International Journal of Semantic Computing
Vol. 13, No. 1 (2019) 135–155
©World Scientific Publishing Company
DOI: 10.1142/S1793351X19500016

World Scientific
www.worldscientific.com

Summarization of Multiple News Videos Considering the Consistency of Audio-Visual Contents*

Ye Zhang[†] and Ryunosuke Tanishige[‡,§]
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
[†]*zhangy@murase.m.is.nagoya-u.ac.jp*
[‡]*tanishiger@murase.m.is.nagoya-u.ac.jp*

Ichiro Ide
Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
ide@i.nagoya-u.ac.jp

Keisuke Doman
School of Engineering, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, 470-0393, Japan
kdoman@sist.chukyo-u.ac.jp

Yasutomo Kawanishi
Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
kawanishi@i.nagoya-u.ac.jp

Daisuke Deguchi
Information Strategy Office, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
ddeguchi@nagoya-u.jp

Hiroshi Murase
Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
murase@i.nagoya-u.ac.jp

News videos are valuable multimedia information on real-world events. However, due to the incremental nature of the contents, a sequence of news videos on a related news topic could be redundant and lengthy. Thus, a number of methods have been proposed for their summarization. However, there is a problem that most of these methods do not consider the consistency

*This paper was left out from IJSC Special Issue on ISM 2017. It went through the review process of that special issue, and is included as a regular paper in this issue.
[§]Present address: Currently at Denso Corporation, Japan.

between the auditory and visual contents. This becomes a problem in the case of news videos, since both contents do not always come from the same source. Considering this, in this paper, we propose a method for summarizing a sequence of news videos considering the consistency of auditory and visual contents. The proposed method first selects key-sentences from the auditory contents (Closed Caption) of each news story in the sequence, and next selects a shot in the news story whose "Visual Concepts" detected from the visual contents are the most consistent with the selected key-sentence. In the end, the audio segment corresponding to each key-sentence is synthesized with the selected shot, and then these clips are concatenated into a summarized video. Results from subjective experiments on summarized videos on several news topics show the effectiveness of the proposed method.

Keywords: News video; summarization; visual concepts.

1. Introduction

Due to the large amount of video data available online, it has become nearly impossible to view all of them even if we had limited them to those retrieved as relevant to a user's interest. Therefore, there is a large demand for efficiently viewing a large amount of video data in a short period of time, which has led to various research activities in the field including TRECVid's "BBC rushes summarization" task organized in 2007 and 2008 [1].

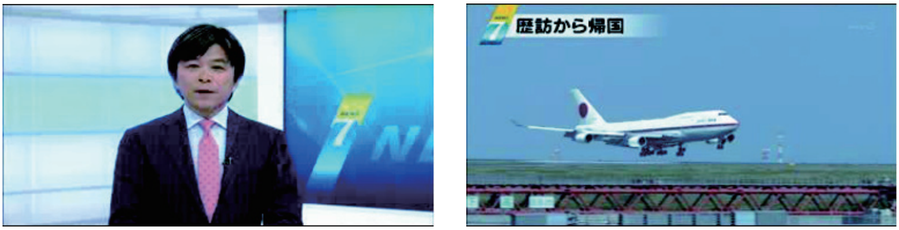
Although it does not contain the most up-to-date works, a comprehensive survey on various video summarization approaches could be found in [2, 3]. After the works covered in these surveys, video summarization based on learning good frames/segments to be included in a summarized video has become a trend. For example, Gygli *et al.* [4] and Potapov *et al.* [5] proposed summarization methods for user generated videos by learning the relations between the original and the summarized videos. Khosla *et al.* proposed a method that selects a frame with good framing learned from Web images based on the assumption that they were photographed so that they should capture the target in a maximally informative way [6]. Meanwhile, Lu and Grauman proposed a method to generate a summarized video by selecting segments such that a subset of visual objects in the previous segment should influence the succeeding segment [7].

While most of these works attempt to summarize a single video, there are some works that attempts to summarize multiple videos. For example, Wang and Merialdo [8] proposed a method for summarizing multiple videos considering the redundancy that exists between them. Ide *et al.* [9] proposed a method for automatically generating a summarized video on a famous person in news by concatenating video footage from important events concerning the person's activities along his/her career. The task setting of the work presented in this paper also falls into this type of video summarization. Meanwhile, there are works on multi-view video summarization, mostly targeting field-sport games [10], stage performances [11], life-logs [12], and lectures [13]. However, in these cases, since the selection of the best angle from multiple synchronized video streams is the key issue, their problem settings differ with the above type of multiple video summarization methods.

Among various kinds of videos, we have been focusing on news videos since they are valuable multimedia information on real-world events. When considering news videos, it is necessary to be handled as a series of events that occur along time rather than individual events. Considering this requirement, Ide *et al.* have proposed an interface that allows the users to track the development of news topics [14] based on a structure built by considering the chronological and semantic relations between news stories. They named the structure a "Topic thread" and according to the statistics shown in the work [14], an average topic thread will be composed of 2,770 s. of video footage. In order to view the development of a news topic from its beginning to the end, it will take on average roughly 45 min. While this will allow us to thoroughly understand the development of a news topic, it will consume too much time for most users who only wishes to roughly grasp an idea on what it was all about. This is the reason that we consider the proposed video summarization method across multiple news videos is necessary even though each news video is essentially a summarized video in itself.

In the case of news videos, since the auditory contents are usually more informative in the sense that they represent the facts concisely compared to the visual contents, the selection of the important auditory contents should precede that of the video contents when generating a summary. This is the main difference of the problem setting compared to the majority of video summarization methods introduced above which generate the summaries solely or mostly based on the selection of visual contents.

In this sense, multiple (text) document summarization methods such as that by Radev *et al.* [15] may serve our purpose better. Thanks to the existence of Closed-Caption (CC) which is a transcript of the auditory contents in a broadcast video, we can process them as text data in most cases. However, in the case of news video summarization, visual contents also need to be considered after the selection of important auditory contents when generating the summarized video due to the fact that they are sometimes inconsistent with corresponding auditory contents, as illustrated in Fig. 1. This is a significant characteristic of news videos that are not



(a) Scene where an anchorperson is speaking (Inconsistent with the auditory contents).

(b) Scene where a plane is landing (Consistent with the auditory contents).

Fig. 1. Visual contents corresponding to the audio contents: "After a series of schedules, Prime Minister Abe arrived at Haneda Airport by Japanese Air Force One."

prominent in most other video genres. As a matter of fact, this issue has already been pointed out by Smith and Kanade [16], and considered in their method in the early days of multimedia contents analysis. However, due probably to the technology available then, their method considered only low-level audio–visual features except for the existence of faces in a scene. Although in their work, it is shown that this approach is effective to some extent, if we do not consider the more high-level visual contents actually present in a scene, it will limit the cases that it could handle properly. Recently, Kumagai *et al.* attempted to detect such inconsistency in news videos based on the relation between audio–visual features [17], but it could only handle monologue (speech) scenes.

Therefore, in this paper, we propose a method of summarizing news videos by selecting shots whose visual contents (Visual Concepts) actually present in a scene are consistent with the auditory contents (key-sentences) decided to be included in the summarized video. We consider that this is especially important when summarizing hours of news videos into a very short video so that users can intuitively grasp the idea of what the news topic was all about; After all, as the saying goes, "Seeing is worth a hundred words" [18].

The remainder of this paper is organized as follows: In Sec. 2, we describe the proposed method. In Sec. 3, we report the result of evaluation experiments. Finally, we conclude the paper in Sec. 4.

2. Summarization of News Videos

We solve the task of summarizing news videos considering the consistency of auditory and visual contents by the process-flow shown in Fig. 2. We expect as input,

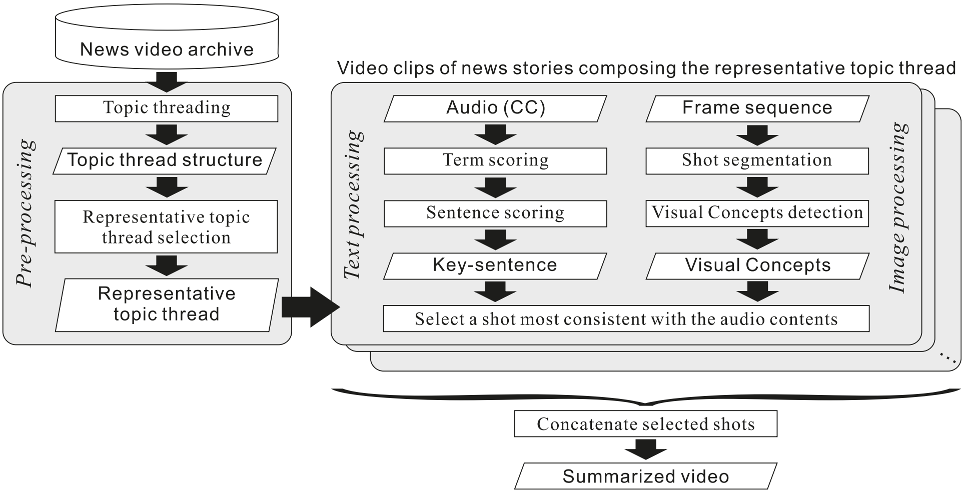

News video archive
Pre-processing
Topic threading
Topic thread structure
Representative topic thread selection
Representative topic thread
Video clips of news stories composing the representative topic thread
Text processing
Audio (CC)
Term scoring
Sentence scoring
Key-sentence
Image processing
Frame sequence
Shot segmentation
Visual Concepts detection
Visual Concepts
Select a shot most consistent with the audio contents
Concatenate selected shots
Summarized video

Fig. 2. Process-flow of the proposed method.

a temporally-ordered sequence of news stories. Both text processing of CC and image processing of each shot corresponding to each news story are applied to the input; First, key-sentences are selected from CC by text processing. Next, visual contents are extracted from shots in the form of Visual Concepts by image processing. Finally, shots whose visual contents are judged as consistent with the key-sentences are selected and then the two are synthesized. In the end, these synthesized video clips are concatenated in order to generate a summarized video.

Here, we make use of Visual Concepts to represent the visual contents of shots. Visual Concepts are representations of high-level visual contents of an image, such as objects, scenes, and activities, whose detections have been actively challenged in the research community in recent years [19–21].

Details of the process are described below, following the definition of terminology that appear in this paper.

2.1. *Terminology*

Below are definitions of important terminology that appear in this paper. First, the following three terms follow the definitions in the Topic Detection and Tracking (TDT) workshop series organized by NIST [22].

- **Event**
  Some incident that occurred at some specific time and place along with all necessary preconditions and unavoidable consequences.
- **News story**
  A topically cohesive segment of news that includes two or more declarative independent clauses about a single event.
- **Topic**
  A seminal event or activity, along with all directly related events and activities.

Next, the following three terms follow the definitions by Ide *et al.* [14].

- **Topic thread**
  A sequence of related news stories chained chronologically. It may contain multiple topics.
- **Topic thread structure**
  A directed graph composed of topic threads originating from a specified news story.

Finally, the following two terms are general concepts in the video processing field.

- **Frame**
  Each of the sequence of still images that compose a video.
- **Shot** Visually continuous sequence of frames.

2.2. *Pre-processing: Selection of a topic thread (news story sequence)*

The proposed method is applied to news videos which are broadcasted with CC. As pre-processing, we construct a directed graph structure representing semantic and temporal relations between news stories called a "topic thread structure" as shown in Fig. 3 by the method proposed by Ide *et al.* [14].

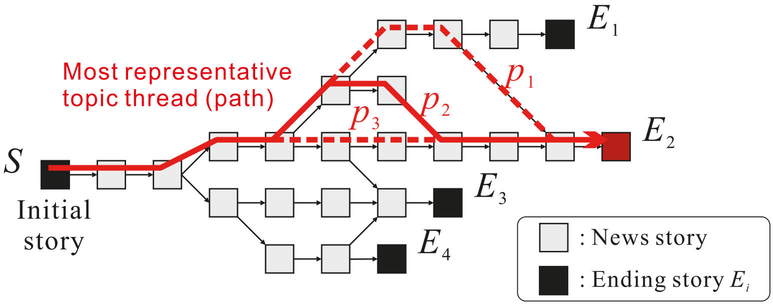


Fig. 3. Example of a topic thread structure originating from the initial story S.

Then, we estimate the most representative sequence of news stories (a path in the graph structure) called a "topic thread" from the structure by Kato *et al.*'s method [23].

For simplicity, in this paper, we will consider these processes as pre-processing and expect a sequence of already selected news stories as input to the proposed method. For reference, each method is briefly introduced below. Please refer to corresponding publications for details.

2.2.1. *Topic threading*

Topic threading is a task to link related news stories according to the development of news topics. Since one topic can develop into another topic, a topic thread could include multiple topics by linking locally related news stories. In the proposed method, we applied the method proposed by Ide *et al.* [14] to first construct a topic thread structure originating from a specified news story, and then select one representative topic thread from it. The topic thread structure represents local relations between individual news stories as directed edges, and at the same time a global trend of topics as a directed graph.

This method first segments news stories based on similarity between adjacent CC sentences, and then constructs the topic thread structure based on the temporal order and the similarity between news stories measured by the cosine distance between the distributions of term frequencies that appear in the CC of news stories. Note that although this process could employ visual information, we only used text information obtained from CC in our work.

2.2.2. *Selection of a representative topic thread*

Although the topic thread structure represents rich information on the topics originating from the specified news story, for a normal user who wishes to understand the main stream of the development of topics, he/she does not necessarily need to be provided with all the related news stories. Thus, we decided to select the most representative topic thread (path) that connects the initial story (root node; S in Fig. 3) and one of the ending stories (leaf nodes; E_i $(i = 1, 2, \ldots, I)$).

For this, we applied the method proposed by Kato *et al.* [23]. This method first selects the most representative ending story E from the leaf nodes. As shown in Fig. 3, there are cases that multiple paths (p_1, p_2 and p_3) exist between the root node S and the selected leaf node E. In such cases, the most representative path is selected among all possible paths.

First, to select the most representative ending story E, for each ending story E_i $(i = 1, 2, \ldots, I)$, the following features are considered with Features 4 and 5 originally proposed by Sawai *et al.* [24]:

(1) Similarity of proper nouns in S and E_i.
(2) Elapse of days between S and E_i.
(3) The number of news stories along the path between S and E_i. In the case where multiple paths exist, the one with the most number of news stories is selected for the counting.
(4) The order of E_i among the sequence of news stories broadcasted in the same program on that day.
(5) Video length of E_i.

After normalization of each feature, their weighted sum is calculated for each ending story, and the one with the highest score (in this example, $E = E_2$) is selected.

Next, in the case where multiple paths exist between S and E (p_1, p_2, and p_3), the same algorithm as above is applied recursively by setting each of the news stories directly succeeding the news story where the branch begins as the initial node, and the news story where the branched paths merge as the ending story.

After the above process, we obtain a news story sequence from the selected most representative topic thread, which is used as the input to the proposed method described in the remaining part of this section.

2.3. ***Text processing***

A key-sentence is selected from each news story based on scoring of terms and sentences considering the rarity of terms and also the appearance of Visual Concept vocabularies. Details of the process are described below.

2.3.1. *Assignment of term scores*

First, the proposed method assigns a score to each noun within a news story. In general, Term Frequency Inverse Document Frequency (TF-IDF) is used as a measure to calculate the rarity of each term in a document. Here, TF-IDF is calculated as the proportion of the frequency of a noun in a news story to the inverse-log frequency of news stories in which the noun appears. In detail, nouns that appear in each news story are scored as follows:

(1) Apply morphological analysis to CCs of all news stories that compose a topic thread, and extract nouns.

(2) Calculate the Term Frequency $\mathrm{tf}_{i,j}$ (TF) in a news story and the Inverse Document Frequency idf_i (IDF) as

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{1}$$

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d|i \in d, d \in D\}|}. \tag{2}$$

Here, $n_{i,j}$ is the frequency of occurrences of a noun i in news story j, and $\sum_k n_{k,j}$ the sum of occurrences of all nouns in news story j. D indicates a set of news stories, $|D|$ the number of news stories, d a news story, and $|\{d|i \in d, d \in D\}|$ the number of news stories that include noun i.

(3) Calculate the TF-IDF value of noun i in news story j as

$$\omega_{i,j} = \mathrm{tf}_{i,j} \cdot \mathrm{idf}_i. \tag{3}$$

In this way, we can assign higher scores to rare nouns, and assign lower scores to frequent nouns when they appear in text. This value $\omega_{i,j}$ is called the "term score" hereafter.

In order to obtain the IDF scores, we analyzed the appearance of nouns in 46,870 CC sentences from the news program "NHK News7" broadcasted during March 16, 2001 and May 16, 2013 in Japan. The analysis involved 2,124,569 nouns in total, and yielded IDF scores for 98,340 unique nouns.

2.3.2. *Assignment of sentence scores*

Next, the proposed method assigns a score to each sentence. We considered that sentences which contain more nouns that represent visual phenomena are more important for the summarization of news videos. Therefore, each sentence is assigned a score based on the term scores and the number of nouns that exist in the Visual Concepts' vocabulary. The sentence score is calculated as the average of term scores in each sentence as

$$S_l = \frac{N+1}{|\mathrm{W}_l|} \sum_{i \in \mathrm{W}_l} \omega_{i,j}. \tag{4}$$

Here, W_l is a set of all nouns in sentence l, and N is the number of nouns that exist in the vocabulary of the set of Visual Concepts employed. Since many synonyms appear in news text, we used a Japanese version of the WordNet [25] to expand the limited vocabulary.

2.3.3. *Selection of the key-sentence based on sentence scores*

Finally, a sentence with the highest sentence score is selected as the key-sentence representing each news story. Note that sentences starting with “First” or “Next” were discarded since they tend to be introductory statements at the beginning of a news story with not much information.

2.4. *Image processing*

After shot segmentation, Visual Concepts are detected from each shot. Details of the process are described below.

2.4.1. *Shot segmentation*

First, an input video is segmented into shots. In the following experiment, we simply used difference of the HSV color histograms between two frames for detecting initial shot boundaries. Note that miss-detections were manually corrected before applying the subsequent process.

2.4.2. *Detection of Visual Concepts*

Then, Visual Concepts are detected from each shot. Considering the computational cost, we assumed that the Visual Concepts detected from the first frame should represent the entire shot. Here, two kinds of Visual Concept detectors are prepared; The first one is a popular detector from the ILSVRC2014 challenge [27] based on the GoogLeNet model which uses a deep neural network [26], and the second one is an SVM-based detector for detecting person-related Visual Concepts trained by the authors. The latter was prepared since the image categories defined in the ILSVRC2014 challenge are confined to only 1,000 out of the 32,326 categories in the entire ImageNet image thesaurus [28] with almost no human-related ones except for clothing; We considered that we should also analyze more detailed attributes of a person, since we are targeting news videos where various people play important roles. Details of each detector and the combination of the two detectors are introduced below.

Visual Concepts from the ILSVRC2014 challenge The GoogLeNet network model for the ILSVRC2014 challenge has shown high performance on image classification (6.67% top-five error) and detection (43.9% mean average precision). We decided to inherit the image classes defined in the challenge as Visual Concepts in this paper, and relied on the publicly-available network model for their accurate detection.

Additional person-related Visual Concepts In order to decide the additional person-related Visual Concepts, we first analyzed the appearance of terms in 46,870 CC sentences from the news program "NHK News7" broadcasted during March 16, 2001 and May 16, 2013 in Japan. Table 1 shows an excerpt of the most frequent person-related terms that appeared in the CCs. To create Visual Concept detectors related to them, image categories in the ImageNet thesaurus were manually assigned as shown in the table. As a result, we defined the following 10 person-related Visual Concepts: Person, Female, Male, Child, Patient, Student, Athlete, Leader, Journalist, and Policeman. For each of them, an SVM classifier was trained using images from corresponding categories in ImageNet as shown in Table 2. We used the Soft-Weighted Bag-of-Features (SWBoF) [29] representation of SIFT features [30] as an input to the SVM. The size of the training data used to train each detector is shown in the right-most column of the table. Note that the same numbers of images were prepared as positive and negative samples for the training of each person-related Visual Concept, thus indicated as "×2" in the table.

Table 1. Frequent person-related terms that appeared in CCs of news programs and ImageNet categories assigned to them (Excerpt from the top 50 frequent nouns).

Rank	Person-related term	Frequency	ImageNet category
8	Person	5996	Person
17	Minister	4514	Political leader
34	Police	3582	Policeman
42	Representative	3393	Representative
43	Reporter	3330	Journalist
50	Prime Minister	3142	Political leader

Table 2. ImageNet categories used for training person-related Visual Concepts.

Visual Concept	ImageNet categories	Training data
Person[a]	Person, Individual, Someone, Somebody, Mortal, Soul	9180 × 2
Female	Female, Female person	1371 × 2
Male	Male, Male person	1407 × 2
Child	Child, Baby	1424 × 2
Patient	Patient	531 × 2
Student	Student, Pupil, Educate	1024 × 2
Athlete	Athlete, Jock	1028 × 2
Leader	Military leader, Religious leader, Political leader, Civic leader, Spiritual leader	1287 × 2
Journalist	Journalist	292 × 2
Policeman	Policeman	816 × 2

Note: [a]The training data for the "Person" Visual Concept is composed of those from all the other person-related Visual Concepts.

Once trained, person-related Visual Concepts are detected from an input image by applying each SVM classifier to it; Visual Concepts corresponding to classifiers yielding confidence values are larger than a pre-defined threshold (defined experimentally as 0.5 in the following) are detected.

For reference, the performance of each person-related Visual Concept detector examined over testing data (200 positive and 200 negative samples separated from the training data) was shown in Table 3. Note that F-score stands for the harmonic mean of precision and recall. We can see that most of the categories show relatively high F-scores.

Table 3. Performance of the person-related Visual Concept detectors.

Visual Concept	F-score	Visual Concept	F-score
Person	0.75	Student	0.76
Female	0.08	Athlete	0.91
Male	0.09	Leader	0.66
Child	0.44	Journalist	0.66
Patient	0.66	Policeman	0.05

Combination of the two Visual Concept detectors In the end, the results from the two Visual Concept detectors are combined as shown in Fig. 4; First, the Person detector in the person-related Visual Concept detector is applied to a given image. If it detects a person, the remaining nine person-related Visual Concept detectors are applied. If not, the ILSVRC2014-based Visual Concept detector is applied.

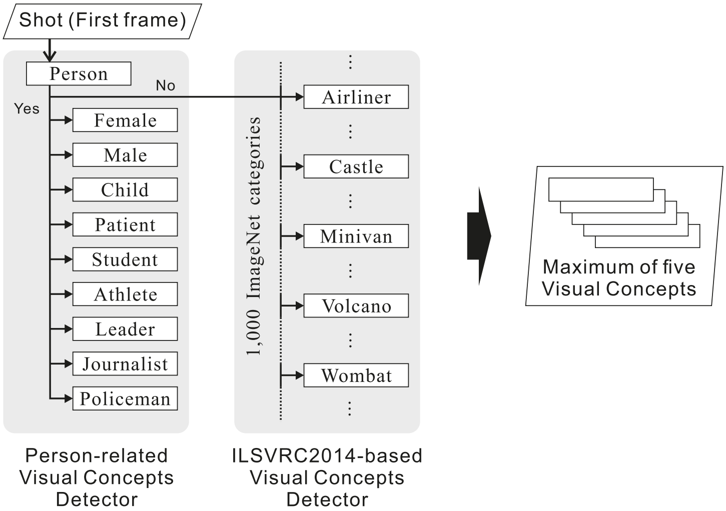

Shot (First frame)
Person
No
Yes
Female
Male
Child
Patient
Student
Athlete
Leader
Journalist
Policeman
1,000 ImageNet categories
Airliner
Castle
Minivan
Volcano
Wombat
Maximum of five Visual Concepts
Person-related Visual Concepts Detector
ILSVRC2014-based Visual Concepts Detector

Fig. 4. Combination of the two Visual Concept detectors.

In the end, a maximum of five Visual Concepts out of the detected ones are labelled to each shot.

2.5. *Generation of a summarized video*

A summarized video is generated based on the key-sentences and the Visual Concepts detected from each shot.

2.5.1. *Selecting shots containing visual contents consistent with auditory contents*

The criteria for selecting shots are as follows:

(1) Select a shot in the news story which includes the most number of Visual Concepts that correspond to the selected key-sentence. Here, in order to cover a wider range of vocabulary, we use WordNet to expand the vocabulary of Visual Concepts in the same way as in Sec. 2.3.
(2) If multiple shots were selected by the above criterion, choose the one closest to the selected sentence in time.

2.5.2. *Editing a summarized video*

Next, the proposed method generates a summarized video in the following procedure:

(1) Sort the key-sentences in temporal order.
(2) Extract audio segments corresponding to the key-sentences.
(3) Synthesize the extracted audio segments and shots selected in Sec. 2.5.1, then concatenate them as shown in Fig. 5.

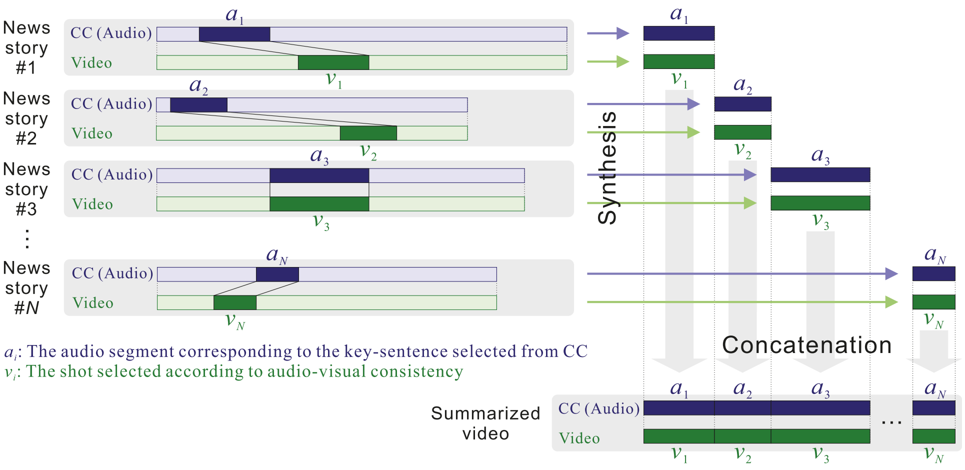


Fig. 5. Editing of a summarized video.

Note that when the length of the selected shot is shorter than that of the audio segment, the next candidate shots according to the selection criteria are concatenated to the selected one. On the other hand, when the length is longer, the remaining part of the selected shot is eliminated.

3. Experiments

In order to evaluate the effectiveness of the proposed method, we performed two experiments:

- Experiment 1: Evaluation of the text-image consistency.
- Experiment 2: Evaluation of the generated summarized video.

Details of each experiment are reported in the following sections.

3.1. *Dataset*

As the video dataset, we used the NII TV-RECS news video archive [31] which consists of news video from a daily evening program "NHK News 7" in Japan, recorded since March 16, 2001 with a total volume of approximately 3000 h of news videos to date.

The news videos used in the experiments were in MPEG-1 format (Resolution: 352 × 180 pixels, Frame-rate: 25 fps) for Experiment 1, and in TS format (Resolution: 1400 × 1080 pixels, Frame-rate: 30 fps) for Experiment 2. Both formats were accompanied with CC in Japanese, which are translated into English for reference, in the examples introduced hereafter.

3.2. *Evaluation on the effect of considering the audio-visual consistency*

3.2.1. *Experimental conditions*

First, we conducted a subjective experiment in order to evaluate the effect of considering the audio–visual consistency by the proposed method. Fifteen video segments corresponding to CC sentences that included at least one Visual Concept vocabulary, and whose contents were not consistent with the corresponding visual contents were selected from news videos broadcasted between January 14 and May 12, 2013.

Thirty-two Computer Science major students in their twenties were asked to freely view the original video segment corresponding to the CC sentence and that synthesized by the proposed method for all 15 CC sentences. The presentation order of the videos was changed randomly per CC sentence in order to avoid bias.

The subjects were then asked to choose the one that visually represented the auditory contents better. Note that the bottom part of the frames was trimmed since it tends to contain excessive text information that could interfere with the purpose of this experiment.

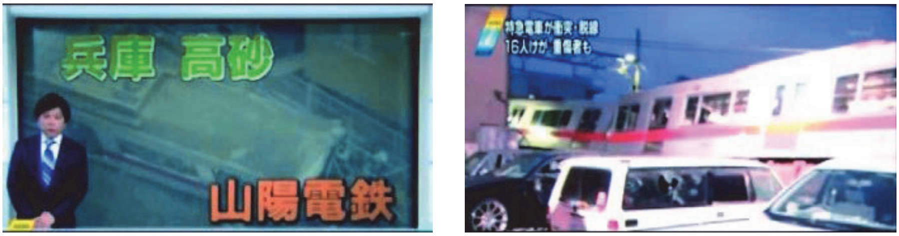
3.2.2. *Results and discussions*

The result was evaluated by the "selection ratio" defined as the ratio of the number of subjects who preferred the video synthesized by the proposed method versus the original video, to the total number of subjects. Table 4 shows the selection ratio of the proposed method.

Table 4. Selection ratio of the video synthesized by the proposed method considering audio–visual consistency versus the original video.

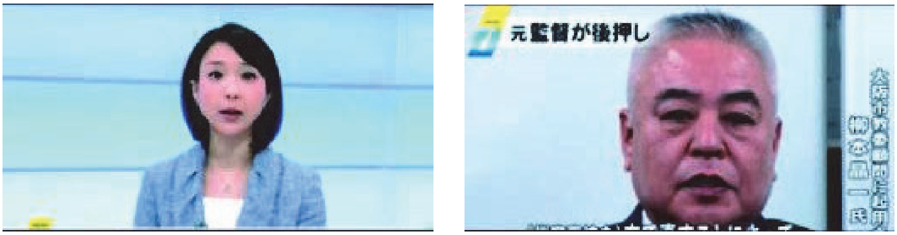
Sentence ID	1	2	3	4	5	6	7	8	9	10
Selection ratio	56%	47%	34%	72%	93%	47%	84%	69%	44%	78%
Sentence ID	11	12	13	14	15			Average		
Selection ratio	59%	91%	84%	84%	94%			**69.2%**		

Figures 6–8 show the original videos and videos synthesized by the proposed method whose selection ratio by the subjects were higher than 90%. We can see that the proposed method synthesized shots that visually represent the contents of the key-sentence better than the original shots.



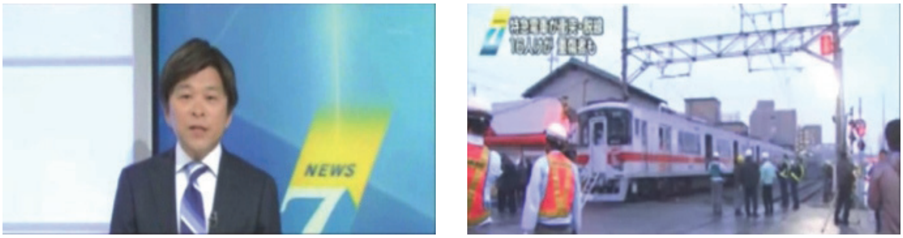
(a) Original (b) Proposed method

Fig. 6. Shots selected for sentence #5: "A Sanyo Railway express train derailed and crashed into the platform after running into a truck at a crossing."



(a) Original (b) Proposed method

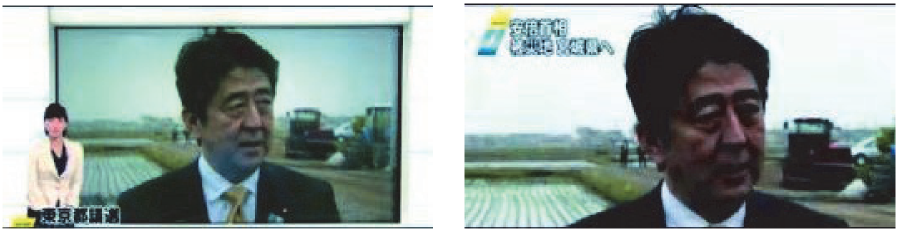
Fig. 7. Shots selected for sentence #12: "The Osaka municipal education committee decided to appoint Mr. Shoichi Yanamoto, the ex-manager of the National volleyball team, as their advisor."



(a) Original (b) Proposed method

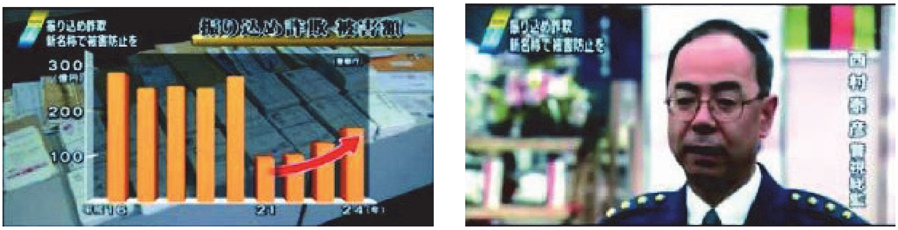
Fig. 8. Shots selected for sentence #15: "The derailed train ran onto the platform."

Figures 9–12 show the original videos and videos synthesized by the proposed method whose selection ratio by the subjects were lower than 50%, i.e. the majority of the subjects considered that the original video was better than the video synthesized by the proposed method.



(a) Original (b) Proposed method

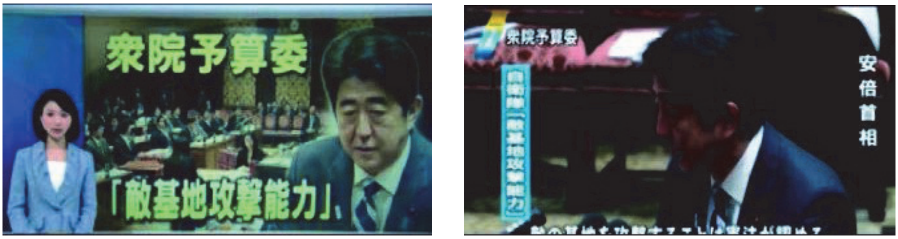
Fig. 9. Shots selected for sentence #2: "The Prime Minister expressed that the upcoming Tokyo Metropolitan Assembly election will be a barometer for the public opinion about his economic policy."



(a) Original (b) Proposed method

Fig. 10. Shots selected for sentence #3: "The police will start boosting the campaign to prevent damage."

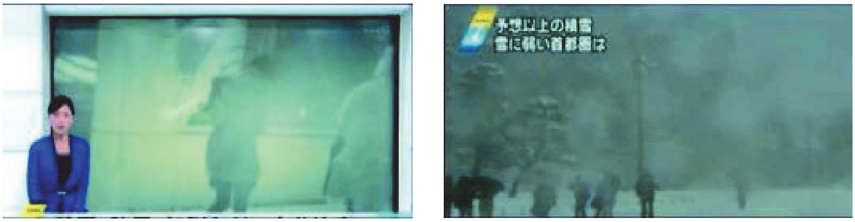
For sentences #2 (Fig. 9) and #6 (Fig. 11), it seems that the inconsistency of the speaker and the voice was considered unnatural. The actual voice in the audio stream was uttered by not the subject of the video but an anchorperson, which seemed to have given the subjects an unnatural impression. To solve this problem, as a special



(a) Original

(b) Proposed method

Fig. 11. Shots selected for sentence #6: "The Prime Minister expressed that at this moment he has no intention to do so, but we may need to consider possessing the ability to attack enemy bases according to international situations."



(a) Original

(b) Proposed method

Fig. 12. Shots selected for sentence #9: "Due to the rapid development of low pressure, snow has fallen along the Pacific coast of Kanto-Koshin and Tohoku areas, and strong wind is blowing along the coast."

case, we should consider using the speaker's original voice when the selected shot contains a monologue, instead. This could be detected by, for example, Kumagai *et al.*'s method [17].

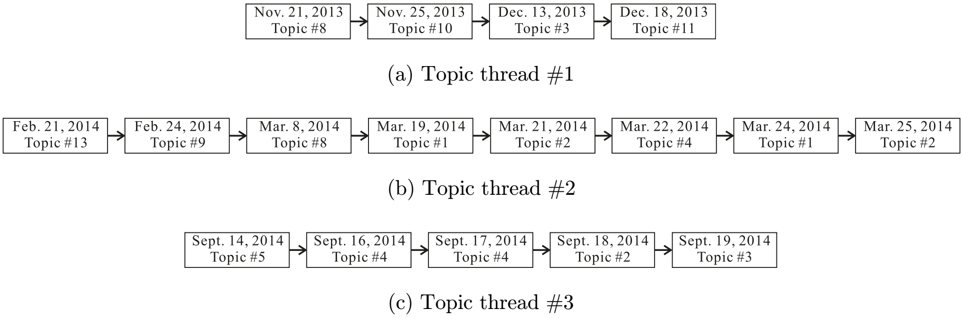
For sentence #3 (Fig. 10), it seems that the statistics of damage caused by fraud shown as a graph was more informative to the subjects than seeing the portrait of the chief of the police that is visually consistent with the term "police". To solve this problem, we should consider putting higher priority on showing graphs and tables.

For sentence #9 (Fig. 12), it seems that the contents of the caption overlayed on the screen being inconsistent with that of the sentence has given the subjects an unnatural impression. Although we trimmed the bottom part of the frame as mentioned in Sec. 3.2.1, there were still many captions in other areas of the frame. To solve this problem, we should detect, recognize, and analyze the overlayed captions and either erase them or consider their contents.

3.3. *Evaluation of the generated summarized video*

3.3.1. *Experimental conditions*

Next, we conducted a subjective experiment to evaluate the quality of the generated summarized video with three topic thread structures shown in Fig. 13. Details of each



(a) Topic thread #1

(b) Topic thread #2

(c) Topic thread #3

Fig. 13. Sequence of news stories that compose the topic threads used in the experiment.

topic thread are shown in Table 5. News videos broadcasted between November 2013 and September 2014 were used as the source.

Table 5. Details of topic threads used in the experiment.

ID	Initial story	Topic	Number of news stories	Number of sentences	Length [s]
1	Nov. 21, 2013	TEPCO's nuclear power restart	4	54	641
2	Feb. 21, 2014	2014 Crimean crisis	8	131	1644
3	Sept. 14, 2014	Scottish independence	5	194	1855

Table 6. Factors considered in each summarization method.

Method	Text processing (Term score calculation)	Image processing (Visual consistency)
Comparison 1	TF-IDF	Not considered (Original shot)
Comparison 2	TF-IDF	Considered
Comparison 3	TF-IDF + Existence of Visual Concept vocabulary	Not considered (Original shot)
Proposed	TF-IDF + Existence of Visual Concept vocabulary	Considered

We compared the proposed method with three different summarization methods shown in Table 6 for comparison. Details of each method are as follows:

- Comparison method 1: Summarization by concatenating shots *originally corresponding to each sentence* selected according to term scores based *only on TF-IDF.*
- Comparison method 2: Summarization by concatenating shots *visually consistent with each sentence* selected according to term scores based *only on TF-IDF.*
- Comparison method 3: Summarization by concatenating shots *originally corresponding to each sentence* selected according to term scores based *on TF-IDF and existence of Visual Concept vocabulary.*
- Proposed method: Summarization by concatenating shots *visually consistent with each sentence* selected according to term scores based *on TF-IDF and existence of Visual Concept vocabulary.*

Fifteen Computer Science major students in their twenties were presented with pairs of videos summarized by all four methods in random order, and then asked to select the one among the pair whose visual contents represented the auditory contents better. In order to reduce the bias on prior knowledge on the topic, the subjects were allowed to familiarize themselves with each topic by reading articles related to the topic before performing the evaluation, in the Japanese version of the online encyclopedia Wikipedia.[a]

3.3.2. *Results and discussions*

The result was evaluated by the "selection ratio" defined as the ratio of the number of subjects who selected the result by the proposed method versus each of the comparison methods, to the total number of subjects.

Table 7 shows the lengths of the videos generated by each summarization method. Note that since the pairs of Comparison methods 1 and 2, and Comparison method 3 and the Proposed method take the same key-sentence selection strategy, respectively, the length for the summarized videos generated by each pair of methods is the same. Also note that the length of the summarized video depends on the length of the audio segment corresponding to the selected key-sentences. Although we could roughly adjust the length of the summarized video by selecting multiple sentences per news story, the proposed method does not expect to generate a summarized video with a length specified in advance.

Table 7. Lengths [sec.] of the videos generated by each summarization method. The percentage in the parentheses indicates the summarization rate.

Method	Topic thread #1	Topic thread #2	Topic thread #3
Comparison 1 Comparison 2	54 (8%)	101 (6%)	73 (4%)
Comparison 3 Proposed	78 (12%)	88 (5%)	43 (2%)

Table 8 shows the selection ratio of the Proposed method versus Comparison methods, respectively. We can see that the proposed method was more effective than Comparison methods 2 and 3 for topics #1 and #2. In these cases, we confirmed that

Table 8. Selection ratio of the proposed method versus comparison methods for the generated summarized videos.

Versus method	Topic thread #1 (%)	Topic thread #2 (%)	Topic thread #3 (%)	Average (%)
Comparison 1	60	60	60	60
Comparison 2	80	80	80	80
Comparison 3	80	100	20	67

[a] Wikipedia, http://ja.wikipedia.org/.

considering the consistency of auditory and visual contents was effective for selecting the key-sentences and selecting shots consistent with them.

However, the selection ratio was significantly low for topic #3 which contained many monologue scenes. We consider the primary cause for this was the inconsistency of the speaker and the voice like in the case of sentences #2 and #6 in the previous experiment in Sec. 3.2.

Meanwhile, the Proposed method was less effective versus Comparison method 1. We consider the primary cause for this was that multiple shots were selected according to the exceptional rule in 2.5.2, which seemed to have given the subjects an unnatural impression. To solve this problem, we should consider selecting only one shot for each sentence and rectify its length by adjusting the frame rate or the audio pitch, instead.

4. Conclusion

In this paper, we proposed a method for summarizing news videos along a news topic thread structure. The proposed method synthesized shots based on the consistency of auditory and visual contents, and generated a summarized video by concatenating them.

Future work includes improvement of Visual Concept detectors and introduction of more detailed Visual Concepts so that the proposed method could perform better. We will also consider incorporating additional editing rules to generate the summarized video. Evaluation on a larger dataset including videos from different news programs, should also be performed.

Acknowledgments

We would like to thank the subjects who participated in the experiments. Parts of this research were supported by the Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and a joint research program with the National Institute of Informatics, Japan. This paper is a revised and extended version of the paper published at the 19th IEEE Int. Symposium on Multimedia (ISM2017).[b]

References

[1] P. Over, A. F. Smeaton and G. Awad, The TRECVID 2008 BBC rushes summarization evaluation pilot, in *Proc. 2nd ACM TRECVid Video Summarization Workshop*, 2008, pp. 1–20.

[2] A. G. Money and H. Agius, Video summarization: A conceptual framework and survey of the state of the art, *J. Vis. Commun. Image Represent.* **19**(2) (2008) 121–143.

[3] B. T. Truong and S. Venkatesh, Video abstraction: A systematic review and classification, *ACM Trans. Multimedia Comput. Commun. Appl.* **3**(1) (2007) 3.1–3.37.

[b] © 2017 IEEE. Reprinted with permission from [32].

[4] M. Gygli, H. Grabner and L. Van Gool, Video summarization by learning submodular mixtures of objectives, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3090–3098.

[5] D. Potapov, M. Douze, Z. Harchaoui and C. Schmid, Category-specific video summarization, in *Proc. 13th European Conf. Computer Vision*, Vol. IV, 2014, pp. 540–555.

[6] A. Khosla, R. Hamid, C.-J. Lin and N. Sundaresan, Large-scale video summarization using Web-image priors, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2698–2075.

[7] Z. Lu and K. Grauman, Story-driven summarization for egocentric video, in *Proc. 2013 IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2714–2721.

[8] F. Wang and B. Merialdo, Multi-document video summarization, in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2009, pp. 1326–1329.

[9] I. Ide and F. Nack, Explain this to me! — A study on automatic recompilation of broadcast news video —, *ITE Trans. Media Technol. Appl.* **67**(2) (2013) 101–117.

[10] X. Wang, K. Hara, Y. Enokibori, T. Hirayama and K. Mase, Personal viewpoint navigation based on object trajectory distribution for multi-view videos, *IEICE Trans. Inf. and Syst.* **E101-D**(1) (2018) 193–204.

[11] M. K. Saini, R. Gadde, S. Yan and W.-T. Ooi, MoViMash: Online mobile video mashup, in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 139–148.

[12] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song and Z.-H. Zhou, Multi-view video summarization, *IEEE Trans. Multimedia* **12**(7) (2010) 717–729.

[13] C. Zhang, Y. Rui, J. Crawford and L.-W. He, An automated end-to-end lecture capture and broadcasting system, *ACM Trans. Multimedia Comput., Commun. Appl.* **4**(1) (2008) 6:1–6:23.

[14] I. Ide, T. Kinoshita, T. Takahashi, H. Mo, N. Katayama, S. Satoh and H. Murase, Efficient tracking of news topics based on chronological semantic structures in a large-scale news video archive, *IEICE Trans. Inf. Syst.* **E95-D**(5) (2012) 1288–1300.

[15] D. R. Radev, H. Jing and M. Budzikowska, Centroid-based summarization of multiple documents, *Info. Process. Manage.* **40**(6) (2004) 919–938.

[16] M. A. Smith and T. Kanade, Video skimming and characterization through the combination of image and language understanding, in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 1997, pp. 775–781.

[17] S. Kumagai, K. Doman, T. Takahashi, D. Deguchi, I. Ide and H. Murase, Speech shot extraction from broadcast news videos, *Int. J. Semant. Comput.* **6**(2) (2012) 179–204.

[18] G. Ban and Z. Ban (eds.), Biography of Zhao Chongguo and Xin Qinji (in Chinese), in *Book of Han*, Vol. 69.

[19] T. Deselaers and A. Hunbury, The visual concept detection task in ImageCLEF2008, in *Evaluating Systems for Multilingual and Multimodal Information Access*, Lecture Notes in Computer Science, Vol. 5706 (2009), pp. 94–109.

[20] M. J. Huiskes, B. Thomee and M. S. Lew, New trends and ideas in visual concept detection, in *Proc. 11th ACM SIGMM Int. Conf. Multimedia Information Retrieval*, 2010, pp. 527–536.

[21] S. K. Divvala, A. Farhadi and C. Guestrin, Learning everything about anything: Webly-supervised visual concept learning, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 3270–3277.

[22] J. G. Fiscus and G. R. Doddington, Topic detection and tracking evaluation overview, in *Topic Detection and Tracking: Event-based Information Organization* (Kluwer Academic Publishers, 2002), pp. 17–31.

[23] K. Kato, I. Ide, D. Deguchi and H. Murase, Estimation of the representative story transition in a chronological semantic structure of news topics, in *Proc. 4th ACM Int. Conf. Multimedia Retrieval*, 2014, pp. 487–490.

[24] R. Sawai, H. Senoo and Y. Shishikui, Proposal and evaluation of a method for calculating news value for creating news digest (in Japanese), *IPSJ Trans. Databases* **2**(2) (2009) 1288–1300.

[25] F. Bond, T. Boldwin, R. Fothergill and K. Uchimoto, Japanese SemCor: A sense-tagged corpus of Japanese, in *Proc. 6th Int. Global Wordnet Conf.*, 2012, pp. 9–16.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Going deeper with convolutions, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, ImageNet large scale visual recognition challenge, *Int. J. Comput. Vision* **115**(3) (2015) 211–252.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A large scale hierarchical image database, in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[29] Y.-G. Jiang, C.-W. Ngo, and J. Yang, Toward optimal Bag-of-Features for object categorization and semantic video retrieval, in *Proc. 6th ACM Int. Conf. Image and Video Retrieval*, 2007, pp. 494–501.

[30] D. G. Lowe, Object recognition from local scale-invariant features, in *Proc. 7th IEEE Int. Conf. Computer Vision*, 1999, Vol. 2, pp. 1150–1157.

[31] N. Katayama, H. Mo, I. Ide and S. Satoh, Mining large-scale broadcast video archives towards inter-video structuring, in *Advances in Multimedia Information Processing*, Lecture Notes in Computer Science, LNCS Vol. 3332, 2004, pp. 489–496.

[32] I. Ide, Y. Zhang, R. Tanishige, K. Doman, Y. Kawanishi, D. Deguchi and H. Murase, Summarization of news videos considering the consistency of auditory and visual contents, in *Proc. 19th IEEE Int. Symp. Multimedia*, 2017, pp. 193–199.