Automatic Generation of Informative Video Thumbnail

Baoquan Zhao*, Hanhui Li*, Ruomei Wang† and Xiaonan Luo‡§
*Nanyang Technological University, Singapore
†Sun Yat-sen University, Guangzhou, China
‡Guilin University of Electronic Technology, Guilin, China
§ Corresponding author: luoxn@guet.edu.cn

2020 8th International Conference on Digital Home (ICDH) | 978-1-7281-9234-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICDH51081.2020.00050

***Abstract*—Thumbnail plays a vital role in boosting the discovery and viewership of online video. Although it can be easily obtained by simply selecting a certain image from the video sequence, most of popular videos in today's video sharing platforms come with elaborately designed custom thumbnails to better showcase the highlight within the video. Unfortunately, both the selection of salient content from thousands of frames and the creation of an eye-catching thumbnail are very time-consuming and require highly specialized skills. In this paper, we present a fully automatic approach for generating informative and eye-catching video thumbnails. The proposed pipeline first splits up a given video into shot units and then selects a set of keyframes using keyframe representativeness and quality assessment metric. After performing salient region detection, a synthetic thumbnail can be obtained by embedding as many as salient regions of keyframes into the carrier image. Experimental results demonstrate the feasibility and effectiveness of our method in informative thumbnail creation.**

***Keywords*-Video summarization; video thumbnail creation; image synthesis;**

I. Introduction

Thumbnails as an important visual medium to showcase content are ubiquitously found in today's video sharing platforms . They, along with video titles and descriptions, often act like advertisement posters to attract the attention of online viewers, spark their curiosity and offer them immediate visual feedback to make relevance judgements [1]. In the early years, video sharing systems like YouTube choose thumbnails from certain marks (e.g., 25%, 50%, 75%) in the video. However, this scheme can easily led to low-quality or less informative images and may put videos at the risk of missing out on a significant proportion of potential audience. Due to the pivotal role thumbnails play in boosting the discovery and viewership, many creators generate elaborately designed thumbnails using professional tools to give a better overview of video content. Unfortunately, it requires highly specialized skills and considerable time and effort to prepare materials, design layout and render, which is a poor choice for amateur users and not suitable for large-scale deployment.

To alleviate burdens on users, growing attention has been paid to the development of effective thumbnail selection and creation mechanisms [2], [3], [4], [5], [6], [7]. Song et al. [5] presented an automatic thumbnail selection system that exploits content relevance and visual aesthetic quality characteristics commonly associated with effective thumbnails. Liu et al. [2] developed an approach to dynamically generate thumbnails for web videos according to use's query. Similarly, Liu et al. [4] proposed a multi-task deep visual-semantic embedding framework for video thumbnail selection. Unlike conventional methods, which usually ignore the side semantic information such as title, description and query associated with the video, they trained an embedding modal using the click-through video and image date to compute the similarity between query and thumbnails. All of these methods mainly focus on single thumbnail selection, which are helpful and efficient to identify catchy thumbnails from tens of thousands of frames in a video. However, video is an information-dense medium and can be a multifaceted combination of activities. For example, a film clip usually contains multiple characters. Simply selecting the 'best' frame as thumbnail using existing schemes cannot guarantee the informativeness and all-sidedness of thumbnail even for a short video, let alone hour long videos. Hence, a question can be raised: can we automatically generate abstractive video thumbnails to reveal more meaningful content of video than a single frame thumbnail? The first thing that comes to mind is image or video summarization [8], [9], [10], [11], [12], which is a topic that has been extensively studied in the past decades. A common practice for image and video summarization is that selecting the most representative visual content and compacting them into a canvas with a particular style. Unfortunately, summarization with a very limited size of canvas has been addressed by these approaches. If we directly apply them to thumbnail generation, it may result in poor readability of content.

In this paper, we introduce a novel framework for automatically generating visually-appealing and informative thumbnails for web videos. In our view, an effective thumbnail should have the following characteristics: representativeness, informativeness and attractiveness. *Representativeness* means that a thumbnail should showcase the most significant content of the video. To achieve this, we perform keyframe clustering and face clustering to identify importance scenes and characters. *Informativeness* requires that a thumbnail should reveal the representative content as much as possible. In our approach, we formulate the visual content compo-

978-1-7281-9234-5/20/$31.00 ©2020 IEEE
DOI 10.1109/ICDH51081.2020.00050

Figure 1. Salient region detection. *Left:* an original keyframe; *Middle:* the detected saliency map of the original keyframe using the method introduced in [13]; *Right:* re-colored saliency map.

sition as an energy optimization problem and obtain an informative thumbnail by embedding salient content into the non-salient regions of carrier image. *Attractiveness* implies that a thumbnail should be visually-pleasing. To reach this goal, we use two different quality assessment models to measure the aesthetic score of frame. To the best of our knowledge, this work is the first attempt to automatically generating abstractive and informative thumbnails by taking the representativeness and visual quality of video content into consideration.

II. Method

The proposed approach consists of the following three main steps: Firstly, we perform shot transition detection to segment a given video into shot units. For each detected shot, frame quality evaluation is sequentially carried out and the one of the best quality is selected as a representative keyframe of the shot. As some shots are quite similar in content, k-means clustering algorithm is further employed to merge shots into shot groups based on the similarities of keyframes. The candidate image set for thumbnail generation can be obtained by selecting one keyframe from each group. In the second step, the salient region of each candidate image is detected. Then the importance of a salient region is measured according to the total shot length in its corresponding group. Finally, a synthetic thumbnail can be generated by embedding as many as salient regions into the non-salient region of carrier image.

A. Candidate Keyframes Selection

Instead of choosing a certain keyframe as content preview thumbnail, the proposed framework starts with extracting a set of most representative and best-quality keyframes from video frame sequence. With a view to improve the efficiency of computation, we first adopt an edge-based shot transition detection algorithm [14], [15], [16], [17] to split up video into basic shot units. To get a stunning thumbnail, the low-quality frames should be avoided when selecting keyframes. And for that, for each detected shot, we extract the best image from it by using two different image quality assessment models according to image content. As human faces are one of the most common and important subjects of videos, we employ a data-driven portrait expression attractiveness ranking model [18] to evaluate the frame quality of human-centric shots. The frame with the highest attractiveness score will be selected as the keyframe of the shot. For the rest of shots, a deep convolutional neural network based model [19], which incorporates joint learning of meaningful photographic attributes and image content information, is used to rank frame aesthetics. Similarly, the frame with the highest aesthetic score is chosen as keyframe of the shot.

So far, we have obtained a set of high quality keyframes of the video. To get a fine-tuning candidate keyframe set for abstractive thumbnail generation, we further perform image clustering algorithm to classify shots into a series of shot groups. There are two major considerations for doing so. The one is, for many videos, there could be tens or hundreds of shots. Without taking content similarities between different keyframes into account, it may easily introduce redundant information into the composite image, which is particularly undesirable especially for thumbnails with strict size constraint. The other is due to the very limited size of the canvas, it is usually not necessary to embedded all candidate keyframes into the carrier image in order to keep the readability of the thumbnail. In other words, only the most representative and significant keyframes are selected for compositing to meet certain constraint conditions. By performing clustering, it will be more accurate to identify salient content of video than using individual shot. As mentioned above, two different types of shots are considered in our framework. Accordingly, we carry out two different clustering strategies. For keyframes of human-centric shots, we extract all facial features from them and then compare the similarity between each pair of faces by calculating facial feature distance. If the distance is large then a predefined threshold, they are grouped into same cluster. For non human-centric shots, k-means clustering algorithm is adopted to classify them into shot groups based on color histogram similarity of their corresponding keyframes. It's worth noting that more sophisticated clustering methods can also be employed to perform this task [20], [21], [22], [23]. Finally, the refined candidate image set can be got by selecting the keyframe with highest quality score in each cluster. And the importance of each cluster can be measured

by the total length of shots in it.

B. Salient Region Detection

Video thumbnails are often presented as low-resolution images in the initial or search result interfaces of video sharing platforms. Although keyframes can be compacted into grid structure layout, as we have seen in traditional video summarization, it could seriously hinder the readability or informativeness. In our method, we adopt a salient visual content composition scheme that selecting one image as the carrier image from the candidate keyframe set and embedding the rest keyframes into the carrier image. This is based on the observation that there are usually less informative regions in a frame and people can still perceive the main idea of image even from a salient patch. In this case, we use a robust salient region detection algorithm [13] to obtain the most significant part of keyframe. As shown in Figure 1, the middle image illustrates the saliency map of a keyframe on its left. The brighter or more intense the color, the higher saliency the region. To create an informative thumbnail, the proposed approach tries to embedded as many as salient content into the non-salient region of the carrier image. For simplicity, we assume that there is one connected region for each keyframe. Intuitively, it can be obtained by detecting the largest contour area after performing adaptive image binarization, for example the Otsu's method [24], on the saliency map. Unfortunately, only the most salient but very small patch is detected using this method, which is usually insufficient to convey recognizable content. Here, we address this by introducing a dual-threshold scheme. Specifically, we use a relative large threshold σ_l to specify the region that disable overlap when embedding as shown by white regions of the right most image in Figure 1. And a small threshold σ_s is also adopted to ensure the integrity and informativeness of salient area , as shown by the gray regions of the right most image in Figure 1. In our experiment, σ_l and σ_s is set to 50 and 5, respectively.

C. Informative Thumbnail Generation

The goal of our thumbnail generation method is to compact as much as representative content of a video into a limited size but visually-pleasing and recognizable image. To achieve this, we select a keyframe from the candidate set as the carrier image of thumbnail and embed as many as salient regions of the rest of keyframes into the non-salient areas of the carrier image. For convenience, we denote the keyframe set for thumbnail generation as $\mathcal{S}$. Let I_i be the carrier image selected from $\mathcal{S}$ and $\hat{\mathcal{S}}$ be the set of the rest of keyframes except I_i. The generation of thumbnail can be formulated as a energy optimization problem. The energy $E(A_i)$ of a generation scheme A_i using I_i as carrier image can be represent as follow:

$$E(A_i) = -(E_{rep}(A_i) + \omega_{rat}E_{rat}(A_i) + \omega_{spa}E_{spa}(A_i)) \tag{1}$$

The first term $E_{rep}(A_i)$ indicates the energy of content representativeness and is defined by:

$$E_{rep}(A_i) = \sum_{I_k \in \mathcal{S}} \alpha_k R(I_k) \tag{2}$$

where $R(I_k)$ is the representativeness score of I_k and measured by the total shot length of its corresponding shot group, and α_k is an indicator variable, taking the value 1 if I_k is selected in scheme A_i and 0 otherwise. The second term $E_{rat}(A_i)$ represents the information loss when scaling salient regions of keyframes in $\hat{\mathcal{S}}$ to fit the canvas:

$$E_{rat}(A_i) = \sum_{I_k \in \hat{\mathcal{S}}} \alpha_k s(I_k) \tag{3}$$

where $s(I_k)$ is a scaling factor that ranges from ℓ_{min} to 1, ℓ_{min} limits the minimum scaling ratio in order to ensure the readability of visual content. In our experiment, ℓ_{min} is set to 0.5. The last term $E_{spa}(A_i)$ indicates the energy of used space using scheme A_i and is measured by:

$$E_{spa}(A_i) = 1 - \frac{\sum_{(x,y)\in I_i^{non'}} 1}{\sum_{(x,y)\in I_i^{non}} 1} \tag{4}$$

where I_i^{non} and $I_i^{non'}$ is the available location set for embedding visual content before and after compositing. The parameter ω_{rat} and ω_{spa} weights the relative importance of the second and last term, respectively. To get a good visual effect, when embedding a irregular salient region into the carrier image, we smooth the overlapped area nearby the boundary along the horizontal and vertical direction by:

$$g(x,y) = (1-\alpha)f_s(x,y) + \alpha f_t(x,y) \tag{5}$$

where $f_s(x,y)$ and $f_t(x,y)$ is the color value of two overlapped areas at position (x,y). By varying α in a linear fashion from 0 to 1, a smooth blending effect can be achieved.

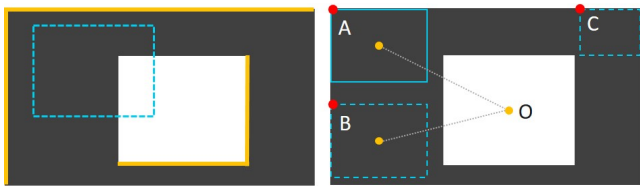

Figure 2. A greedy strategy for energy optimization. *Left:* the salient map binarized with threshold ρ_s of a carrier image and the white region visualizes its bounding box; The dashed rectangle is the bounding box of the salient map binarized with threshold ρ_l of a keyframe in $\hat{\mathcal{S}}$; *Right:* the search process of the scaling ratio and location of the dashed rectangle.

The search space for optimization of $E(A_i)$ is extremely huge. To solve the minimization problem efficiently, the proposed method adopts a greedy heuristic strategy as follows: firstly, a carrier image I_i is selected from $\mathcal{S}$ and its salient region is binarized with the larger threshold ρ_l. We obtain the initial canvas I_{can} for searching solution by setting the

value of all pixels in the bounding box of the binarized image to 255 and the rest to 0, as shown in Figure 2(left). The the keyframes in $\hat{\mathcal{S}}$ are successively selected to check whether it can be embedded into the canvas in accordance with the representativeness score order from high to low. For a candidate salient region bounding box $\hat{I}_{roi}$ of the keyframe in $\hat{\mathcal{S}}$, as shown by the dashed rectangle in in Figure 2(left), we find possible search areas of its upper-left corner. In our approach, a pixel at (x, y) will be tested for embedding if the following conditions are met:

$$\begin{cases} \sum_{p,q\in\{0,1\}} f(x+p, y+p) = 0 \\ \sum_{t\in\{0,1\}} f(x-t, y+t-1) > 0 \\ 0 < x < W-1, 0 < y < H-1 \end{cases} \tag{6}$$

Alternatively,

$$\begin{cases} \sum_{p,q\in\{0,1\}} f(x+p, y+p) = 0 \\ xy = 0, x < W-1, y < H-1 \end{cases} \tag{7}$$

where $f(x, y)$ represents the pixel of the canvas I_{can} at position (x, y), and W and H is the width and height of I_{can}, respectively. In this way, the possible locations of upper-left corner of $\hat{I}_{roi}$ in the canvas can be identified, which are highlighted with yellow lines in Figure 2(left).

Next, we check whether $\hat{I}_{roi}$ can be embedded into the canvas under the following constrains: one is that the the scaling factor $s(\hat{I}_{roi})$ should be in $[\ell_{min}, 1]$; the other one is $\hat{I}_{roi}$ should not overlap with pixels with a value of 255 in I_{can}. To further save computational cost, $s(\hat{I}_{roi})$ is restricted to a finite set of values $\mathcal{R} = \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Initially, we check whether there are feasible solutions from all the possible locations identified in last step without scaling $\hat{I}_{roi}$. If not, we decrease the scaling factor with a fixed step size of 0.1 and then repeat the above procedure. The searching will be terminated if there are feasible solutions or all the elements in $\mathcal{R}$ and $\hat{\mathcal{S}}$ have been checked. This strategy encourages that $\hat{I}_{roi}$ is embedded into the canvas with a large scaling factor. For example, point A will be more favourable for embedding than point C in Figure 2. In most cases, there are more than one solution for a scaling factor. As shown in the right image of Figure 2, both point A and B are feasible solutions. In order to keep spatial continuity of available embedding areas, we select the one of which the center of $\hat{I}_{roi}$ is farthest from that of salient region bounding box of carrier image. Then, the canvas will be updated by setting the value of all the pixels in the occupied region to 1. The generation process using I_i as a carrier image will be terminated if there are no more suitable keyframes in $\hat{\mathcal{S}}$ or available space in I_{can} for embedding. Finally, the optimized solution can be obtained by examining the energy of thumbnails generated by using different carrier images:

$$A^* = argmin_{I_i\in\mathcal{S}} E(A_i) \tag{8}$$

Figure 3 presents the generated thumbnails using the proposed method.

Figure 3. The generated thumbnails using the proposed method.

III. User Study

Although video summarization has been intensively studied, there is no standard criteria for performance evaluations. A common practice [25], [26] for comparison is to perform user studies to obtain a subjective assessment. In order to evaluate the satisfaction of the proposed thumbnail generation method, we carried out the experiment by selecting 12 videos collected from Yahoo Screen [5] in the subject of travelling, cooking, news and film. As this work is the first attempt to generate synthetic thumbnail, we only evalate whether it is more informative than conventional single-frame based thumbnail selection method (YahooThumbnail) [5], and compare the content informativeness, readability and aesthetic with grid structure layout (GSL) and Autocollage [25] using the same keyframes as used in the generation of the proposed method. 12 participants (8 men and 4 women) have been recruited to take part in the survey, in the age range from 21 to 30. All of them major in computer sciences and education levels are above B.Sc. They were asked to rate the following questions with 1 to 5, where 5 means very satisfied and 1 means very dissatisfied. The average scores are:

- Do you think the content of thumbnail is representative of video?
  YahooThumbnail (4.13), Ours (4.25)
- Do you think the thumbnail is informative?
  GSL (2.16),Autocollage (4.02), Ours (4.19)

- Do you consider the content of thumbnail is easy to read?
  GSL (1.52), Autocollage (3.57), Ours (4.14)
- Do you believe the presentation of thumbnail is visually-pleasing?
  GSL (1.87), Autocollage (4.50), Ours (4.36)
- What's your overall satisfaction about the thumbnail?
  GSL (1.35), YahooThumbnail (3.74), Autocollage (4.06), Ours (4.32)

The results demonstrate the effectiveness of the proposed framework. Specifically, the content representativeness of the our method (4.25) than peer method (4.13), which is mainly because we introduced a content-adaptive clustering scheme to precess different types of shot separately. Although the same content is used to generate thumbnail,the informativeness and readability of different presentation styles vary. This is because that the content becomes unrecognisable when compacting several images into a small canvas using GSL (2.16, 1.52) and Autocollage (4.02, 3.57). The thumbnail generated by Autocollage (4.50) outperforms our method (4.36) in terms of aesthetics. We plan to refine it in the further by taking design theory into account. In general, the proposed method achieves a high level of user satisfaction (4.32), which can be used for generating informative and effective video thumbnails.

IV. Conclusion

In this paper, we introduced a fully automatic approach for generating informative and visually-appealing thumbnails. Compared with traditional thumbnail selection methods, which mainly focus on identifying the 'best' image from the video frame sequence as the video thumbnail, the proposed method generates thumbnail by compositing the most representative and high-quality visual content into a carrier image. Such kind of thumbnails reveal video salient content in a more comprehensive manner and could have a positive impact on the discovery and viewership of videos. In the further, we plan to fine-tune the thumbnail by exploring deep learning based content extraction methods such as salient object segmentation and portrait matting.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (61902087) and in part by the Natural Science Foundation of Guangxi (2018GXNSFAA294127).

References

[1] S. J. Cunningham and D. M. Nichols, "How people find videos," in *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2008, pp. 201–210.

[2] C. Liu, Q. Huang, and S. Jiang, "Query sensitive dynamic web video thumbnail generation," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 2449–2452.

[3] W. Zhang, C. Liu, Z. Wang, G. Li, Q. Huang, and W. Gao, "Web video thumbnail recommendation with content-aware analysis and query-sensitive matching," *Multimedia tools and applications*, vol. 73, no. 1, pp. 547–571, 2014.

[4] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3707–3715.

[5] Y. Song, M. Redi, J. Vallmitjana, and A. Jaimes, "To click or not to click: Automatic selection of beautiful thumbnails from videos," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 659–668.

[6] B. Zhao, S. Xu, S. Lin, X. Luo, and L. Duan, "A new visual navigation system for exploring biomedical open educational resource (oer) videos," *Journal of the American Medical Informatics Association*, vol. 23, no. e1, pp. e34–e41, 2016.

[7] B. Zhao, S. Lin, X. Qi, Z. Zhang, X. Luo, and R. Wang, "Automatic generation of visual-textual web video thumbnail," in *SIGGRAPH Asia 2017 Posters*. ACM, 2017, p. 41.

[8] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 3, no. 1, p. 3, 2007.

[9] Z. Yu, L. Lu, Y. Guo, R. Fan, M. Liu, and W. Wang, "Content-aware photo collage using circle packing," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 2, pp. 182–195, 2014.

[10] T. Wang, T. Mei, X.-S. Hua, X.-L. Liu, and H.-Q. Zhou, "Video collage: A novel presentation of video sequence," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 1479–1482.

[11] J. Kim, C. Gray, P. Asente, and J. Collomosse, "Comprehensible video thumbnails," in *Computer Graphics Forum*, vol. 34, no. 2. Wiley Online Library, 2015, pp. 167–177.

[12] Y.-J. Liu, C. Ma, G. Zhao, X. Fu, H. Wang, G. Dai, and L. Xie, "An interactive spiraltape video summarization," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1269–1282, 2016.

[13] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2814–2821.

[14] D. Adjeroh, M. C. Lee, N. Banda, and U. Kandaswamy, "Adaptive edge-oriented shot boundary detection," *Eurasip Journal on Image and Video Processing*, vol. 2009, no. 1, pp. 1–13, 2009.

[15] B. Zhao, S. Lin, X. Luo, S. Xu, and R. Wang, "A novel system for visual navigation of educational videos using multimodal cues," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1680–1688.

[16] B. Zhao, S. Lin, X. Qi, R. Wang, and X. Luo, "A novel approach to automatic detection of presentation slides in educational videos," *Neural Computing and Applications*, vol. 29, no. 5, pp. 1369–1382, 2018.

[17] B. Zhao, S. Xu, S. Lin, R. Wang, and X. Luo, "A new visual interface for searching and navigating slide-based lecture videos," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 928–933.

[18] J.-Y. Zhu, A. Agarwala, A. A. Efros, E. Shechtman, and J. Wang, "Mirror mirror: Crowdsourcing better portraits," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, p. 234, 2014.

[19] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 662–679.

[20] G. Zhang, H. Sun, Y. Zheng, G. Xia, L. Feng, and Q. Sun, "Optimal discriminative projection for sparse representation-based classification via bilevel optimization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1065–1077, 2019.

[21] Y. Zheng, X. Wang, G. Zhang, B. Xiao, F. Xiao, and J. Zhang, "Multi-kernel coupled projections for domain adaptive dictionary learning," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2292–2304, 2019.

[22] G. Zhang, Y. Zheng, and G. Xia, "Domain adaptive collaborative representation based classification," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30 175–30 196, 2019.

[23] G. Zhang, H. Sun, F. Porikli, Y. Liu, and Q. Sun, "Optimal couple projections for domain adaptive sparse representation-based classification," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5922–5935, 2017.

[24] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[25] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "Autocollage," in *ACM transactions on graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 847–852.

[26] T. Mei, B. Yang, S.-Q. Yang, and X.-S. Hua, "Video collage: presenting a video sequence using a single image," *The Visual Computer*, vol. 25, no. 1, pp. 39–51, 2009.