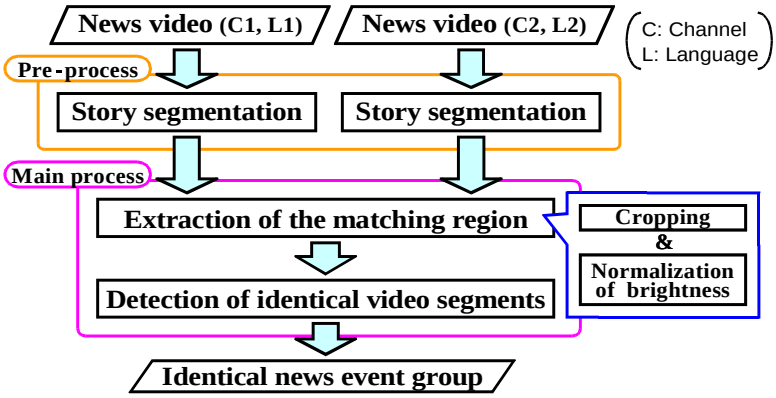

Fig. 2. Retrieval flow of identical news events using image information

often inserted by each news program. Therefore, they may be judged as different video segments despite using the same source video. Setting the threshold to a low degree of similarity is one method for avoiding this. But such a strategy increases the number of false detections. So in our work, since telops and logos are usually not inserted in the center of each frame, the central region of each frame is clipped as the matching region.

However, the problem of matching news video is not limited to such video editing. There is also a problem regarding the color differences of video between broadcast stations. In our work, since RGB pixel values are used as features, we cope with this critical problem by normalizing the brightness by histogram equalization [6] in the whole frame. In the following experiment, normalization is performed by changing the RGB value into a YUV value and then performing histogram equalization on Y, which is the brightness value, and returning it to an RGB value. In the following process, the RGB color features are those after the normalization.

3.3 DETECTION OF IDENTICAL VIDEO SEGMENTS

The brute force frame-by-frame comparison method can be considered the simplest method of detecting arbitrary pairs of identical video segments between videos. However, with this method the computation time increases drastically in proportion to the length of videos. So, we use the fast identical segment detection method proposed in [5] that compresses feature dimensions using PCA. Its basic idea is to detect identical video segments as quickly and accurately as possible by 1) comparing the features of short video fragments instead of a frame, and by 2) compressing the dimension of feature vectors. The first approach makes the comparison efficient and at the same time robust to noise. Meanwhile, the second approach reduces computation time together with i/o time and storage

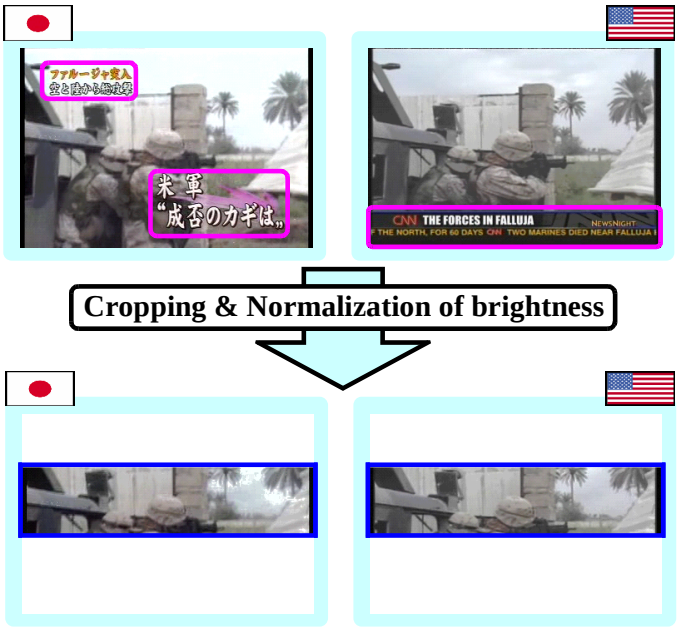

Fig. 3. Extraction of matching region (Top: original frame images, Bottom: extracted region for matching)

space, which is a significant problem when processing a long video stream. As the feature vector, all the RGB values of pixels from a group of consecutive frames are extracted from a video stream and compressed spatiotemporally by PCA. Secondly, the candidates of identical video segments are detected by comparisons using compressed spatiotemporal feature vectors for every video segment. Identical video segments are correctly detected from video segments obtained as candidates by comparison in the original (high-dimension) feature space.

4 EXPERIMENT

Following the process introduced in Section 3, we conducted an experiment that compared the detection results using only text information with the results using image information together with text information and verified the effectiveness of making use of the existence of near-duplicate video segments and the possibility of improving the retrieval performance.

4.1 DATA SET

The experiment used news videos broadcast in Japan (NHK: News7) and in the U.S. (CNN: NewsNight Aaron Brown) in November 2004. The U.S. news

video was provided by TrecVid2005 [7], and the corresponding CC texts were obtained from CNN's web page [3]. One Japanese news video (News7: half-hour) and two English news videos (NewsNight: one hour each) within a ±24-hour time range were treated as one group, where matching processing is performed within the group to cover the time differences between the two countries and the time of the news transfer. In verification of this experiment, the accuracy of the identical news event pair detected by near-duplicate video segment detection is judged manually from the contents of actual video and text data. In the following experiment, the result of four such groups is shown.

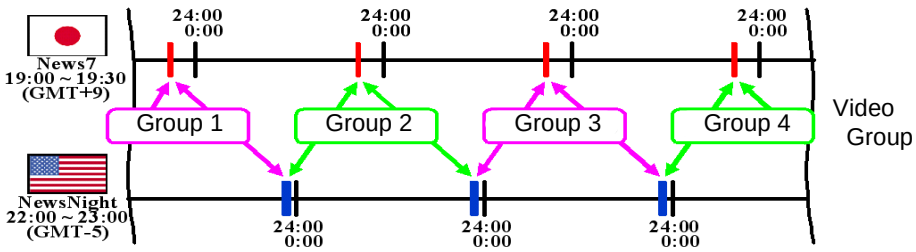

Fig. 4. Relation of broadcast time between a Japanese and U.S. news program

Table 1. Broadcast time of video data in each video group (All data are from November 2004)

	NHK News7	CNN NewsNight (previous)	CNN NewsNight (after)
Group 1	9th 19 : 00 - 19 : 30	8th 22 : 00 - 23 : 00	9th 22 : 00 - 23 : 00
Group 2	10th 19 : 00 - 19 : 30	9th 22 : 00 - 23 : 00	10th 22 : 00 - 23 : 00
Group 3	11th 19 : 00 - 19 : 30	10th 22 : 00 - 23 : 00	11th 22 : 00 - 23 : 00
Group 4	12th 19 : 00 - 19 : 30	11th 22 : 00 - 23 : 00	12th 22 : 00 - 23 : 00

4.2 EXPERIMENTAL PROCEDURE

In the experiment, the detection of identical news events only using image information was performed based on the procedure shown in Fig. 2 to evaluate the detection performance by near-duplicate video segment detection. For story segmentation of the Japanese news video, since CC texts exist, the method proposed by Ide et al. [3] was used, which is based on the similarity of keyword

[3] http://transcripts.cnn.com/TRANSCRIPTS/asb.html

vectors between CC sentences. On the other hand, for story segmentation of the English news, participant shared data of TRECVid2005 were used. For identical video segment detection, parameters were set to detect segments longer than four seconds. Furthermore, to reduce false detection, we precisely matched the candidates from low-dimension matching using color correlograms [9]. As a result, event pairs in which identical video segments exist are detected as identical events. However, because short segments in headlines cause false detections, they were ignored in the evaluation.

Second, we detected identical news events using only text information. This experiment compared detection results using image and text information and using only text information. CC texts of each video were used as text information. Identical news events were detected from them based on the following process. In addition, English text was translated automatically by commercial MT software [4].

Step 1. Story segmentation was performed as in the experiment using image information (See Sect. 3.1).

Step 2. The character strings that serve as keywords were extracted from the CC text of each event by morphological analysis (JUMAN 5.1 [8] was used).
- The following two kinds of character strings were extracted as keywords.
  - Nouns: "personal pronoun" and "formal noun" were ignored, "prefix" and "postfix" in front and behind of words were combined, and the combined noun sequence was regarded as a keyword.
  - Undefined words

Step 3. After extracting *keywords*, the inner product of the frequency-of-appearance vectors was computed among all stories.
- Only a keyword that appears twice or more is matched.
- Only when an inner product is larger than a threshold, the story pair is detected as stories discussing an identical news event.

4.3 RESULTS

The following are the experiment results. The number of identical news events detected with only image information, only text information, and image information together with text information is shown in Table 2, where "Total" denotes the number of detected news events, "Correct" denotes the number of detected identical news events, "Groundtruth" denotes the manually provided answers by us, "Recall" denotes the ratio of "Correct" to "Groundtruth," and "Precision" denotes the ratio of "Correct" to "Total." Similarity by image information in addition to text information is defined as

$$R_{fusion}(S_i, S_j) = \alpha R_{text}(S_i, S_j) + (1 - \alpha) R_{image}(S_i, S_j), \quad (1)$$

where $R_{text}(S_i, S_j)$ and $R_{image}(S_i, S_j)$ are the similarities between stories S_i and S_j by text and image information respectively and α is a constant to balance

[4] The Translation Professional V10 [TOSHIBA]

the importance of their similarities. In this paper, $R_{text}(S_i, S_j)$ is defined as the inner product of keyword frequency vectors between stories S_i and S_j, and $R_{image}(S_i, S_j)$ is defined as

$$R_{image}(S_i, S_j) = \frac{1}{2}\left(\frac{M_{ij}}{L(S_i)} + \frac{M_{ji}}{L(S_j)}\right), \tag{2}$$

where $L(S_i)$ and $L(S_j)$ are the time length of stories S_i and S_j, respectively, M_{ij} is the total time length of identical video segments in S_i that also appear in S_j, and M_{ji} is the total time length of identical video segments in S_j that also appear in S_i. Moreover, the simplest way to select constant α is to let $\alpha = 0.5$. From this result, we can see that four identical news events that couldn 't be detected by text information were detected by near-duplicate video segment detection, and recall became 94% using both types of information. So we conclude that the proposed method makes it possible to obtain better retrieval performance by using image information together with text information.

Table 2. Detection results of identical news events

	Total	Correct	Groundtruth	Recall	Precision
Image only	11	9	18	50%	82%
Text only	35	14	18	78%	40%
Image and text	20	17	18	94%	85%

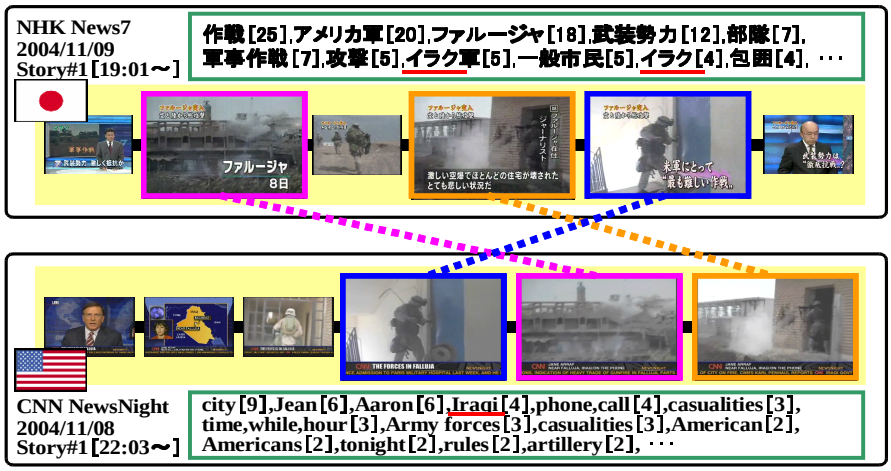

Fig. 5. Example of detected identical news events

4.4 DISCUSSION

First, we consider the detection accuracy of identical news events by image information. The precision of detection by near-duplicate video segments detection was 82% and the recall was 50%. Considering recall, since identical video segments are not always included in identical news events, this result is also sufficient. On the other hand, although precision is 82%, the following two kinds of false detection exist:

- Trailers for upcoming news before commercials
- False detection of identical video segments

Although eliminating the former case is difficult, it is possible to remove the latter case by introducing other image features.

We also considered whether the method can be applied to news videos from countries other than Japan and the U.S. After applying the method to other data from TRECVid, some identical news events were detected from Chinese and Lebanese programs base on the existence of near-duplicate video segments.

5 Conclusion

In this paper, we proposed a cross-lingual retrieval method of identical news events that uses image information in addition to text information and verified the effectiveness of making use of the existence of near-duplicate video segments and the possibility of improving retrieval performance. In addition, although text information is used for the story segmentation of the proposed method, it can also be implemented by a method that uses image information. That is, even when text information cannot be used, detecting identical news events is possible only using image information. Moreover, there are near-duplicate detection methods which use the SIFT features to detect near-duplicate video segments taken from slightly different angles or in different sizes [11]. However, since those methods compare still images, there are problems that huge computation time required to compare all the frame pictures in video segments, thus motion features of objects may not be taken into consideration.

Future study includes a more effective way to use image and text information together by investigating the effect of weighting and the combination of image and text information. Moreover, the implementation of an application that can actually be used for retrieval of identical news events should also be considered.

Acknowledgments

Parts of this work were supported by the Grants-In-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology. The video data used in the experiment were provided by the US National Institute of Standards and Technology, and the National Institute of Informatics through a joint research project. This work is developed based on the MIST library (*http://mist.suenaga.m.is.nagoya-u.ac.jp/*).

References

1. Yasuhiko Watanabe, Yoshihiro Okada, Kengo Kaneji, and Yoshitaka Sakamoto: "Multimedia database system for TV newscasts and newspapers," Advanced Multimedia Content Processing: First Int. Conf., AMCP'98, Osaka, Japan, November 1998. Procs., pp. 208–220, Nov. 1998.
2. Ken Araya, Tatsuhiko Tsunoda, Takumi Ooishi, and Makoto Nagao: "A retrieval method of relevant newspaper articles using word's cooccurrence frequency and location," Trans. of Information Processing Society of Japan, vol. 38, no. 4, pp. 855–862, Apr. 1997. (in Japanese)
3. Ichiro Ide, Hiroshi Mo, Norio Katayama, and Shin'ichi Satoh: "Threading news video topics," Proc. Fifth ACM SIGMM Intl. Workshop on Multimedia Information Retrieval, pp. 239–246, Nov. 2003.
4. Kunio Kashino, Takayuki Kurozumi, and Hiroshi Murase: "A quick search method for audio and video signals based on histogram pruning," IEEE Trans. Multimedia, vol. 5, no. 3, pp. 348–357, Sept. 2003.
5. Ichiro Ide, Kazuhiro Noda, Tomokazu Takahashi, and Hiroshi Murase: "Genre-adaptive near-duplicate video segment detection," Proc. 2007 IEEE Int. Conf. on Multimedia and Expo, pp. 484–487, July. 2007.
6. Graham Finlayson, Steven Hordley, Gerald Schaefer, and Gui Yun Tian: "Illuminant and device invariant colour using histogram equalisation," Pattern Recognition, vol. 38, issue 2, pp. 179–190, Feb. 2005.
7. TREC Video 2005 Evaluation, http://www-nlpir.nist.gov/projects/tv2005/tv2005.html.
8. "Japanese morphological analysis system JUMAN version 5.1," http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html, Sept. 2005.
9. Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih: "Image indexing using color correlograms," Proc. IEEE Conf. on Computer Vision and Pattern Recognition '97, pp. 762–768, June 1997.
10. Yun Zhai, Alper Yilmaz, and Mubarak Shah: "Story segmentation in news videos using visual and text cues," Image and Video Retrieval, Fourth Int. Conf., CIVR 2005, pp. 92–102, July 2005.
11. Chong-Wah Ngo, Wan-Lei Zhao, and Yu-Gang Jiang: "Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation," ACM Multimedia, pp. 845–854, Oct. 2006.