Summarization of News Videos Considering the Consistency of Auditory and Visual Contents

Ichiro Ide*, Ye Zhang†¶, Ryunosuke Tanishige†¶, Keisuke Doman‡*,
Yasutomo Kawanishi*, Daisuke Deguchi§*, and Hiroshi Murase*

*Graduate School of Informatics, Nagoya University, Nagoya, Japan, 464–8601
Email: {ide, kawanishi, murase}@i.nagoya-u.ac.jp
†Graduate School of Information Science, Nagoya University, Nagoya, Japan, 464–8601
Email: {zhangy, tanishiger}@murase.m.is.nagoya-u.ac.jp
‡School of Engineering, Chukyo University, Toyota, Japan, 470–0393
Email: kdoman@sist.chukyo-u.ac.jp
§Information Strategy Office, Nagoya University, Nagoya, Japan, 464–8601
Email: ddeguchi@nagoya-u.jp
¶Currently at DENSO CORPORATION.

***Abstract*—Since news videos are valuable sources of multimedia information on real-world events, there is a demand for viewing them efficiently. However, there is a problem that summarization methods based on auditory contents do not take into account the visual contents. In the case of news videos, due to its presentation style where audio contents and visual contents do not necessarily come from the same source, this could severely decrease the amount of informative visual contents included in the generated summarized video. Thus, we propose a method for summarizing a sequence of news videos considering the consistency of both 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 then selects a shot within the news story whose "Visual Concepts" detected from the visual contents are the most consistent with the key-phrase. Finally, the audio segment corresponding to each key-phrase is overlapped onto the selected shot, and then concatenated to generate a summarized video. The effectiveness of the proposed method was confirmed on several news topics through a subjective experiment.**

I. Introduction

Due to the tremendous amount of video data available online, it has become nearly impossible to view all of them even if we had limited to those retrieved as relevant to a user's interest. Therefore, there is a demand for efficiently viewing a large amount of video data in a short period of time, which has lead 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] and [3]. Since then, 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 based on learning the relations between the original video and the summarized video. Khosla et al. proposed a method to select a frame with good framing learned from Web images based on the assumption that they were taken 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 consider summarizing a single video, there are some works that try to summarize multiple videos. For example, Wang et al. [8] proposed a method for summarizing multiple videos considering the redundancy that exist between them. Our task setting introduced below also falls into this type of video summarization.

Among various kinds of videos, we have been focusing on news videos since they are valuable sources of 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. Following this necessity, we have proposed a structuring method that allows the users to track the development of news topics [9] based on both the chronological and semantic relations of news stories. We named such a structure "Topic thread" and according to the statistics shown in the work [9], an average topic thread will be composed of 2,770 sec. 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 minutes. While this allows us to thoroughly understand the development of a news topic, it is too time consuming for most users who only wishes to roughly grasp an idea on what it was all about. This is the reason why 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

978-1-5386-2937-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ISM.2017.33

(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."

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 Radev et al. [10]'s method may serve our purpose better. Thanks to the existence of Closed-Caption (CC) which are transcripts of the auditory contents in a broadcast video, we can process its auditory contents as text data in most cases. However, in the case of news video summarization, visual contents also needs 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 feature of news videos that are not prominent in most other video genres. As a matter of fact, this issue has already been pointed out by Smith and Kanade [11], 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 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 [12], 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 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" [13].

The remainder of this paper is organized as follows: In Section II, we describe the proposed method. In Section III, we report the result of an evaluation experiment. Finally, we conclude the paper in Section IV.

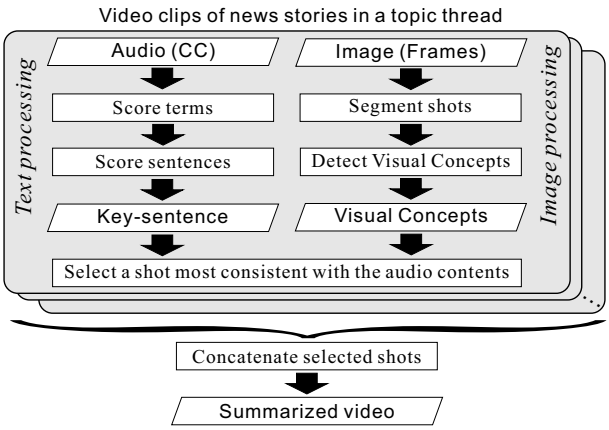


Fig. 2. Process-flow of the proposed method.

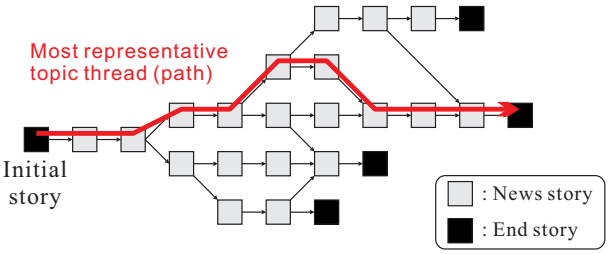


Fig. 3. Example of a topic thread structure.

II. Summarization of News Videos

A. Approach

We solve the task of summarizing news videos considering the consistency of auditory and visual contents by the process-flow shown in Fig. 2. First, a temporally ordered sequence of news stories is input. Then, both text processing of CCs in each news story and image processing of each shot, are conducted. Finally, shots whose visual contents are consistent with the key-sentences selected from the auditory contents (i.e. CCs) are selected, and then they are concatenated in order to generate a summarized video. Here, we use Visual Concepts to represent the visual contents of shots. Visual Concepts represent high-level visual contents of an image, such as "person", "animal", and "location".

B. Pre-processing: Selection of a Topic Thread (News Story Sequence)

The proposed method is applied to news videos which are broadcasted with CCs. 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 us previously [9]. This method constructs the structure based on the temporal order and the cosine distance between term frequency distributions that appear in CCs of news stories.

Then, we estimate the most representative sequence of news stories called a "topic thread" from the structure by Kato et

al.'s method [14]. This method selects the most representative topic thread (path) that connects the initial story (root node) and one of the end stories (leaf nodes) according to certain features of the topic thread such as duration and density of news stories.

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

C. Text Processing

First, the proposed method assigns a score to each word 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 term in a news story to the inverse-log frequency of news stories in which the word appears. In detail, terms 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

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

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

Here, $n_{i,j}$ is the frequency of occurrences of a term i in news story j, and $\sum_k n_{k,j}$ the sum of occurrences of all terms 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 term i.

3) Calculate the TF-IDF value of the term as

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

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

Next, the proposed method assigns a score to each sentence. We considered that sentences which contain more terms 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 terms that exist in the Visual Concepts' vocabulary. The sentence score is calculated as the average of term scores in each sentence as

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

Here, W_l is a set of all terms in sentence l, and N is the number of terms that exist in the Visual Concepts' vocabulary. Since many synonyms appear in news text, we used a Japanese version of the WordNet [15] to expand the Visual Concepts' vocabulary.

Finally, a sentence with the highest score is selected as the key-sentence representing each news story.

TABLE I
IMAGENET CATEGORIES USED FOR TRAINING PERSON-RELATED CLASSIFIERS

Classifier	ImageNet categories
Person	Person, Individual, Someone, Somebody, Mortal, Soul
Female	Female, Female person
Male	Male, Male person
Child	Child, Baby
Patient	Patient
Student	Student, Pupil, Educate
Athlete	Athlete, Jock
Leader	Military leader, Religious leader, Political leader, Civic leader, Spiritual leader
Journalist	Journalist
Policeman	Policeman

D. Image Processing

First, an input video is segmented into shots. In the following experiment, we simply used HSV color histogram for detecting shot boundaries. Then, Visual Concepts are detected from each shot. Considering the computational cost, we assumed that the Visual Concepts detected from the first frame represent the entire shot.

In the following experiment, two kinds of Visual Concept detectors were used. The first one was the GoogLeNet detector which uses a deep neural network [16]. Although this detector can detect various Visual Concepts, we considered that we should also analyze more detailed attributes of a person, since we are targeting news videos where people play important roles.

Thus, we constructed additional Visual Concept detectors related to a person. We defined the following ten person-related Visual Concepts after analyzing the term frequency in the CCs of news programs during 2001 and 2013: 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 the ImageNet database [17] as shown in Table I. Here, we used the Soft-Weighted Bag-of-Features (SWBoF) [18] representation of SIFT features [19]. Once trained, the classifiers are applied in two steps; First only the person classifier is applied. If a person is detected, then the remaining nine attribute classifiers are applied.

The two detectors were used in combination, and the top five classes of the detection results are used in the subsequent processing for each shot.

E. Generation of a Summarized Video

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

1) Selecting Shots 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

TABLE II
SELECTION RATIO OF THE PROPOSED METHOD FOR TEXT-IMAGE CONSISTENCY

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

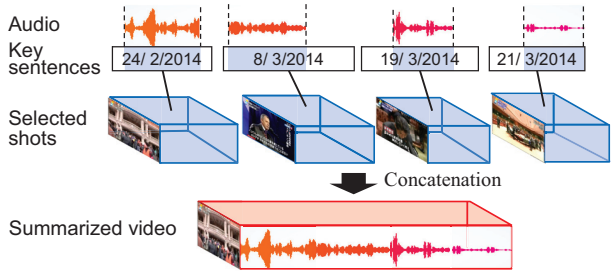


Fig. 4. Synthesizing audio segments and shots.

(a) Original (b) Proposed method

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

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 Section II-C.

2) If multiple shots were selected by the above criterion, choose the one closest to the selected sentence in time.

2) Editing a Summarized Video: Next, the proposed method generates a summarized video in the following procedure.

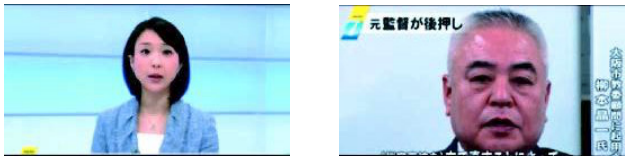
1) Sort the selected sentences in temporal order.
2) Extract audio segments corresponding to the key sentences.
3) Synthesize the extracted audio segments and shots selected in II-E1, then concatenate them as shown in Fig. 4.

Note that when the length of the selected shot is shorter than the audio segment's length, 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.

III. EXPERIMENTS

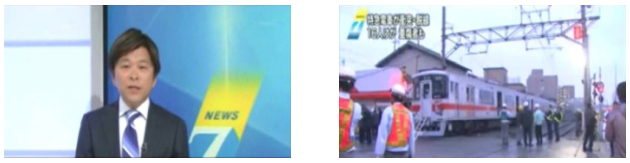
In order to evaluate the effectiveness of the proposed method, we performed two experiments:

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



(a) Original (b) Proposed method

Fig. 6. 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. 7. Shots selected for sentence 15: "The derailed train ran onto the platform."

Details of each experiment are reported in the following sections.

A. Dataset

As the video dataset, we used the NII TV-RECS news video archive [20] which consists of news video from a daily evening program "NHK News 7" recorded since Mar. 2001 with a total volume of approximately 3,000 hours of footage to date.

B. Evaluation of the Text-Image Consistency

1) Experimental Conditions: First, we conducted a subjective experiment to evaluate the quality of the text-image consistency by the proposed method. Fifteen 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, and these videos were used as the source.

Thirty-two Computer Science major students in their twenties were asked to freely view the original shot corresponding to the sentence and the shot selected according to the visual consistency with the sentence, and then asked to choose the one that visually represented the contents of the sentence better. Note that the bottom part of the video was trimmed since it tends to contain too much text information that could interfere with the purpose of this experiment.

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 to the total

TABLE III
TOPIC THREAD STRUCTURES USED IN THE EXPERIMENT

Topic ID	Initial story	Topic	# of news stories	# of sentences	Length [sec.]
1	November 21, 2013	TEPCO's nuclear power restart	4	54	641
2	February 21, 2014	2014 Crimean crisis	8	131	1,644
3	September 14, 2014	Scottish independence	5	194	1,855

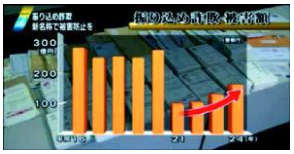
TABLE IV
FACTORS CONSIDERED IN EACH SUMMARIZATION METHOD

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

(a) Original

(b) Proposed method

Fig. 8. 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. 9. Shots selected for sentence 3: "The police will start boosting the campaign to prevent damage."

number of subjects. Table II shows the selection ratio of the proposed method.

Figures 5 to 7 show the original shots and shots selected by the proposed method for sentences whose selection ratio by the subjects were higher than 90%. We can see that the proposed method selected shots that visually represent the contents of the key-sentence better than the original shots.

Figures 8 to 11 show the original shots and shots selected by the proposed method for sentences whose selection ratio by the subjects were lower than 50%.

For sentences 2 (Fig. 8) and 6 (Fig. 10), 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 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 [12].

For sentence 3 (Fig. 9), it seems that the statistics of damage

(a) Original

(b) Proposed method

Fig. 10. 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. 11. 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."

caused by fraud shown as a graph was more informative to the subjects than seeing 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. 11), 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 III-B1, 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.

C. Evaluation of the Generated Summarized Video

1) Experimental Conditions: Next, we conducted a subjective experiment to evaluate the quality of the generated summarized video with three topic thread structures detailed in Table III. News videos broadcasted between November 2013

TABLE V
LENGTHS [SEC.] OF THE VIDEOS GENERATED BY EACH SUMMARIZATION METHOD. THE PERCENTAGE IN THE PARENTHESES INDICATES THE SUMMARIZATION RATE.

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

TABLE VI
SELECTION RATIO OF THE PROPOSED METHOD VS. COMPARISON METHODS FOR THE GENERATED SUMMARIZED VIDEOS

Vs. method	Topic 1	Topic 2	Topic 3	Average
Comparison 1	60%	60%	60%	60%
Comparison 2	80%	80%	80%	80%
Comparison 3	80%	100%	20%	67%

and September 2014 were used as the source.

We compared the proposed method with three different summarization methods shown in Table IV 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 shown 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 Wikipedia articles related to the topic before performing the evaluation.

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 vs. each of the comparison methods, to the total number of subjects.

Table V 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 VI shows the selection ratio of the proposed method vs. 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 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 Section III-B.

Meanwhile, the proposed method was less effective vs. comparison method 1. We consider the primary cause for this was that multiple shots were selected according to the exceptional rule in II-E2, 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, instead.

IV. CONCLUSION

In this paper, we proposed a method for summarizing news videos along a news topic thread structure. The proposed method selected 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.

ACKNOWLEDGMENT

First of all, we would like to thank the subjects who participated in the experiments. We would also like to thank the National Institute of Informatics, Japan for allowing us the usage of the NII TV-RECS news video archive. 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.

REFERENCES

[1] P. Over, A. F. Smeaton, and G. Awad, "The TRECVid 2008 BBC rushes summarization evaluation," in *Proceedings of the Second ACM TRECVid Video Summarization Workshop*. Vancouver, BC, Canada: ACM, October 2008, pp. 1–20.

[2] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, February 2008.

[3] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 3, no. 1, pp. 3.1–3.37, 2007.

[4] M. Gygli, H. Grabner, and L. V. Gool, "Video summarization by learning submodular mixtures of objectives," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA: IEEE, June 2015, pp. 3090–3098.

[5] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Computer Vision —ECCV2014 Thirteenth European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8694. Zurich, Switzerland: Springer-Verlag, September 2014, pp. 540–555.

[6] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using Web-image priors," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, June 2013, pp. 2698–2075.

[7] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, June 2013, pp. 2714–2721.

[8] F. Wang and B. Merialdo, "Multi-document video summarization," in *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo*. New York, NY, USA: IEEE, June 2009, pp. 1326–1329.

[9] 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 Transactions on Information and Systems*, vol. E95-D, no. 5, pp. 1288–1300, May 2012.

[10] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based summarization of multiple documents," *Information Processing and Management*, vol. 40, no. 6, pp. 919–938, November 2004.

[11] M. A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding," in *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*. San Juan, Puerto Rico: IEEE, June 1997, pp. 775–781.

[12] S. Kumagai, K. Doman, T. Takahashi, D. Deguchi, I. Ide, and H. Murase, "Speech shot extraction from broadcast news videos," *International Journal of Semantic Computing*, vol. 6, no. 2, pp. 179–204, June 2012.

[13] G. Ban and Z. Ban, Eds., *Biography of Zhao Chongguo and Xin Qinji (in Chinese)*, ser. Book of Han, vol. 69.

[14] K. Kato, I. Ide, D. Deguchi, and H. Murase, "Estimation of the representative story transition in a chronological semantic structure of news topics," in *Proceedings of the Fourth ACM International Conference on Multimedia Retrieval*. Glasgow, Scotland, UK: ACM, April 2014, pp. 487–490.

[15] F. Bond, T. Boldwin, R. Fothergill, and K. Uchimoto, "Japanese SemCor: A sense-tagged corpus of Japanese," in *Proceedings of the Sixth International Global Wordnet Conference*, Matsue, Shimane, Japan, January 2012, pp. 9–16.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA: IEEE, June 2015, pp. 1–9.

[17] 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," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, April 2015.

[18] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Toward optimal Bag-of-Features for object categorization and semantic video retrieval," in *Proceedings of the Sixth ACM International Conference on Image and Video Retrieval*. Amsterdam, The Netherlands: ACM, July 2007, pp. 494–501.

[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2. Kerkyra, Corfu, Greece: IEEE, September 1999, pp. 1150–1157.

[20] 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 —PCM2004, Fifth Pacific Rim Conference in Multimedia, Tokyo, Japan, November / December 2004 Proceedings, Part II*, ser. Lecture Notes in Computer Science, K. Aizawa, Y. Nakamura, and S. Satoh, Eds., vol. 3332. Tokyo, Japan: Springer-Verlag, June 2004, pp. 489–496.