Multimed Tools Appl (2016) 75:15975–15991
DOI 10.1007/s11042-015-2909-6

A framework for automatic static and dynamic video thumbnail extraction

Jiwon Choi[1] **· Changick Kim**[1]

Received: 25 November 2014 / Revised: 27 May 2015 / Accepted: 24 August 2015 /
Published online: 11 September 2015

Abstract Video thumbnails enable users to see quick snapshots of video collections. To display the video thumbnails, the first frame or a frame selected by using simple low level features in each video clip has been set to the default thumbnail for the sake of computational efficiency and implementation simplicity. However, such methods often fail to represent the gist of the clip. To overcome this limitation, we present a new framework for both static and dynamic video thumbnail extraction. First, we formulate energy functions using the features which incorporate mid-level information to obtain superior thumbnailing. Since it is considered that frames whose layouts are similar to others in the clip are relevant in video thumbnail extraction, scene layouts are also considered in computing overall energy. For dynamic thumbnail generation, a time slot is determined by finding the duration showing the minimum energy. Experimental results show that the proposed method achieves comparable performance on a variety of challenging videos, and the subjective evaluation demonstrates the effectiveness of our method.

Keywords Video thumbnail · Video keyframe extraction · Thumbnail sequence extraction · Content-aware video thumbnailing

✉ Changick Kim
changick@kaist.ac.kr

Jiwon Choi
g1choi@kaist.ac.kr

[1] Department of Electrical Engineering, Korea Advanced Institute of Science and Technology(KAIST), Daejeon 305-701, Korea

1 Introduction

With the development of handheld digital devices, such as digital camera, digital camcorder, or mobile phones, it is universal to capture memorable moments of human life in an effortless manner. Besides, the captured photos and videos can be easily shared via social network services such as Facebook or YouTube. In order to access tremendous amounts of stored visual data, thumbnail images are provided as a preview of the corresponding contents for effective browsing and searching. According to the study in [5], the displayed thumbnails strongly influence the users' behavior in searching and browsing. Figure 1 shows examples that may produce completely different content understanding according to given thumbnails. The first frames selected as thumbnails from various video clips are shown in Fig. 1a, which are still adopted in most of digital devices. However, the first frame may be blurry due to camera motion and often fail to represent the gist of the video clip. On the other hand, the thumbnails shown in Fig. 1b provide better understanding of the video at a glance. Therefore, selection of the semantically representative frame is essential in video thumbnailing.

The primary goal of the video thumbnailling is extracting the most representative frame, which abstracts the content of the video clip. Existing video thumbnailing methods are classified into two main types: *static* and *dynamic* [10]. The static thumbnailing techniques extract a single frame from a video clip to describe the content of the sequence. Conventional methods usually focus on the individual frame quality for keyframe extraction, such as the level of blur, contrast, motion, etc [9, 10, 24]. Recently, more advanced approach has been developed to extract the semantically meaningful keyframe that reflects the theme of the video [8]. But it requires additional information such as video title or tags to obtain representative visual samples from a data. Also empirically thresholds are necessary for candidate keyframe selection. Similarly, video summarization extracts several keyframes to describe the whole content of the video and shows them in the form of a storyboard for effective browsing and searching. Therefore, video summarization is highly related to the studies of video thumbnailing by sharing keyframe extraction techniques. Since the static thumbnail extraction techniques may face the limitation in describing the

Fig. 1 Comparisons of thumbnails and its effect on understanding the video content. **a** first frame selected as a thumbnail, **b** thumbnail extraction results using our method

object movement in a video clip, the dynamic thumbnailing techniques can be considered to generate a video thumbnail by extracting consecutive frames for a few seconds [3, 16]. However, their relatively high computational complexity may limit their use in practice [6, 21].

In this paper, we propose an automatic static and dynamic video thumbnail extraction method which incorporates mid-level information such as location and size of semantic objects as well as low-level information related to scene quality. Therefore, we explore such information to formulate corresponding energy terms. Also, we give preference to the frames whose layouts are similar to other frames when calculating final energy. Finally, the proposed method automatically extracts the representative frame which has a minimum energy cost among all frames.

The contribution of this paper is summarized as follows: (a) We propose an algorithm for static and dynamic video thumbnail extraction. To this end, we formulate the energy terms assessing the mid-level characteristics as well as scene quality. (b) We assume that the frames whose layouts are similar to others are relevant in describing the video. We calculate the proposed scene binary pattern (SBP) descriptor for each frame. Then, we compute the probability of each SBP value, by counting the frequency in the clip. It is used to give preference to those frames in thumbnail extraction.

The remainder of this paper is organized as follows: After Section 2 reviews related work, Section 3 addresses cues to construct energy terms. Section 4 presents the proposed static and dynamic thumbnailing methods. Section 5 discusses experimental results, followed by conclusion in Section 6.

2 Related work

2.1 Video summarization

The basic framework of the video summarization can be briefly described as follows: First, the video sequence is divided into multiple shots by applying shot boundary detection or scene change detection algorithms. For each shot, a single representative frame is extracted as the keyframe. Then, the keyframes are presented in temporal order to build a storyboard.

Earlier work on the video summarization has concentrated on the keyframe extraction by using low-level visual features such as color, texture, shape, and motion [15, 17]. However, the keyframes are selected without regard to its semantic content in these bottom-up approaches. Recently, more advanced approaches have been developed to select keyframes by using semantic analysis [1, 15, 16, 23]. Almeida et al. [1] design a video summarization system for online application, which exploits HSV color histogram directly built in the compressed domain. The system allows user interaction to control the quality of the summaries. Wang et al. [23] present an event driven web video summarization approach based on tag localization and key-shot mining. Ma et al. [15, 16] extract keyframes using a visual attention model for semantic analysis. The employed visual attention model is based on saliency, face, and camera motion. Ngo et al. [18] propose a unified approach for video summarization based on the analysis of video structure and video highlights. Yong et al. [26] present a keyframe extraction method that models semantic context extracted from video frames. To represent the semantic context, low-level features are extracted in blockwise from image segment.

2.2 Video thumbnailing

Unlike video summarization usually displaying several keyframes in the form of a storyboard on a large screen, video thumbnailing aims at displaying a single keyframe or a short video due to limited display space and memory constraint. Lee et al. [12] extract a thumbnail from the H.264/AVC bitstreams in frequency domain directly while considering error propagation. Jiang and Zhang [10, 11] present a spatiotemporal vector quantization method to generate a video thumbnail, where the video time density function and the ICA-based feature extraction method are employed to explore the temporal and the spatial characteristics of video frames, respectively. Another interesting feature is considered in the system, where a frame containing flash illumination is automatically selected as a thumbnail [20]. They take notice that flash illumination is generated at the interesting instant while recording. However, the video with flash illumination is not general in personal video recording. Gao et al. [8] present a video thumbnailing algorithm which reflects the theme of the video content. They notice that the quality-based thumbnail may not be semantically representative. First of all, candidate keyframes are extracted using visual features such as color, motion, face, image quality, and so on. Then, to build the visual theme model, some sample images are obtained by searching visual database using the video tags. The candidate keyframes are compared to the theme model for a semantic ranking, and the highest ranked keyframe is selected as the video thumbnail. Several studies [7, 14, 28] report that there exists an intention gap between the author generated video thumbnail and the user's query. In order to apply the intention of the user to the thumbnail, Liu et al. [14] propose a query sensitive web video thumbnail generation method, which not only consider visual contents, but also meet the preference of the user. Another approach for web video thumbnail to meet the user's preference is presented in [28]. The system recommends thumbnails which satisfy both video owners and browsers on the basis of image quality assessment, image accessibility analysis, video content representativeness analysis and query sensitive matching [28]. Craggs et al. [7] present ThumbReels for query-sensitive web video previews. In order to create a preview that contains a users query, the viewers in crowd-sourced temporally tag videos whilst watching them. Al-Hajri et al. [2] provide a variable-sized thumbnail to represent the popular content using viewing statistics derived from personal or crowd-sourced histories of video consumption for fast navigation of the video. Note that these web video thumbnailing methods [7, 14, 28] require user's query, thus produce query sensitive thumbnail results.

3 Cues to construct energy terms

We focus on the personal home video for video thumbnailing, since it captures real-life events and the usage of unedited videos recorded by consumers is dramatically increasing. We observe that these video contents consist of a single shot, because it is hard to edit while recording using the mobile devices. Therefore, we conclude that it is not necessary to employ scene change detection or shot boundary detection algorithms. In the following subsections, we present seven visual cues to extract representative frame which satisfies both frame quality and the semantic level of the content description. Then we formulate an energy function based on these visual features in each frame for thumbnail extraction.

3.1 Face location (FL) and size (FS)

Face information acts as a primary visual cue, especially in the selection of the representative frame of the video clip. In detail, we obtain the information of face location and size in each frame by using the Viola-Jones face detector [22]. We consider it is more meaningful if a face is located relatively close to the center of the image. Also, we believe that the bigger size of the face is more important than the smaller one in the scene. Thus, we define two energy terms to describe the face information as follows:

$$E_{FL}(i) = \frac{\sqrt{\left(x_{i,c} - W/2\right)^2 + \left(y_{i,c} - H/2\right)^2}}{\sqrt{(W/2)^2 + (H/2)^2}}, \tag{1}$$

$$E_{FS}(i) = 1 - \left(\frac{\sum_{m=1}^{M_i} W_{i,face}^m \times H_{i,face}^m}{W \times H} \right)^2, \tag{2}$$

where M_i is the number of detected faces in the ith frame, and $x_{i,c}$, $y_{i,c}$ represent the center of the largest face among all detected faces in the ith frame. $W_{i,face}^m$, and $H_{i,face}^m$ represent the width and height of the mth face in the ith frame, and W, H denote the width and height of the video. Note that, as the face in the ith frame is located close to the center of the image, $E_{FL}(i)$ gets close to 0. Also, when the face covers the most of the scene in the ith frame, $E_{FS}(i)$ gets close to 0. Note that if faces are not detected in the frame, both $E_{FL}(i)$ and $E_{FS}(i)$ are equal to 1.

3.2 Object location (OL) and size (OS)

In order to deal with other object classes in the image, we adopt the objectness measure proposed in our prior work [4]. In [4], the local regions in the image are categorized into one of three classes: natural, man-made, and object. We obtain the information of object location and size from the classified object region in each frame. Figure 2 shows the classification

(a) (b)

Fig. 2 Examples of detected object region using [4]. Note that the object regions are overlaid in green

results of the object region in the test images. Similar to the face feature mentioned above, we compute the energy terms of the object location and size as follows:

$$E_{OL}(i) = \frac{\sqrt{\left(x_{i,c} - W/2\right)^2 + \left(y_{i,c} - H/2\right)^2}}{\sqrt{(W/2)^2 + (H/2)^2}}, \tag{3}$$

$$E_{OA}(i) = 1 - \left(\frac{area_{i,obj}}{W \times H}\right)^2, \tag{4}$$

where $x_{i,c}$, and $y_{i,c}$ represent the center of the largest object region among all detected object regions in the ith frame, and $area_{i,obj}$ denotes the area of all detected object region in the ith frame.

3.3 Frame difference (FD)

We observe that the object movement becomes a valuable cue for determining the quality and representativeness of the frame. For instance, an object with fast movement is less preferred than the object with focused one with limited movement. While numerous motion estimation algorithms are available in literature, it requires a high computational cost to estimate motion vectors for every frame. Thus, instead of calculating the motion vectors of all sequences, we simply compute frame difference between current and previous frame and normalize to define $E_{FD}(i)$:

$$E_{FD}(i) = \frac{\sum_{x,y} |I(x, y, i) - I(x, y, i-1)|}{W \times H}, \tag{5}$$

where $I(x, y, i)$ represents the normalized pixel intensity at (x, y) in the ith frame. Therefore, if an object is steadily focused, $E_{FD}(i)$ gets close to 0.

3.4 Focus blurness (FB)

Since blurred images are not desirable for keyframe selection, we attempt to reject the blurred images by assigning higher energy. In our energy term, we compute focus blurness as described in [13] to measure the blurness of the image. Based on the assumption that blurred version of the original image loses high frequency components, it is expected to produce little difference between the inherently blurry image and the blurred version of it. Therefore, we define the energy term $E_{FB}(i)$ as follows:

$$E_{FB}(i) = 1 - \frac{\sum_{x,y} |I(x, y, i) - g(x, y) * I(x, y, i)|}{W \times H}, \tag{6}$$

where $g(x, y)$ denotes the Gaussian function and "$*$" denotes the convolution operation. Therefore, blurry images get close to 1, while well-focused images get close to 0.

3.5 Scene steadiness (SS)

In selecting the representative frame of a video clip, it is important to infer the photographer's intention. Without the help of additional user interaction, we pay attention to repeated frames or relatively steady scenes to analyze the representativeness of the video sequence. Assuming that steady scenes contain more meaningful moments and are highly likely to share similar layouts in the video, we measure the frequency of the similar layouts. Inspired

93	110	129
122	117	107
151	127	121

0	0	1
1		0
1	1	1

(a) (b) (c)

Fig. 3 An example of SBP. **a** original image, **b** mean of each 3 × 3 blocks, **c** thresholding of (**b**). The SBP of the sample scene is 47 (00101111_2)

from the LBP feature in [19], we propose the scene binary pattern (SBP) for indexing each frame.

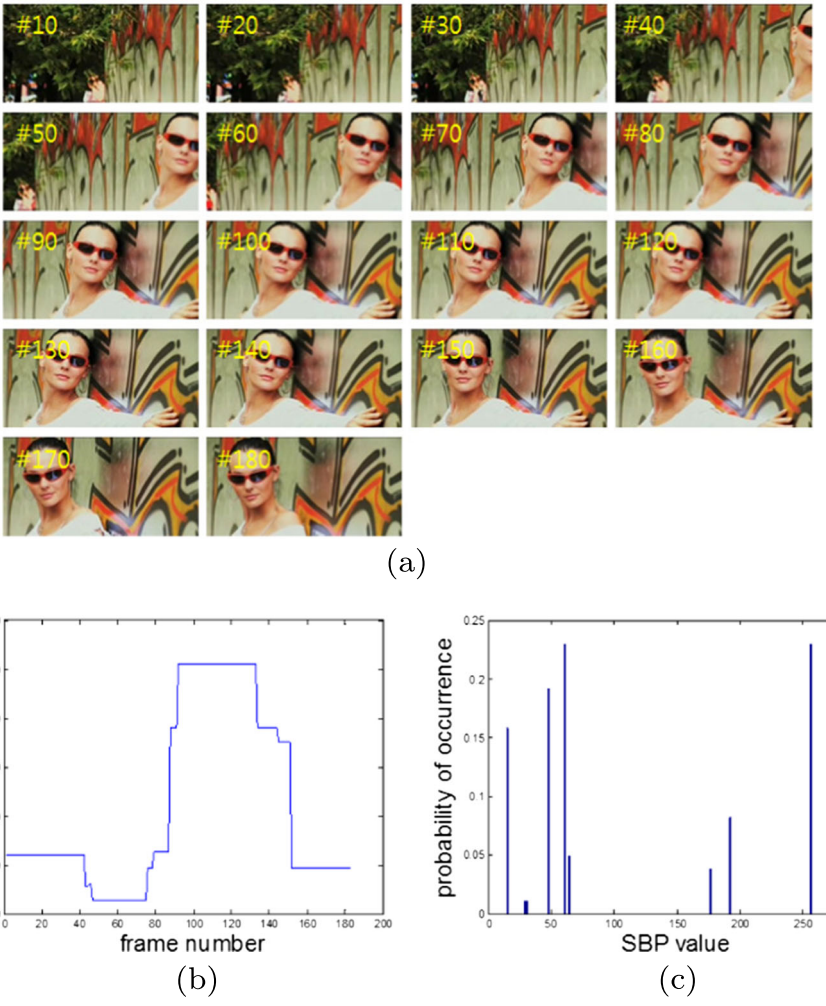

Fig. 4 An example of SBP values and its frequency according to the scene. **a** sub-sampled frame of test video **b** SBP values of each frame, **c** the probability of the occurrence of each SBP value

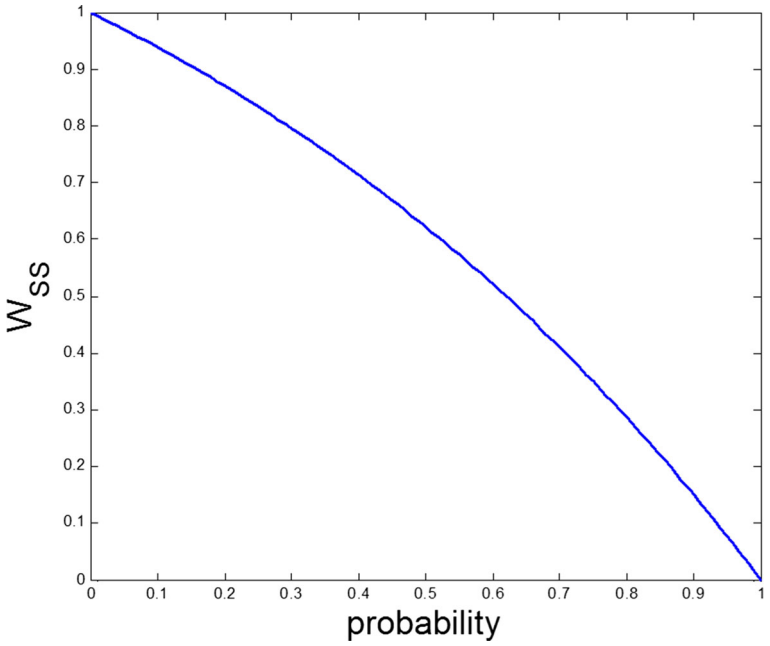

Fig. 5 A plot of W_{SS} according to the probability, $p(SBP)$

To this end, we first divide each frame into 3×3 blocks. In each block, we calculate the average of gray levels, as illustrated in Fig. 3b. Then, we assign a binary value of each block by thresholding the 3×3 neighborhood of each block $g_{i,n}(n = 0, ..., N-1)$ centered at the block $g_{i,c}$ of the ith frame. Therefore, as depicted in Fig. 3c, we calculate the SBP of ith frame as follows:

$$SBP(i) = \sum_{n=0}^{N-1} s\left(g_{i,n} - g_{i,c}\right) 2^n, \tag{7}$$

where

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}. \tag{8}$$

The histogram of the scene steadiness with SBP levels in the range $[0, L-1]$ is a discrete function $h(r_l) = t_l$, where r_l is the lth SBP level and t_l is the number of frames in the sequence having SBP level r_l. Thus, the probability of the occurrence of SBP level is given by $p(r_l) = t_l/T$, for $l = 0, 1, ..., L-1$, and T is the total frame number of the video clip. Figure 4 shows an example of SBP values and its probability of occurrence of the test video sequence. Given the probability of the scene, we assign high weights to scenes with low frequency, while low weights are assigned to steady scenes. We obtain the weight of scene steadiness at the ith frame as follows:

$$W_{SS}(i) = -\frac{\exp(p\,(SBP\,(i))) - 1}{\exp(1) - 1} + 1, \tag{9}$$

where the weight is estimated by inverse modeling. Note that $W_{SO}(i)$ ranges from 0 to 1, and illustrated in Fig. 5.

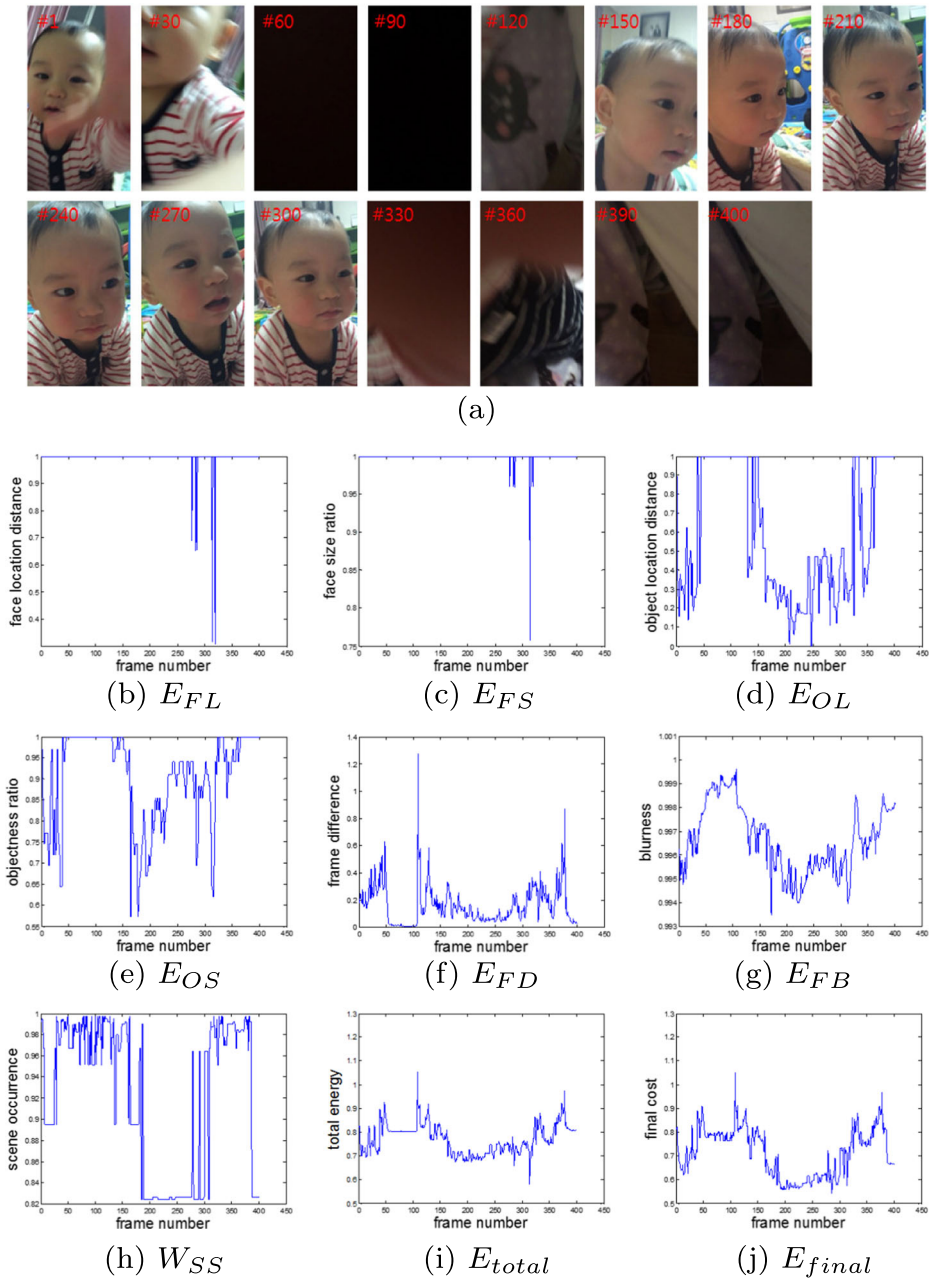

Fig. 6 An example of test video and corresponding energy terms on every frame. **a** sub-sampled frames, **b** face location distance, **c** face size ratio, **d** object location distance, **e** object size ratio, **f** frame difference, **g** focus blurness, **h** scene steadiness, **i** total energy, **j** final final energy by (12)

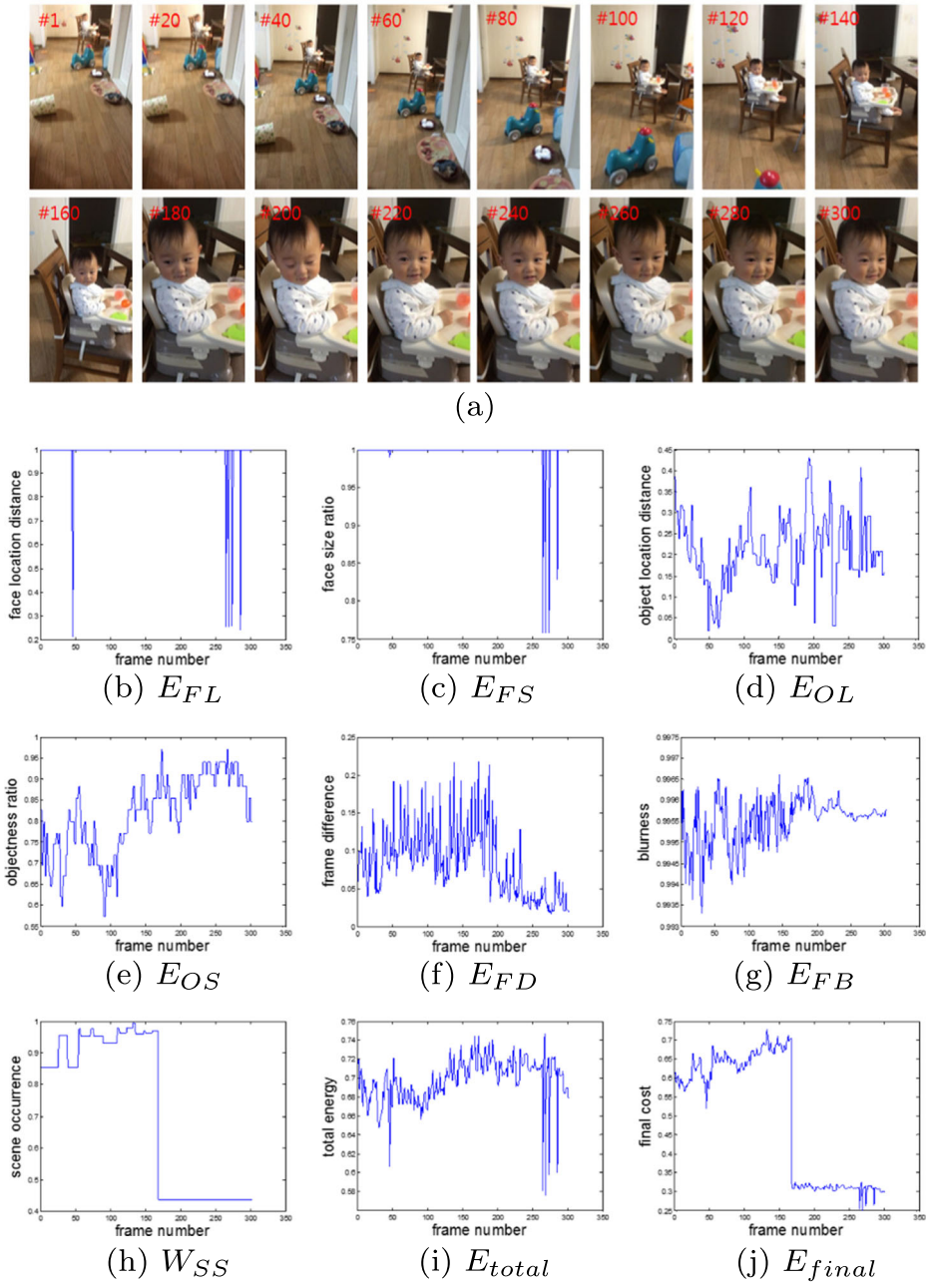

Fig. 7 An example of test video and corresponding energy terms on every frame. **a** sub-sampled frames, **b** face location distance, **c** face size ratio, **d** object location distance, **e** object size ratio, **f** frame difference, **g** focus blurness, **h** scene steadiness, **i** total energy, **j** final energy by (12)

4 Video thumbnail extraction

4.1 Static thumbnail extraction

Next, we formulate a energy function based on the energy terms obtained from several visual cues (see Figs. 6 and 7, b to g) described in the previous section. We express the total energy function in the ith frame as the weighted summation of component energy terms as follows:

$$E_{total}(i) = \lambda_1 E_{FL}(i) + \lambda_2 E_{FS}(i) + \lambda_3 E_{OL}(i) + \lambda_4 E_{OS}(i) + \lambda_5 E_{FD}(i) + \lambda_6 E_{FB}(i), \quad (10)$$

where

$$\sum_{j=1}^{6} \lambda_j = 1. \quad (11)$$

Here, λ_j is weight of each energy term, and the weight parameters are empirically tuned to obtain a satisfying result. We set $\lambda = [0.15, 0.15, 0.15, 0.15, 0.2, 0.2]$ in our experiment.

In order to give preference to the steady scenes in the thumbnail extraction, we apply the weight of scene steadiness W_{SS} to the total energy function as follows:

$$E_{final}(i) = W_{SS}(i) \cdot E_{total}(i) \quad (12)$$

Finally, the proposed method automatically extracts a single frame which has the minimum energy, $\arg\min_i E_{final}(i)$. Figures 6 and 7 describe the effect of the scene steadiness in the final energy.

4.2 Dynamic thumbnail extraction

While static thumbnailing extracts a single representative frame with the minimum energy, dynamic approaches seek consecutive frames that represents the video clip. In our case, we extract consecutive frames that represent the following equation:

$$\arg\min_i \sum_{k=i}^{i+dur-1} E_{final}(k), 0 < i \leq T - dur, \quad (13)$$

where i denotes the ith frame, T is the total frame number of the video clip, and dur, which is set by the user, denotes the number of consecutive frames.

5 Experiments

5.1 Dataset and Details of Experiments

We collected a total of 13 videos from Youtube, [25] and personal collection, with video resolutions ranging from 640×480 to 1920×1080. The collected videos include both indoor and outdoor scenes, which are generally captured by mobile users. Table 1 shows the details of test videos with total frame length. Although many attempts have been made, there is no standard criteria to evaluate the performance of thumbnailing algorithm. Therefore, to evaluate the performance of the algorithm, we performed extensive subjective evaluation of the proposed method on the collection of videos.

We performed a user study on the subjective preference similar to that used in [14, 27, 28]. We asked 20 participants in total for the study. In this study, users were shown original

Table 1 Summary of test videos

Video no.	Total frames	Frame index of the static thumbnail	1st frame's index of the dynamic thumbnail
Video 1	183	92	90
Video 2	192	70	101
Video 3	495	168	175
Video 4	361	90	102
Video 5	1752	1566	1374
Video 6	249	1	1
Video 7	870	348	354
Video 8	839	228	228
Video 9	302	269	211
Video 10	562	272	267
Video 11	401	288	288
Video 12	1338	941	869
Video 13	1346	1094	1117

video and asked to compare our result with the thumbnail taken at the first frame of the video clip, which is usually adopted in various mobile applications. Then, they were asked to give a score: better, the same, or worse, which means if the thumbnails obtained by our algorithm are better than, the same with or worse than the default thumbnails.

Also, we performed subjective evaluation as described in[27]. During the evaluation, the subjects watched the original sequence in advance for evaluation of each thumbnail. Then, the participants were asked to give a score from 1 to 10 for four items: image quality, accessibility, representativeness and the overall evaluation.

Fig. 8 Some comparison results. **a**, **e** thumbnails from 1st frame, **b**, **f** thumbnails obtained by [8], **c**, **g** thumbnails obtained by [26], **d**, **h** thumbnails obtained by our method

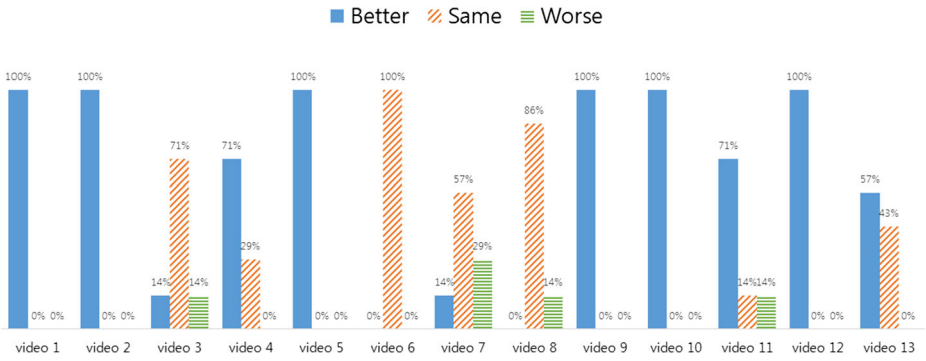

Fig. 9 The subjective evaluations of the user preference task

In addition, we conducted another user study to evaluate the proposed dynamic video thumbnailing. We adopt two items for performance evaluation: informativeness and enjoyability [15, 23]. The participants are asked to give a score (1 to 10) to evaluate the informativeness and the enjoyability, respectively. Note that the higher score indicates the more satisfaction on thumbnails. Each test video has three associated thumbnails, one with 90 (3 sec.) consecutive frames from the original video, and the others with 180 (6 sec.), and 300 (10 sec.) consecutive frames from the original video, respectively. Note that if the length of the original video is less than 300 frames, the original sequence is used in the 300 frame test.

Table 2 Subjective evaluation results of thumbnail image quality and accessibility

Video	Image quality				Accessibility			
No.	Original	Gao's [8]	Yong's [26]	Ours	Original	Gao's [8]	Yong's [26]	Ours
1	5.14	7.67	7.82	8.29	3.57	8.00	7.45	7.57
2	4.86	6.67	7.16	8.00	3.86	7.33	7.28	7.75
3	8.14	5.33	6.02	7.00	6.86	7.00	7.08	7.13
4	5.57	5.67	5.41	6.38	6.29	6.00	6.24	6.88
5	1.86	4.00	7.96	8.13	1.43	6.33	7.88	8.75
6	6.00	6.00	6.02	7.00	7.14	7.33	7.04	7.88
7	6.71	4.33	8.04	7.13	7.29	5.33	7.16	6.63
8	5.29	4.33	6.85	6.75	6.71	5.00	6.92	6.38
9	3.14	6.67	5.98	8.50	5.00	7.33	6.45	8.25
10	5.14	6.00	8.15	6.75	3.43	7.67	8.06	6.88
11	4.71	4.00	8.26	7.50	5.29	5.67	7.62	6.00
12	2.29	4.67	7.48	7.75	2.71	5.00	7.02	7.00
13	2.29	5.67	6.75	7.38	4.43	5.33	6.96	8.88
Average	4.70	5.46	7.07	7.43	4.92	6.41	7.17	7.38

Table 3 Subjective evaluation results of thumbnail representativeness and overall evaluation

Video	Representativeness				Overall evaluation			
No.	Original	Gao's [8]	Yong's [26]	Ours	Original	Gao's [8]	Yong's [26]	Ours
1	3.29	8.00	7.46	7.86	4.57	8.00	7.38	7.57
2	3.71	8.00	7.54	7.75	3.86	7.67	7.48	8.00
3	7.14	6.00	6.74	7.63	7.29	6.00	6.42	7.50
4	5.86	5.00	5.12	6.25	6.14	5.33	5.64	6.63
5	1.86	6.00	8.02	8.75	1.43	5.33	7.96	8.75
6	6.29	5.67	5.82	7.38	6.57	6.33	6..55	7.63
7	6.43	4.67	7.76	6.75	6.57	4.67	7.82	6.50
8	6.71	4.67	6.96	6.50	6.29	4.33	7.04	7.00
9	4.14	8.00	7.04	8.75	3.86	7.33	6.83	8.63
10	4.29	7.00	7.94	6.88	4.57	7.67	8.04	7.25
11	5.29	6.00	7.83	6.75	5.43	5.33	7.95	7.00
12	3.14	4.67	7.64	7.88	2.29	5.00	8.02	8.13
13	4.29	5.00	6.14	8.63	3.29	5.67	6.88	8.75
Average	4.80	6.05	7.08	7.52	4.78	6.05	7.29	7.64

5.2 Results and discussion

Here, we present detailed experimental results to demonstrate the performance of the proposed method. We first show static and dynamic thumbnail extraction results in Table 1, by reporting the extracted frame index which has the minimum cost.

Figure 8 shows qualitative comparisons of static thumbnails. We compare our approach with the thumbnails from the 1st frame, Gao's [8] and Yong's [26]. Note that we used the keyframe selection module for unedited videos in the framework presented in [8].

As shown in Fig. 8, the proposed method qualitatively outperforms the existing methods.

Figure 9 shows the results of the subjective preference task. As reported in Fig. 9, the thumbnails generated by our method are generally better than or comparable to others.

Table 2 and 3 shows the subjective evaluation results of the default thumbnail, Gao's [8] results, Yong's [26] results, and proposed thumbnailing. In subjective preference task, the participants tend to focus on the principal role of the thumbnailing, which extracts the most representative frame without the photographers additional explanation. Therefore, the similar distributions appear in Fig. 9 and the representativeness score of the Table 3. The overall evaluation scores in Table 4 come from considering the image quality, accessibility, and representativeness. In particular, the proposed method gratifies the general requirements in thumbnail, which not only allows for the image quality of the thumbnail, but also satisfies the accessibility and the representativeness of the video content.

The comparisons of dynamic video thumbnailing results with various thumbnail lengths are reported in Table 4. From the results, we have the following observations.

– The subjects consider that the 90-frame sequence thumbnails are more enjoyable than the longer sequence of that, because the 90-frame sequence is considered to be sufficient to understand or estimate the content of the video.

Table 4 Performance evaluation of dynamic thumbnails according to thumbnail duration

	Enjoyability			Informativeness		
Video No.	3 sec. (90)	6 sec. (180)	10 sec. (300)	3 sec. (90)	6 sec. (180)	10 sec. (300)
1	8.57	8.80	8.80	7.42	7.59	7.79
2	8.53	8.80	6.97	7.41	7.90	7.99
3	8.10	6.23	8.43	6.37	6.92	7.09
4	8.41	8.31	8.07	9.10	9.22	9.68
5	9.37	8.90	7.33	8.88	9.22	9.46
6	8.47	8.43	8.80	8.21	8.56	8.89
7	9.33	8.80	7.70	9.33	9.56	9.51
8	9.60	9.53	7.70	8.67	8.67	9.02
9	8.43	8.07	9.90	9.02	9.17	9.46
10	8.89	8.07	6.60	8.56	9.14	9.25
11	8.52	8.07	9.53	7.05	7.62	8.39
12	9.60	9.90	8.07	8.07	8.29	8.56
13	9.08	8.67	6.97	8.73	8.89	9.14
Average	8.84	8.51	8.07	8.22	8.52	8.79

– As expected, there is a trade-off between enjoyability and informativeness. However, we believe that the gaps between average scores of informativeness in Table 4 are acceptable for real world applications.

6 Conclusion

This paper presents an automatic static and dynamic video thumbnailing method through the content-based scene analysis. The proposed method uses features which allows for the image quality and the semantically meaningful representation of the video content. Also, we assume that the steady scenes are more informative, which is calculated based on the SBP. Both static and dynamic thumbnails are automatically extracted, by computing the minimum energy cost. Carefully designed experiments have demonstrated the effectiveness of the proposed method.

References

1. Almeida J, Leite NJ, Torres RdS (2012) Vison: Video summarization for online applications. Pattern Recogn Lett 33(4):397–409
2. Al-Hajri A, Fong M, Miller G, Fels S (2014) Fast forward with your vcr: Visualizing single-video viewing statistics for navigation and sharing. In: Proceedings of the 2014 Graphics Interface Conference, pp 123–128
3. Benini S, Migliorati P, Leonardi R (2007) Hidden markov models for video skim generation. In: Eighth International Workshop on Image Analysis for Multimedia Interactive Services. IEEE, pp 6–6
4. Choi J, Jung C, Lee J, Kim C (2014) Determining the existence of objects in an image and its application to image thumbnailing. IEEE Signal Process Lett 21(8):957–961

5. Christel MG (2006) Evaluation and user studies with respect to video summarization and browsing. In: Electronic Imaging 2006. International Society for Optics and Photonics, pp 60730M–60730M
6. Cotsaces C, Nikolaidis N, Pitas I (2006) Video shot detection and condensed representation: a review. IEEE Signal Process Mag 23(2):28–37
7. Craggs B, Scott MK, Alexander J (2014) ThumbReels: query sensitive web video previews based on temporal, crowdsourced, semantic tagging. In: Proceedings the 32nd annual ACM Conference on Human Factors in Computing Systems, pp 1217–1220
8. Gao Y, Zhang T, Xiao J (2009) Thematic video thumbnail selection. In: Proceedings of IEEE International Conference on Image Processing (ICIP). IEEE, pp 4333–4336
9. Gong Y, Liu X (2000) Generating optimal video summaries. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), vol 3. IEEE, pp 1559–1562
10. Jiang J, Zhang X-P (2010) A novel video thumbnail extraction method using spatiotemporal vector quantization, Proc. of the 3rd International Workshop on Automated Information Extraction in Media Production. ACM, pp. 9–14
11. Jiang J, Zhang X-P (2011) Video thumbnail extraction using video time density function and independent component analysis mixture model. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 1417–1420
12. Lee K-J., Lee W-J., Jeong J-C. (2014) An enhanced error compensation method for thumbnail generation in H. 264/AVC bitstreams. In: Proceedings of the 4th IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp 236–239
13. Li H, Ngan KN (2007) Unsupervized video segmentation with low depth of field. IEEE Trans Circuits Syst Video Technol 17(12):1742–1751
14. Liu C, Huang Q, Jiang S (2011) Query sensitive dynamic web video thumbnail generation. In: Proceedings of IEEE International Conference on Image Processing (ICIP). IEEE, pp 2449–2452
15. Ma Y-F, Hua X-S, Lu L, Zhang H-J (2005) A generic framework of user attention model and its application in video summarization. IEEE Trans Multimed 7(5):907–919
16. Ma Y-F, Lu L, Zhang H-J, Li M (2002) A user attention model for video summarization. In: Proceedings of the tenth ACM international conference on Multimedia. ACM, pp 533–542
17. Money AG, Agius H (2008) Video summarisation: A conceptual framework and survey of the state of the art. J Vis Commun Image Represent 19(2):121–143
18. Ngo C-W, Ma Y-F, Zhang H-J (2005) Video summarization and scene detection by graph modeling. IEEE Trans Circuits Syst Video Technol 15(2):296–305
19. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
20. Sundaram S, Velisavljevic V, Qin Y (2011) Hotflashes: Thumbnailing videos of social gatherings by detecting camera flash illuminated frames. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp 1–4
21. Truong BT, Venkatesh S (2007) Video abstraction: A systematic review and classification. ACM Trans Multimed Comput Commun Appl (TOMCCAP) 3(1):3
22. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
23. Wang M, Hong R, Li G, Zha Z-J, Yan S, Chua T-S (2012) Event driven web video summarization by tag localization and key-shot identification. IEEE Trans Multimed 14(4):975–985
24. Wang T, Mei T, Hua X-S, Liu X-L, Zhou H-Q (2007) Video collage: A novel presentation of video sequence. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp 1479–1482
25. Wang Y.-S, Liu F, Hsu P-S, Lee T-Y (2013) Spatially and temporally optimized video stabilization. IEEE Trans Vis Comput Graph:1
26. Yong S-P., Deng JD, Purvis MK (2013) Wildlife video key-frame extraction based on novelty detection in semantic context. Multimed Tools Appl 62(2):359–376
27. Zhang W, Liu C, Huang Q, Jiang S, Gao W (2012) A novel framework for web video thumbnail generation. In: Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), IEEE, pp 343–346
28. Zhang W, Liu C, Wang Z, Li G, Huang Q, Gao W (2013) Web video thumbnail recommendation with content-aware analysis and query-sensitive matching. Multimed Tools Appl:1–25

Jiwon Choi received the B.S. degree in Electronic Engineering and Computer Science from Kyungpook National University, Daegu, Korea in 2008. She is currently pursuing the Ph.D. degree at the Computational Imaging Lab., Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. Her current research interests are image/video understanding, computer vision, pattern recognition, and image processing.

Changick Kim was born in Seoul, Korea. He received the B.S. degree in Electrical Engineering from Yonsei University, Seoul, the M.S. degree in Electronics and Electrical Engineering from Pohang University of Science and Technology (POSTECH), Pohang, Korea, and the Ph.D. degree in Electrical Engineering from the University of Washington, Seattle, in 1989, 1991, and 2000, respectively. From 2000 to 2005, he was a Senior Member of Technical Staff at Epson Research and Development, Inc., Palo Alto, CA. From 2005 to 2009, he was an Associate Professor in the School of Engineering, Information and Communications University (ICU), Daejeon, Korea. Since March 2009, he has been with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, where he is currently a Professor. His research interests include multimedia signal processing and image/video understanding.