TEXTUAL DESCRIPTION-BASED VIDEO SUMMARIZATION FOR VIDEO BLOGS

Mayu Otani, Yuta Nakashima, Tomokazu Sato, Naokazu Yokoya

Nara Institute of Science and Technology
{otani.mayu.ob9, n-yuta, tomoka-s, yokoya}@is.naist.jp

ABSTRACT

Recent popularization of camera devices, including action cams and smartphones, enables us to record videos in everyday life and share them through the Internet. Video blog is a recent approach for sharing videos, in which users enjoy expressing themselves in blog posts with attractive videos. Generating such videos, however, requires users to review vast amount of raw videos and edit them appropriately, which keeps users away from doing so. In this paper, we propose a novel video summarization method for helping users to create a video blog post. Unlike typical video summarization methods, the proposed method utilizes the text, which is written for a video blog post, and makes the video summary consistent with the content of the text. For this, we perform video summarization by solving an optimization problem, in which an objective function involves the content similarity between the summarized video and the text. Our user study with 20 participants has demonstrated that our proposed method is suitable to create video blog posts compared with conventional methods for video summarization.

***Index Terms*—** Video blog, video summarization, user study

1. INTRODUCTION

Recently, camera devices, including action cams and smartphones, have spread widely, and many users enjoy recording their everyday experiences in videos. An attractive way to share these videos with others is authoring a video blog post, which is a blog post with videos supporting it. Actually, many active users have uploaded videos in their blog posts [1].

Most of such videos appear to be in the web-TV show style, which can be easily edited from a small number of raw videos by capturing them based on a prepared scenario in a preliminarily structured way, *e.g.*, giving a presentation on something. Unfortunately, it is difficult to apply this style to most videos capturing everyday experiences; most people take videos without preliminarily preparing a scenario for final video to be uploaded and consequently get a cluttered set of videos. This makes video editing, especially for video blogs, much cumbersome. A post in a video blog usually consists of text with a certain story, *e.g.*, telling prominent events during a user's trip, as well as a video that may include some scenes to support the story, as shown in Fig. 1. Therefore, to

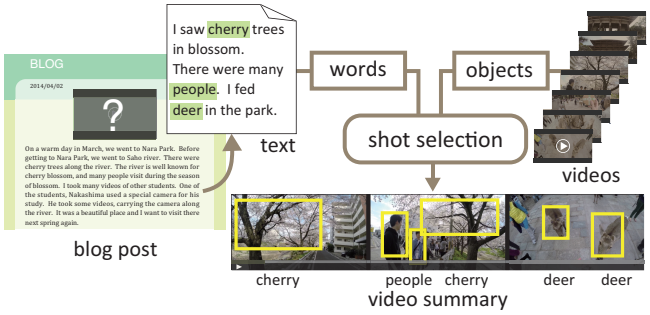


Fig. 1. Our method regards text as a bag of word objects and videos as ones of visual objects.

generate a desirable video, the user needs to review her/his video set and find a subset suitable for what she/he wants to present in the video blog post.

A potential approach to alleviating this video editing process is to adopt video summarization [1]. Video summarization is a technique to produce a compact representation of a vast amount of videos. However, existing methods for video summarization are not suitable for authoring video blog posts. Most of the methods set their goal to generate a video summary that satisfies preliminarily designed criteria such as content coverage or important/interesting events [2]. Although some methods take user's preference into account [3, 4], they do not offer a control over the content in video summaries.

In this paper, we propose a novel method for video summarization that leverages the text in its user's blog post for determining the video subset to be included, considering that the text fully describes the user's story, *i.e.*, what she/he wants to present in the video blog post. Our main contributions are summarized as follows.

- To offer a control over the video summary, we propose a novel textual description-based framework for video summarization that uses text as its basis for video subset selection. This framework is suitable for video blogs because their posts usually consist of text and videos, and thus the users can skip video editing process without any additional burden, since writing text is much easier than video editing.
- To achieve textual description-based video summarization, we propose a new criterion for video shot selec-

tion, which is based on nouns in the text and objects in the original video set as shown in Fig. 1.

- We formulate the problem of video summarization as an optimization problem, which can be efficiently solved by the dynamic programming algorithm.
- Our user study with 20 participants has demonstrated the advantages of our proposed method over conventional approaches for video summarization.

2. RELATED WORK

Video summarization generates a compact representation from a vast amount of videos, which include diverse types of videos such as sports videos [4], news programs [5], consumer videos [6], as well as video retrieval results [7]. Video summarization techniques represent such videos by a set of keyframes [3, 8] or a sequence of shots [9, 10].

Conventional video summarization methods use low-level visual features, such as color and motion [5, 6, 8, 9]. One major approach for video summarization is to reduce redundancy in output summaries. For example, Gong and Liu proposed a cluster-based method in [5], which groups a set of frames in the input videos into several clusters using low-level features and extracts keyframes closest to each cluster's center. Another major approach is to monitor temporal changes of features [8, 9]. Laganière *et al.* [9] extract shots where spatio-temporal features have salient changes.

A video usually contains audio signals, and some methods make use of this additional information for video summarization [10, 11, 12]. Ma *et al.* [11] proposed visual and audio attention models to detect important shots in a video. The method proposed by Taskiran *et al.* [10] uses the speech transcript to achieve a video summary covering maximum semantic content.

The structures of stories and events are critical cues for comprehensible video summaries [13, 14]. Lu and Grauman [14] suggested an observation that some objects leading to another event are important to tell the story in a video.

Some methods utilize prior knowledge on target video sets, *e.g.*, by detecting a predetermined set of objects or events [4, 15, 16]. In the case of videos of certain types of sports, some specific events, such as scoring, are essential for comprehending the entire game, and thus a good video summary should cover them. Babaguchi *et al.* [4] defined significant events in an American football game and developed the method to pick scenes considering a user's preference, such as favorite players or teams.

Some research efforts have been dedicated to take advantage of external information [4, 7, 17, 18]. Sang and Xu [17] proposed a method for movie summarization that uses the script for the movie to retrieve characters and their dialogues. The script aligned with the movie provides shots with important character activities. Another powerful external information is the Internet. To find preferable scenes that are worth to be included in the video summary, Khosla *et al.* [18] trained classifiers to find them from web images.

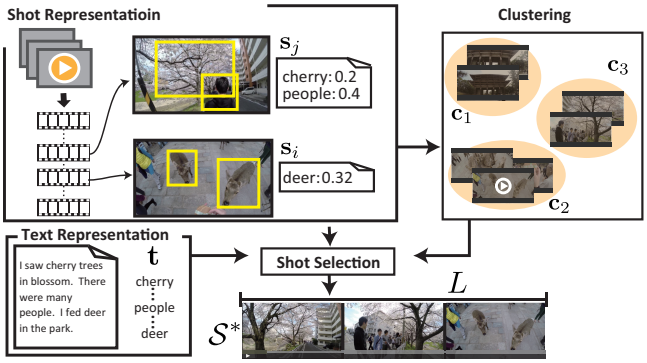


Fig. 2. Overview of our video summarization method.

Our method also utilizes text as external information. Being different from the existing methods, which use text for video content analysis, ours uses the text to determine the content of the output video summary so that the video summary represents the user's story well. This is a novel framework for video summarization, because it provides users with control over the content of the video summary in some extent.

3. TEXTUAL DESCRIPTION-BASED VIDEO SUMMARIZATION

Our method generates a video summary based on the input text that is to be uploaded as a video blog post together with the video summary. Considering this purpose, a video summary must satisfy the following requirements.

1. It must contain scenes relevant to the text.
2. It must include various scenes as long as they are compatible with Requirement 1.
3. It must not be redundant.

Requirement 1 is particularly essential for video blogs, and the others are common to most video summarization methods.

Fig. 2 shows an overview of our proposed method. Our method takes as input a video set V, text T written by the user, and the maximum length L of the video summary. The proposed method first segments each unedited video in V into short shots and annotates each of them with visual objects that appear in the shot, considering they are essential to represent its content. It also extracts nouns from T. The shots are then grouped into several clusters based on the visual object annotation and side information associated with the shots such as timestamp and geo-location. These clusters are expected to loosely correspond to the scenes that the original videos are captured. Using these clusters, annotated visual objects and nouns in the input text, we define an objective function that involves the content similarity between the input text and a set of shots, which is then maximized using the dynamic programming. Our video summary is composed by arranging the extracted shots based on their timestamp.

Our proposed method uses clusters to set preference in order to increase the reliability of the content similarity evaluation: For example, different instances of the same object category may or may not suit with the input text if they are captured at different places and times. Suppose a user inputs text on cherry blossom on a riverside, and there are video shots containing cherry blossom captured on the riverside and at a park. The former shot is relevant to the input text, but the latter is not. Observing this, we set preference to all shots in each cluster based on the content similarity between the cluster and the input text.

3.1. Text Representation

As mentioned above, we presume that objects are essential cues for measuring the content similarity; therefore, in this work, we represent the input text T by a set of nouns, each of which is associated with a certain object o_n ($n = 1, 2, \ldots, N$). For this, the proposed method applies speech tagging to T for extracting nouns and lemmatization to them using [19]. We then remove predefined stop words in them, which discards words that hardly contribute to the representation of T. The input text T is represented as a vector $\mathbf{t} = (t_1, t_2, \ldots, t_N)$, where $t_n = 1$ indicates T contains a noun associated with the object o_n and $t_n = 0$ otherwise.

3.2. Shot Representation

For evaluating the content similarity between the text and a set of shots, which are defined as a short video segment with consistent visual objects, the proposed method represents each shot by a set of visual objects appearing in the shot.

The proposed method first divides each of the original video in V into short shots. Considering the consistency of visual objects, we employ a video segmentation method based on keypoint matching [20]. This method uses the number of matches in successive frames, which roughly indicates if the same objects are included in these frames, as a cue for segmentation. Namely, the method divides the video if the number of matches is a local minimum.

We then annotate each shot with their visual objects with respective bounding boxes. Based on our assumption of the consistency of visual objects in a shot, we extract the middle frame of each shot as its keyframe and annotate it. Techniques for automatic image/video annotation based on general object detection, *e.g.*, a method proposed in [21], can be applied. In our framework, since we have a list of possible objects in the video summary as $\mathbf{t}$, it might be possible to train object detectors for them automatically using the Internet images [22]. In this paper, however, we manually annotate the shots to demonstrate the potential performance of our method with perfect annotation and bounding boxes.

The proposed method then computes a weight value for each visual object, which can be deemed as the importance of the visual object. Various methods can be used for measuring the importance such as in [23]; in this work, we set the weight value solely based on the position and size of the object in the frame for simplicity. More specifically, letting Ω_n be the region surrounded by the bounding box for the object o_n in the frame, we define its weight value s_n as

$$s_n = \int_{\mathbf{x} \in \Omega_n} \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\mathbf{x}, \tag{1}$$

where $\mathcal{N}$ represents the Gaussian with the mean $\boldsymbol{\mu}$ being the frame center position and the predefined variance $\boldsymbol{\Sigma}$. The i-th shot is denoted by a vector of weight values $\mathbf{s}_i = (s_{i,1}, \ldots, s_{i,N})$, and the entire video set by $S = \{\mathbf{s}_i | i = 1, \ldots, I\}$ where I is the number of shots in V.

3.3. Objective Function

The proposed method summarizes videos by finding a subset $S^* \subset S$, which maximizes an objective function involving the content similarity between T and S^* as well as the redundancy of S^*, *i.e.*,

$$f(S, \mathbf{t}) = \mathrm{Sim}(S, \mathbf{t}) + \beta \mathrm{Cvrg}(S, \mathbf{t}). \tag{2}$$

The first term represents the content similarity roughly corresponding to Requirement 1, and the second term the coverage of content related to the scenes described in the text, corresponding to Requirement 2. The redundancy of S^* stated in Requirement 3 is indirectly incorporated into the second term. This objective function can be approximately maximized using the dynamic programing algorithm [24]. The following sections detail our objective function.

3.3.1. Clustering-based Shot Preference

The content similarity solely based on each shot is not reliable; therefore, we introduce the preference for each shot based on the scene, in which it is recorded. We group the shots into several clusters, each of which corresponds to a certain scene, and calculate the similarity between the objects in each clusters and the input text T as the preference.

For clustering the shots, we adopt affinity propagation [25] with the following criterion for $i \neq j$.

$$A(\mathbf{s}_i, \mathbf{s}_j) = \exp\left[-\frac{\lambda \min(|\tau_i - \tau_j|, \theta)}{M}\right] + \gamma J(\mathbf{s}_i, \mathbf{s}_j), \tag{3}$$

where τ_i denotes the temporal frame index of the keyframe for the i-th shot, M the number of frames in S, and $J(\cdot, \cdot)$ the weighted Jaccard similarity defined as

$$J(\mathbf{s}_i, \mathbf{s}_j) = \frac{\sum_n \min(s_{i,n}, s_{j,n})}{\sum_n \max(s_{i,n}, s_{j,n})}. \tag{4}$$

λ, θ, and γ in Eq. (3) are parameters. We set $A(\mathbf{s}_i, \mathbf{s}_i)$ to the median of $A(\mathbf{s}_i, \mathbf{s}_j)$. The k-th cluster is represented by $\mathbf{c}_k = (c_{k,1}, c_{k,2}, \ldots, c_{k,N})$ where $c_{k,n} = 1$ means o_n is included in it (*i.e.*, the cluster contains at least one shot that gives $s_{i,n} > 0$) and $c_{k,n} = 0$ otherwise. The preference p_i for the i-th shot is defined as $p_i = J(\mathcal{C}_i, \mathbf{t})$ where $\mathcal{C}_i$ is the cluster to which the i-th shot belongs. The preference p_i is high when $\mathcal{C}_i$ has similar content as $\mathbf{t}$, which is used for defining $\mathrm{Sim}(S, \mathbf{t})$ and $\mathrm{Cvrg}(S, \mathbf{t})$.

On a warm day in March, we went to Nara Park. Before getting to Nara Park, we went to Saho river. There were cherry trees along the river. The river is well known for cherry blossom, and many people visit during the season of blossom. I took many videos of other students. One of the students, Nakashima used a special camera for his study. He took some videos, carrying the camera along the river. It was a beautiful place and I want to visit there next spring again.

We went to Nara Park. A lot of deer were around the Nandaimon. There were also a few cracker shops, and many tourists enjoyed feeding deer. I bought some crackers and deer immediately gathered around me.

Nandaimon is a famous gate in the Nara Park. I saw a statue of Nandaimon. There were many people.

Fig. 3. Texts used in experiment. From top to bottom: Text 1, Text 2, and Text 3.

3.3.2. Content Similarity

Intuitively, a set of shots S and the input text T should have high content similarity when they share many objects. Based on this observation, we define the content similarity as

$$\mathrm{Sim}(S, \mathbf{t}) = J(\boldsymbol{\phi}(S), \mathbf{t}), \tag{5}$$

where $\boldsymbol{\phi}(S)$ is the sum of all $\mathbf{s}_i$ in S weighted by p_i, *i.e.*,

$$\boldsymbol{\phi}(S) = \sum_{\mathbf{s}_i \in S} p_i \mathbf{s}_i. \tag{6}$$

This term encourages including shots that contain visual objects referred in T. When the shared objects have large weights or the preference of a shot is high, the shot is more likely to be included. It also penalizes shots with visual objects that are already included enough because the content similarity decreases when weight values of the visual objects exceed those of the text.

3.3.3. Content Coverage

For Requirement 2, we need higher coverage of the content related to the scenes in T, which indirectly encourages to reduce the redundancy of the video summary S^* as well by allowing visual objects that are not in T but in the clusters with high similarity to T, considering Requirement 1. By this, we encode our observation that inclusion of visual objects that appear in scenes relevant to T provides a more complete sense of the surrounding situation during video capturing, even if the visual objects are not included in T. Content coverage is thus defined as

$$\mathrm{Cvrg}(S, \mathbf{t}) = J(\boldsymbol{\phi}(S), \boldsymbol{\psi}(\mathbf{t})), \tag{7}$$

where $\boldsymbol{\psi}(\mathbf{t})$ gives the set of visual objects that appear in the clusters relevant to T, *i.e.*, the n-th element of $\boldsymbol{\psi}(\mathbf{t})$ is 1 when object o_i is included in clusters $\{\mathbf{c}_k | J(\mathbf{c}_k, \mathbf{t}) \geq \rho\}$, where ρ is a predetermined threshold.

4. EXPERIMENTAL RESULTS

We evaluated the proposed method by user study with 20 participants to verify the following points.

Table 1. Input text and methods to be evaluated.

	Input	Method
(a)	No text	Uniform sampling.
(b)	No text	Cluster-based.
(c)	Text 1	Ours.
(d)	Text 1	Description-based w/o content coverage.
(e)	Text 1	Description-based w/o content coverage and preference.
(f)	Text 2	Ours.
(g)	Text 2	Description-based w/o content coverage.
(h)	Text 2	Description-based w/o content coverage and preference.
(i)	Text 3	Ours.
(j)	Text 3	Description-based w/o content coverage.
(k)	Text 3	Description-based w/o content coverage and preference.

1. Our objective function is designed to make a video summary similar to the input text as well as to include various objects indirectly related to it. We confirm if this objective function is suitable for video blogs.
2. We verify if the objective function, which encodes the redundancy and coverage criteria, actually works to make the content of video summaries similar to the input text.

In our evaluation, we used a video set containing 42 videos, which are 80 min in total, capturing a short trip in a day. The video set includes various scenes in, *e.g.*, a car, a riverside, a park. Fig. 3 shows the three paragraphs, each of which can be deemed as text for a blog post. We generated video summaries based on each input text with $L = 20$ sec, and each output by the proposed method is shown in Fig. 4. The parameters were empirically set to $\boldsymbol{\Sigma} = \mathrm{diag}(8w, 8h)$, where w and h are the width and the height of the frame, $\beta = 0.25$, $\lambda = 5$, $\theta = 36000$, $\gamma = 0.25$, and $\rho = 0.1$.

To clarify the advantage of the proposed method, we compare it with several baselines (Table 1). (a) is generated by uniform sampling that results in a naive summary including 10 shots sampled at uniform intervals. (b) is a cluster-based summary, which includes exemplar shots of the clusters derived in Sec. 3.3.1. The shots are selected to contain as many visual objects as possible in the video summary. These baselines represent summaries without consideration of user's story. In order to investigate the performance of our textual description-based method with Text1, Text2, and Text3 ((c), (f), and (i)), we also compared it to some variants of ours. In (d), (g), and (j), the second term $\mathrm{Cvrg}(S, \mathbf{t})$ is turned off, and in (e), (h), and (k), the preference in Sec. 3.3.1 was ignored as well. They were generated so that each of them was about 20 sec long.

4.1. Suitability to Video Blog Post

For evaluating the suitability of the our video summarization method to video blog posts, our subjects watched 11 video summaries generated by methods in Table 1. They also reviewed a video blog post as shown in Fig. 5. They were then asked to score each video in terms of how well the video suits with the blog post. The score ranges from 1 to 5, where 1 means that the video definitely does not suit with the blog post, and 5 means that it suits very well. The subjects were

Fig. 4. Keyframes of shots in the video summaries by the proposed method.

Fig. 5. An example video blog post shown to subjects in user study.

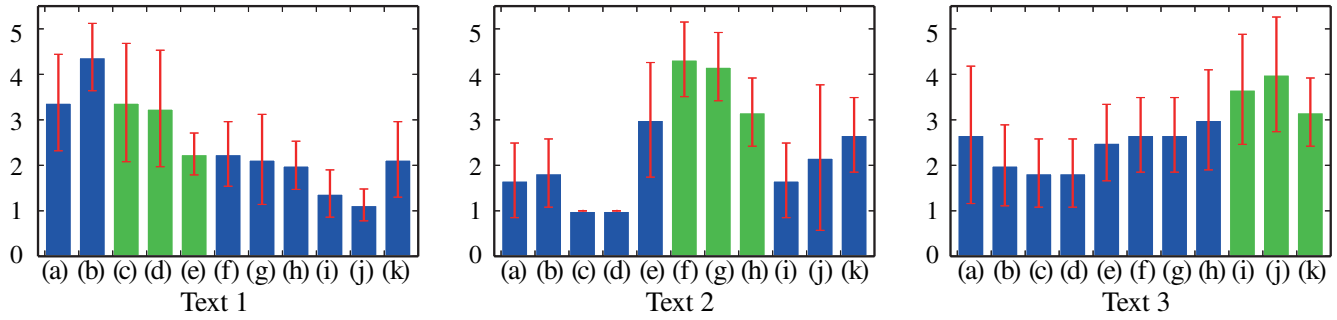


Fig. 6. Average scores on suitability to video blog post with Text 1, Text 2, and Text 3. The methods in Table 1 were evaluated. Video summaries that take the same input text as each video blog post are colored in green.

divided into three groups, and a video blog post with different text in Fig. 3 was assigned to each group to evaluate the subjects' responses to different input text.

Fig. 6 shows the results for Text 1, Text 2, and Text 3, which demonstrate that the proposed method got positive responses for Text 2 and Text 3. For Text 1, the video summary by clustering (b) outperformed ours. This is because the cluster-based video summary (b) includes many shots related to the Text 1 by chance. The summary (b) has another advantage for Text 1. The summary (b) includes shots showing the people heading where the scene in Text 1 was captured. These shots are regarded as leading shots proposed in [14], which smoothly lead to the next scene and improve comprehensibility. Although inclusion of leading shots is not designed in the cluster-based video summary (b), they caused the high score for Text 1. These results for Text 1, Text 2, and Text 3 also show the effects of the content coverage and the preference based on clustering, *i.e.*, content coverage does not affect the score much, but the preference gives significant improvement of suitability. In conclusion, the subjects basically preferred our method for video blogs to other methods, but inclusion of leading shots may improve the suitability.

4.2. Verification of Objective Function

To verify our objective function, we asked the subjects to watch a video containing middle two seconds of all input videos in our video set and to score each method in terms of the following three aspects:

i). How well the video represents the input text (similarity between text and video summary).

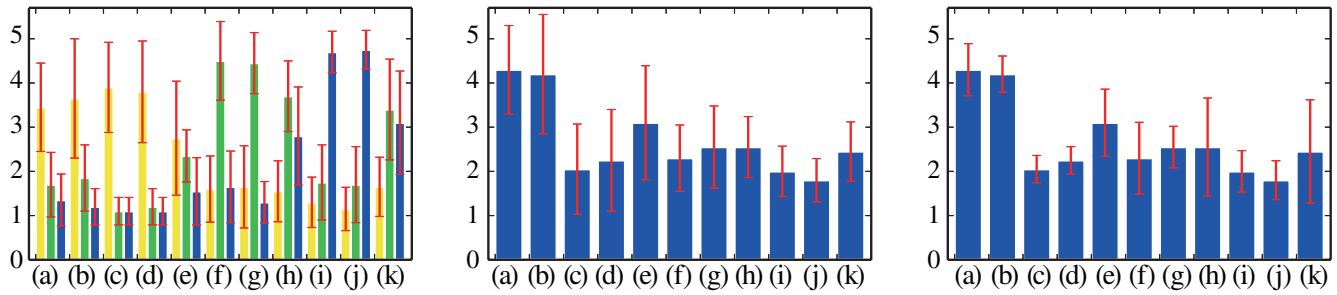
ii). How redundant the video is.

iii). How well the video covers the content of the entire videos.

Figs. 7 (i)–(iii) show the results. As for (i), the result indicates our summaries gain the highest average scores for corresponding input text, which means that our objective function sufficiently works to include the content similar to the input text. In terms of (ii) redundancy and (iii) content coverage, our method got a lower score than others. On the redundancy, although the proposed method tries to reduce the redundancy, it still uses multiple shots from similar scenes because the candidate shots are restricted to ones relevant to the content of the text. On the content coverage, this result is expected because our method does not cover the entire video set but only a subset determined by the input text.

5. CONCLUSION

In this paper, we have proposed a novel method for video summarization that uses textual description to control the content in resulting videos. The proposed method suits to generating video for video blog posts. We have designed an objective function that encodes our observation that a good



(i) Simirality to the text (ii) Redundancy (iii) Coverage of the entire content

Fig. 7. Average scores for questions (i) to (iii). In (i), yellow is for the Text 1, green for the Text 2, and blue for the Text 3.

video summary contains similar content to the text in the blog post. The user study successfully demonstrated that our proposed method is advantageous over the conventional methods in terms of suitability to video blogs. Our future work includes investigation of preferred transition effects for video summarization. Another interesting research direction is to facilitate object detection techniques for automatically annotating original videos with leveraging the input text.

6. REFERENCES

[1] W. Gao, Y. Tian, T. Huang, and Q. Yang, "Vlogging: A survey of videoblogging technology on the web," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 15:1–15:57, 2010.

[2] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput., Commun., and Appl.*, vol. 3, no. 1, art. 3, pp. 1–37, 2007.

[3] A. M. Ferman and A. M. Tekalp, "Two-stage hierarchical video summary extraction to match low-level user browsing preferences," *IEEE Trans. Multimedia*, vol. 5, no. 2, pp. 244–256, 2003.

[4] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted american football video by highlight selection," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 575–586, 2004.

[5] Y. Gong and X. Liu, "Video summarization using singular value decomposition," *in Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 174–180, 2000.

[6] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," *in Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2513 – 2520, 2014.

[7] R. Hong, J. Tang, H.-K. Tan, C.-W. Ngo, S. Yan, and T.-S. Chua, "Beyond search: Event-driven summarization for web videos," *ACM Trans. Multimedia Comput., Commun., and Appl.*, vol. 7, no. 4, pp. 35:1–35:18, 2011.

[8] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," *in Proc. ACM Int. Conf. on Multimedia*, pp. 211–218, 1998.

[9] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu, "Video summarization from spatio-temporal features," *in Proc. ACM TRECVID Video Summarization Workshop*, pp. 144–148, 2008.

[10] C.M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E.J. Delp, "Automated video program summarization using speech transcripts," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 775–791, 2006.

[11] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," *in Proc. ACM Int. Conf. on Multimedia*, pp. 533–542, 2002.

[12] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[13] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 15, no. 2, pp. 296–305, 2005.

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

[15] F. Wang and C.-W. Ngo, "Rushes video summarization by object and event understanding," *in Proc. Int. Workshop on TRECVID Video Summarization*, pp. 25–29, 2007.

[16] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," *in Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1346–1353, 2012.

[17] J. Sang and C. Xu, "Character-based movie summarization," *in Proc. ACM Int. Conf. on Multimedia*, pp. 855–858, 2010.

[18] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using web-image priors," *in Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2698–2705, 2013.

[19] S. Bird, "NLTK: The natural language toolkit," *in Proc. Conf. of Int. Committee on Computational Linguistics and the Association for Computational Linguistics*, pp. 69–72, 2006.

[20] C.-R. Huang, H.-P. Lee, and C.-S. Chen, "Shot change detection via local keypoint matching," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1097–1108, 2008.

[21] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," *in Proc. IEEE Int. Conf. on Computer Vision*, pp. 89–96, 2011.

[22] R Fergus, L Fei-Fei, and P Perona, "Learning object categories from google's image search," *in Proc. Int. Conf. on Computer Vision*, vol. 2, pp. 1816–1823, 2005.

[23] Y. Nakashima and N. Yokoya, "Inferring what the videographer wanted to capture," *in Proc. IEEE Int. Conf. on Image Processing*, pp. 191–195, 2013.

[24] R. McDonald, "A study of global inference algorithms in multi-document summarization," *in Proc. European Conf. on IR Research*, pp. 557–564, 2007.

[25] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.