Thumbnail Image Selection for VOD Services

Chun-Ning Tsao[1], Jing-Kai Lou[2], and Homer H. Chen[1]
[1]*National Taiwan University*
[2]*KKStream Limited*
b03902017@ntu.edu.tw, kaelou@kkstream.com, homer@ntu.edu.tw

Abstract

The booming of video on demand (VOD) provides numerous TV series and movies for users to watch anywhere at any time. As the amount of video contents available for viewing grows explosively, thumbnail image representation of video works as a surrogate and facilitates quick and easy video retrieval. Unlike previous work, which considers representativeness as the main criterion for thumbnail image selection, the work described in this paper incorporates attractiveness as an additional criterion. Our idea is based on the observation that thumbnail image for VOD services should not only convey the gist of the video but also intrigue the users. We propose a two-stage method that efficiently utilizes visual features to select a thumbnail image from each TV series. The effectiveness of the proposed method is verified by a subjective test. The results provide further insight into the user preference.

Keywords: Video content analysis, video retrieval, key-frame extraction, image aesthetic assessment, TV series

1. Introduction

In recent years, the booming of video on demand (VOD) has enabled users to select and watch videos on an Internet-enabled appliance anywhere and at any time. With the booming of VOD services, the number of videos to choose from has significantly grown, so is the time required to search for desirable content [1]. Generating an appropriate surrogate, either in the form of text or thumbnail image, for each video is one way to address the problem.

In this work, we focus on the generation of thumbnail image for TV series. Typically, thumbnail images for VOD services can be generated by either editor selection or automatic selection. The former involves intensive manual labor. As there are typically more than one hundred thousand frames in a one-hour video, the editor selection is time-consuming. On the other hand, the latter involves automatic selection of a random or a fixed frame (such as the middle frame). It is quick but the selected frame may be low quality or meaningless (irrelevant to the main plot of the video).

To address the issue, we propose to automatically select thumbnail images on the basis of both representativeness and attractiveness. Our approach is motivated by the observation that VOD users either have a specific target and directly search for it or keep roaming without a specific target until an attractive video appears. Therefore, thumbnail images should not only convey information (such as topic or vibe) of the video to the users, but also have the appeal for most users to attract their attention. These two criteria for thumbnail image selection are the pillars to support efficient media retrieval with engaging viewing experiences.

The employment of representativeness as a criterion for thumbnail image selection involves a measurement of the relevance between a frame and a video from which the frame is extracted. A representative frame should hold a strong connection with the gist of the video. For example, a frame showing the interaction between a couple would be a fine choice for a romantic film. Likewise, for a war film, a frame showing what happens in the battlefield would be a fine choice. Therefore, a representative frame would emphasize the scene from which the video was taken [3]. Normally, directors express such representativeness by showing the scene for a long duration. Moreover, we believe the presence of characters is a prerequisite for a representative frame since characters are the key element for TV series, as reported in a previous user study [2].

The attractiveness criterion involves a measurement of the appeal of a frame. In this work, the attractiveness of a frame is measured by its quality and shot type. The former involves some objective attributes such as sharpness, brightness, contrast, and saturation. The latter is characterized by the distance of the subject to the camera (in other words, the subject's relative size in the frame). As far as the appeal is concerned, our experiences show that medium to close-up camera shots work better than other types of camera shots. Typically, medium shots show a character from about the waist up to the face to emphasize the character and the interaction between characters or

978-1-7281-1198-8/19/$31.00 ©2019 IEEE
DOI 10.1109/MIPR.2019.00018

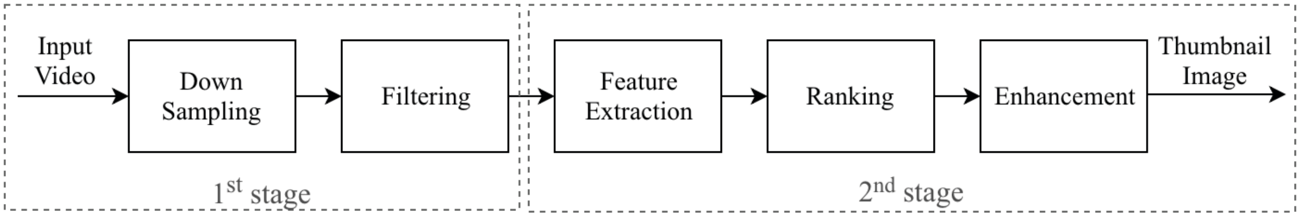


Fig. 1. Flow chart of the proposed method.

between characters and the scene. Close-up shots are taken in a close distance to reveal the details (such as facial emotions and reactions) of the characters. Such camera shots are attractive since they focus on the characters. They are also representative because characters in these camera shots usually play critical roles in the TV series.

In short, we devise a method to automatically select a thumbnail image from each TV series for VOD services. We use representativeness and attractiveness as the two criteria for frame selection and utilize various visual features in a hierarchical way to implement them. Visual features are considered since they can be extracted without additional data such as transcript or audio. We conduct a subjective test to verify the effectiveness of the proposed method and provide further observations of user preference to guide future studies on video abstraction for TV series.

2. Related Work

2.1. Key-frame extraction

Key-frame extraction is similar to thumbnail image selection in many aspects. Traditional key-frame extraction employed the representativeness as the main criterion to extract a set of frames from a video such that the extracted frames represent the salient content of the video. Clustering is a popular approach to key-frame extraction, and various methods have been proposed [3]-[5]. Most methods first segment a video into shots and then extract the key-frames from each shot [6].

However, traditional key-frame extraction focuses on the relevance between a frame and a video without considering the attractiveness of the frame. In our method, we take both representativeness and attractiveness into account. Besides, we select only one frame instead several frames from a video.

2.2. Image aesthetic assessment

In the field of image aesthetic assessment, many computational schemes have been developed to measure the human-perceived aesthetic quality of an image (usually, a photo) [7]. To set the photographic and psychological aesthetics rules, various hand-crafted features have been proposed [8]. Low-level features such as blurriness, lightness, and colors are often used [9]-[11]. Features at the abstraction level, such as scene, composition, and object presence, are also considered [11]-[13]. A comprehensive review of these approaches can be found in the literature [7]. In this work, we further exploit some well-studied hand-crafted features to measure the attractiveness of a frame.

3. Proposed Method

As shown in Fig. 1, the proposed method can be divided into two stages. The first stage includes the down sampling and the filtering steps, and the second stage includes the feature extraction, the ranking, and the enhancement steps. Most of the frames of the input video are filtered out based on certain simple rules in the first stage so that the time-consuming feature extraction in the second stage can be performed on only a small number of frames. The second stage involves a series of operations to select the most representative and attractive frame based on the extracted features. The selected frame is further processed and output as the thumbnail image.

3.1. Down sampling

The down sampling step includes four operations. The first operation is applied to discard the first and last ten percent of the input video, which are often the opening and ending sessions. These two sessions usually contain the theme song, the review, and the preview that are not the main content of the video. In the second operation, the remaining input video is sampled at one frame per second to reduce the number of frames to be analyzed and the redundancy between consecutive frames. The third operation is shot detection, which compares every pair of consecutively sampled frames and checks whether the difference between them is greater than a given threshold. The difference is calculated by

$$D(i,i+1) = \frac{\sum_c \sum_{x=1}^{x=X} \sum_{y=1}^{y=Y} |P_i^c(x,y) - P_{i+1}^c(x,y)|}{XY}, \qquad (1)$$

where i and $i+1$ are the frame indices, X and Y, respectively, denote the width and the height of the frame, $c \in \{H,S,V\}$ denotes the color channel, and $P_i^c(x,y)$ denotes a pixel of frame i in channel c. We choose the HSV color space since it is widely applied for shot detection.

The fourth operation of the down sampling step is applied to sort the shots by length and select the top 100 longest shots. The operation is based on the observation that long shots are usually more stable and hence easier to find sharper frames than short shots. Besides, as discussed in Section 1, long shots usually are more representative than short shots [5].

3.2. Filtering

In this step, frames with poor quality or without presence of characters are removed. A number of features such as sharpness, saturation, brightness, contrast, and embedded subtitle are used to assess the quality of a frame. These features are selected by VOD editors.

The existence of embedded subtitle in a frame is represented by a binary value, same for the existence of characters. On the other hand, sharpness, saturation, brightness, and contrast are represented by numerical values and filtered with a threshold. The detail of the threshold setting is described in Section 3.2.6.

For minimum data representation, at most one frame is extracted from each shot selected in the down sampling step. We start from the middle frame of a shot and check whether all its features pass the corresponding thresholds. If yes, we extract the frame and move to the next shot. If not, we discard the frame and check the next frame until all frames in the shot are processed. If all frames in a shot fail the test, no frame is selected from the shot.

3.2.1. Sharpness. The sharpness of a frame can be obtained by converting the frame to a grayscale image, convolving the grayscale image with a Laplacian filter, and computing the variance of the filtered image [14]. This basic method is also applied to measure the strength of the bokeh effect produced by using optics with shallow depth of field. Typically, this effect is produced to make the background blur and the object of interest sharp. It improves the attractiveness of an image.

More specifically, a frame is first evenly sliced into 5×5 blocks, and the sharpness of each block is computed. Then, the effect of shallow depth of field for the entire frame is measured by

$$(H_1 + H_2) - (L_1 + L_2), \qquad (2)$$

where H_1 and H_2 denote the two highest sharpness values of the blocks, and L_1 and L_2 denote the two lowest ones.

3.2.2. Saturation and Brightness. The computation of saturation and brightness is straightforward. First, a frame is converted to the HSV color space. Then, the saturation and the brightness of the frame are obtained by computing the average pixel value of the S channel and the V channel, respectively.

3.2.3. Contrast. Given the brightness I of each of the 5×5 blocks of a frame, the contrast of the frame is computed by

$$\frac{(I_{max} - I_{min})}{(I_{max} + I_{min})}, \qquad (3)$$

where I_{max} and I_{min}, respectively, denote the maximal and the minimal block brightness. Compared to the original Michelson contrast, which is a pixel-based scheme, the block-based scheme described here alleviates the impact of pixels with extreme values on the contrast measurement.

3.2.4. Embedded subtitles. Production companies may embed subtitle in the source video, but the appearance of subtitle affects the attractiveness of a frame. Therefore, frames with subtitle are discarded.

To detect the existence of embedded subtitle, we compute the sharpness of the bottom center block by the basic method described in Section 3.2.1. Then, we check whether the sharpness value is higher than an empirically determined threshold. This simple approach to subtitle detection works effectively because the subtitle is usually placed in the bottom center of a frame. When a frame is embedded with subtitle, the sharpness of the bottom center block would be high.

3.2.5. Presence of characters. As discussed in Section 1, the presence of characters is a prerequisite for representativeness. Therefore, frames without face are discarded. For the remaining frames, the bounding boxes of the detected faces are passed to the feature extraction step. We use the pre-trained CNN face detector in dlib [15] because it has high accuracy.

3.2.6. Threshold setting. The thresholds for frame removal based on sharpness, saturation, brightness, and contrast are obtained by analyzing the distribution of data annotated by VOD editors. The data consist of 16 videos, each annotated by two editors. We down sample each video in the same way as that described in Section 3.1 and extract the middle frame from each shot selected in the down sampling step. Then, we ask editors to label the frames good or bad according to the frame quality.

To discard blurred, low-saturation, underexposed, or low-contrast frames, we determine the threshold of each feature f as follows:

$$\begin{gathered} T_f = \max V_f \\ s.t.\ P(v_f \geq V_f | B) \geq 2 \times P(v_f \geq V_f | G), \end{gathered} \qquad (4)$$

where T_f denotes the threshold of feature f, V_f denotes the specific feature value to be determined, v_f denotes any value of f, B denotes the bad frames, and G denotes the good frames.

To discard the overexposed frames, we determine another threshold T_b of brightness in a similar way,

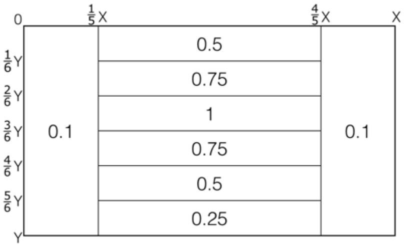


Fig. 2. Face position weights

$$T_b = \min V_b$$
$$s.t.\ P(v_b \le V_b|B) \ge 2 \times P(v_b \le V_b|G), \qquad (5)$$

where subscript b denotes the brightness feature.

3.3. Feature extraction

In this step, three features are extracted to construct the ranking metric. These features are face size score, face position score, and cluster size.

3.3.1. Face size score. We assume that large faces are more important than small faces and define the face size score by

$$S_F(k) = \frac{(s_k - s_{min})}{(s_{max} - s_{min})}, \qquad (6)$$

where k denotes the face index, s_k denotes the size of the kth face in a frame, and s_{max} and s_{min}, respectively, denote the maximal and minimal size of the face in all frames. The “size” of a face means the perimeter of the square bounding box generated by face detection.

3.3.2. Face position score. Professional cinematographers usually place characters’ faces in certain areas of a video frame. The position of a face reflects the importance of a character [16]. Normally, faces in the upper center of a frame attract more attention than those in other areas. The face position score is obtained by

$$S_P(k) = w_{pos}^k, \qquad (7)$$

where w_{pos}^k denotes the position weight of the region where the center of face k locates. The position weights are defined in Fig. 2.

3.3.3. Cluster size. Frames extracted from different shots may be similar if the scene is repeatedly shown in the video. The goal of clustering is to identify the common scene. If a cluster is big, it implies that the corresponding scene appears in many shots. As discussed in Section 1, frames of such shots are potential representative frames. The size of the cluster containing frame i is calculated by

$$S_C(i) = |J|, J = \{j|\ sim(i,j) > T_S, j \ne i\}, \qquad (8)$$

where i and j are frame indices, T_S is a given threshold, and $\mathrm{sim}(i,j)$ denotes the similarity measurement function

Table 1. Threshold values

corresponding feature	step	value(s)
frame difference	shot detection	90
block sharpness	subtitle detection	400
sharpness	filtering	90
brightness	filtering	70, 200
saturation	filtering	35
contrast	filtering	0.3
frame similarity	computing cluster size	0.75

$$sim(i,j) = Inter\big(H(i), H(j)\big) \times C(i,j), \qquad (9)$$

where $Inter(H(i), H(j))$ denotes the intersection function of two histograms, $H(i)$ denotes the HSV histogram of frame i with 16, 3, 3 bins for H, S, and V channels, respectively, and $C(i,j)$ is a constraint function with a binary output to judge whether the interval between two frames is shorter than ten thousand frames. This constraint effectively reduces the false positive rate of the similarity measurement.

3.4. Ranking

In this step, frames are ranked according to the following score metric:

$$S_T(i) = \sqrt{S_C(i)} * \Big(\sum_k (\sqrt{S_F(k)} * P_F(k)\,)\Big), \qquad (10)$$

where $S_T(i)$ denotes the total score of frame i. For each frame i, we sum up the scores of all faces found in the frame. According to the score metric, both close-up shots with a single large face and medium shots with multiple faces receive high scores.

3.5. Enhancement

To further improve the quality of the thumbnail image without generating artifact, we select the sharpest frame from the shot that contains the top frame in the ranking list as the final thumbnail image. The sharpness metric described in Section 3.2.1 is used in this step.

Note that other image enhancement techniques such as image sharpening and contrast enhancement can be applied to this step.

4. Evaluation

4.1. Experimental setup

Table 1 shows all threshold values used in our experiment. Note that since the resolution of frames discussed in Section 3.2.6 is 320 × 180, we resized each frame to this resolution before the filtering step. The CNN face detector in dlib [15] was applied using the default setting. Our software was implemented in Python.

Table 2. Ranking results

TV series	Netflix or Iqiyi	Editor selection	Proposed
SD	2.40±0.73*	1.26±0.48	2.34±0.64
EL	2.13±0.79	1.64±0.76	2.23±0.77
DW	2.09±0.83	1.94±0.77	1.97±0.84
EM	2.11±0.85	1.95±0.83	1.94±0.75
WU	2.10±0.81	1.79±0.77	2.10±0.83
BK	2.30±0.74	1.33±0.59	2.37±0.65

* Mean±standard deviation

Table 3. Percentage of responses in favor of the proposed method over the editor selection

TV series	Percentage	TV series	Percentage
SD	11%	EM	52%
EL	31%	WU	40%
DW	48%	BK	14%

4.2. Questionnaire

Since the assessment of thumbnail image is highly subjective, we conducted a subjective test using an online questionnaire to verify the effectiveness of the proposed method.

The test data consist of six TV series: *Sweet Dreams* (denoted by SD), *Eternal Love* (EL), *Days We Stared at the Sun 2* (DW), *The Ex-Man* (EM), *Wake Up 2* (WU), and *Black Knight* (BK). The first four episodes of each TV series were used in the experiment, so there were 24 videos in total. The length of each video is about one hour, and the frame rate is 29.97 fps.

In the questionnaire, three thumbnail images generated by distinct methods for each video were simultaneously displayed with no indication of which method was used for which image. The three images were generated by the proposed method, the VOD editors, and the commercial software of Netflix or Iqiyi, which are two leading VOD service providers. Although neither company offered the entire six TV series, we were able to obtain DW and WU from Netflix and the others from Iqiyi.

For each video, the subjects were asked to rank the three thumbnail images displayed each time. Then, they were asked to assign a representativeness score and an attractiveness score to each image. The scores range from 1 (poor) to 5 (perfect). To perform the evaluation on the same basis, the subjects were asked whether they have seen each TV series. Moreover, to understand how the subjects assess the thumbnail images, the subjects were asked to describe the rationale behind their ranking.

4.3. Results and discussions

A total of 77 subjects participated in the subjective test. 28 subjects are video experts and the others are ordinary users of VOD services.

Table 4. Representativeness scores and attractiveness scores

TV series		Netflix or Iqiyi	Editor selection	Proposed
SD	R	2.97±1.19	3.74±1.06	2.91±0.98
	A	2.66±1.16	4.22±0.80	3.10±0.99
EL	R	3.07±1.07	3.73±0.89	3.29±0.99
	A	3.16±1.16	3.66±0.97	3.19±1.04
DW	R	3.17±1.05	3.19±1.01	3.43±0.97
	A	3.20±1.07	3.32±1.01	3.31±0.97
EM	R	3.25±1.05	3.84±0.95	3.73±0.91
	A	3.42±1.09	3.69±1.01	3.77±0.96
WU	R	3.15±1.08	3.45±1.07	3.27±1.01
	A	3.12±1.19	3.25±1.07	3.23±1.08
BK	R	3.05±1.05	3.94±0.91	3.09±0.90
	A	3.07±1.00	3.98±0.92	3.13±0.94

R: representativeness, A: attractiveness

The results of ranking are shown in Table 2. For each TV series, the mean and standard deviation are computed for all four episodes. As the values represent the ranking, smaller value is better. It is not surprising that the editors have the best performance. However, our proposed method performs comparably to the editors for DW and EM. Note that the standard deviation for each method and TV series is big in Table 2, which implies there exists significant disagreement between the subjects. To further compare our proposed method with the editors, we compute the percentage of responses in favor of our proposed method over the editor selection. The results are shown in Table 3. A 50% means the two methods are even. In this regard, our proposed method is comparable to the editors for DW and EM and performs closely to the editors for WU. However, the editors perform significantly better than our method for SD and BK. We can see from the representativeness and attractiveness scores in Table 4 that the editors perform quite well for these two TV series.

We can also see from Table 4 that the proposed method outperforms Netflix/Iqiyi for most TV series except SD and is comparable to the editors for DW and EM, which is consistent with the results shown in Table 2. The representativeness score and the attractiveness score are highly correlated (with a correlation coefficient equal to 0.74). This means that a representative thumbnail image is attractive to the subjects as well, and vice versa.

We divide the subjects into two groups: One has seen the TV series and the other has not. The numbers of "seen" subjects are 3, 14, 12, 11, 22, and 8 for the six TV series. We can see from the results shown in Table 5 that there is no significant difference ($p < 0.05$) between these two groups, which implies that prior viewing experience does not affect the performance assessment.

In Table 6, we show the most frequent keywords found in the rationales behind the subjects' evaluation. Four observations are made. First, the composition, which refers to the arrangement of visual elements such as the subject and the background in an image, is indeed a key factor for ranking. Second, the attractiveness outweighs the representativeness in the rationale behind ranking. Third, the

Table 5. Performance Comparison

Question	Subjects	Netflix or Iqiyi	Editor selection	Proposed
Ranking	seen	2.15±0.82	1.68±0.75	2.18±0.77
	unseen	2.20±0.80	1.65±0.76	2.16±0.77
R	seen	3.19±1.11	3.67±1.06	3.38±1.01
	unseen	3.09±1.08	3.64±1.01	3.27±0.99
A	seen	3.22±1.15	3.67±1.00	3.34±1.04
	unseen	3.08±1.13	3.71±1.02	3.28±1.02

R: representativeness, A: attractiveness

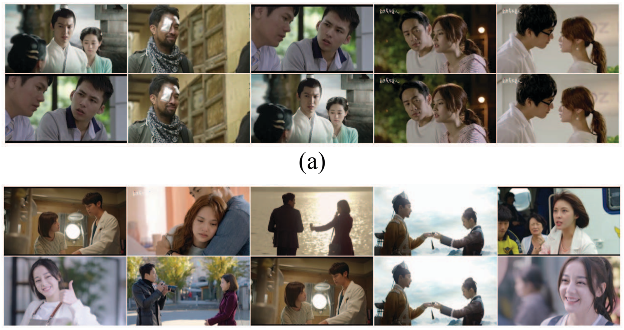
Table 6. The top-5 keywords for each question

Question	Keyword: Frequency
Ranking	composition: 27, attractiveness: 21, facial emotion: 15, actor: 12, character appearance: 11
R	topic conveying: 11, character interaction: 10, relation to the plot: 10, leading characters: 9, crucial scenes: 8
A	composition: 18, character appearance: 15, facial emotion: 13, storytelling: 9, character interaction: 9

R: representativeness, A: attractiveness

interaction between characters is an important element for both representativeness and attractiveness. Finally, leading characters with gorgeous face are ideal candidates for thumbnail image selection.

Thumbnail images with the five highest representativeness scores and attractiveness scores are shown in Fig 3. We can see that the keywords listed in Table 6 indeed describe the characteristics of these images quite well.



(a)

(b)

Fig. 3. Thumbnail images with the five highest representativeness scores (1st row) and attractiveness scores (2nd row). (a) Images generated by the proposed method. (b) Images generated by the editors.

5. Conclusion

We have described a two-stage method that utilizes visual features to automatically select thumbnail images from TV series. It improves viewing experience and facilitates efficient video retrieval. This project was launched in response to the request of a VOD services provider. The subjective test shows the proposed method performs better than the existing solutions of leading VOD service providers. The thumbnail images generated by the proposed method are comparable to those selected by VOD editors. The analysis of the rationales behind subjects' ranking provides further insight of the key elements accounting for user preference.

6. References

[1] Ericsson Consumer Lab, *TV & Media 2017 Study*, 2017, [Online]. Available: https://www.ericsson.com/en/trends-and-insights/consumerlab/consumer-insights/reports/tv-and-media-2017#keyfindings.

[2] D. Wei, G. Marchionini, and D. Soergel, "Multimodal surrogates for video browsing," *Proc. ACM Conf. Digital Libraries*, 1999, p.893.

[3] X. Zeng, W. Hu, W. Li, X. Zhang, and B. Xu, "Key-frame extraction using dominant-set clustering," *Proc. IEEE Int. Conf. Multimedia Expo*, 2008, pp. 1285 -1288.

[4] S.E.F. Avila, A.P.B. Lopes, A. Luz Jr., and A. A. Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[5] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," *Proc. IEEE Int. Conf. Image Processing*, vol. 1, pp. 866-870, 1998.

[6] C. Sujatha and U. Mudenagudi, "A study on keyframe extraction methods for video summary," *Proc. Int. Conf. Computational Intelligence Communication Networks*, 2011, pp. 73-77.

[7] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80-106, 2017.

[8] S. Ma, J. Liu, and C. W. Chen, "A-Lamp: Adaptive layout aware multi-patch deep convolutional neural Network for Photo Aesthetic Assessment," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2017, pp. 722–731.

[9] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," *Proc. European Conf. Computer Vision*, 2006, pp. 288-301.

[10] K. Y. Lo, K. H. Liu, and C. S. Chen, "Assessment of photo aesthetics with efficiency," *Proc. Int. Conf. Pattern Recognition*, 2012, pp. 2186–2189.

[11] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2006, pp. 419-426.

[12] S. Dhar, V. Ordonez, and T. Berg, "High level describable attributes for predicting aesthetics and interestingness," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2011, pp. 1657-1664.

[13] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 271-280.

[14] J. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martinez, and J. Fernández-Valdivia, "Diatom autofocusing in brightfield microscopy: A comparative study," *Proc. Int. Conf. Pattern Recognition*, 2000, pp. 314-317.

[15] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

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