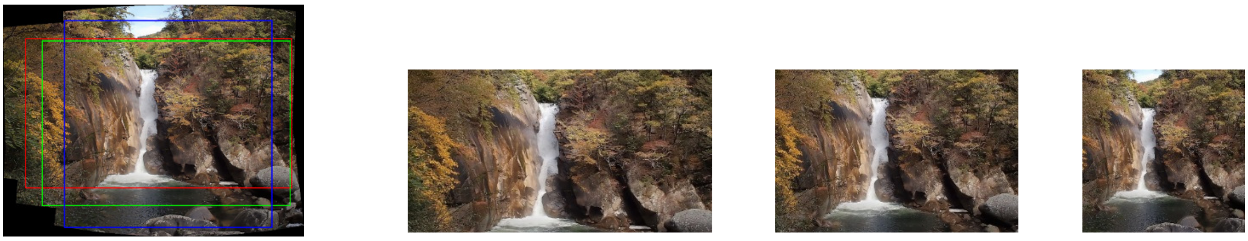
Auto-Framing Based on User Camera Movement

Tomoya SAWADA
Interdisciplinary Graduated School,
University of YAMANASHI

Masahiro TOYOURA
Interdisciplinary Graduated School,
University of YAMANASHI

Xiaoyang MAO
Interdisciplinary Graduated School,
University of YAMANASHI



(a) An image generated by stitching the frames obtained during the user's camera movement.

(b) Result of aspect ratio 16:9.

(c) Result of aspect ratio 3:2.

(d) Result of aspect ratio 1:1

Figure 1: Optimal composition search results for an image at different aspect ratios

ABSTRACT

We propose a novel approach to assisting users with searching the optimal composition of a photograph. In existing studies, the process of detecting the object in a given photo occurred via only image processing, however the result does not always include the object of user's interest. A major technique contribution of our approach is to exploit the user's motion to understand the user's subjective interest in a scene. User's subjective interest and objective structure information of the scene are combined to estimate the best composition based on aesthetic measures. We named this system Auto-Framing. The evaluation result shows that estimated optimal composition closes to the ground-truth. We will embed our technique in an actual camera to enable both automatic detection of compositions and real-time guidance functionality.

CCS CONCEPTS

・Computing methodologies → Computational photography

KEYWORDS

Auto-Framing, optimal composition, user's interest estimation, spatio-temporal analysis, saliency map,

1 INTRODUCTION

Thanks to the advance of computer vision and image processing techniques, anyone can take a photograph of reasonable quality with ease. For example, equipped with the modern face detection technique, many digital cameras provide the face-priority auto focus function, which can set the focus and appropriate exposure automatically. In addition to setting the focus, the composition is yet another important factor to be considered when taking a photo. To the best of our knowledge, currently there are still no effective techniques for helping users to obtain photos with optimal composition, though some methods have been proposed for post-processing a photograph to have the best composition. Understanding how to compose photos generally requires considerable experience and a refined aesthetic taste. For regular users, then, composing photos well can be a challenge. The composition of a photo also depends on the user's intended subject. For this study, we focused on how users move the camera, which is a form of interaction between the user and the target scene when attempting to compose a photo. Using camera movement to estimate the user's subject, the proposed auto-framing technique deduces the subject-image context and automatically captures photos with the optimal composition under an objective aesthetic measure. Users naturally move their cameras around when preparing to take photos. By

CGI '17, June 27-30, 2017, Yokohama, Japan
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5228-4/17/06…$15.00
http://dx.doi.org/10.1145/3095140.3095158

basing our technique around that instinctive process, we establish a system that enables well-composed photography without placing any additional burden on the user. Figure 1 demonstrated the results generated with the proposed method. Figure 1(a) is the master image generated by stitching all the frames obtained during user's camera movement. Figure 1(b) to (d) are the results of optimal composition for given aspect ratios.

2 RELATED RESEARCH

One of the earliest techniques related to the photo composition was the method by Nielsen et al. [1], a system that detected undesirable artifacts in digital photos (lens obstructions by fingers and straps, for example) and automatically cropped the images accordingly. However, the cropping process did not account for photo composition. While Nielsen et al. focused on identifying and eliminating artifacts, our study defines auto-framing as a technique for estimating the user's intended subject, determining a composition that places the subject in the optimal position under some given aesthetic measure, and then automatically cropping the image. While our study aims to develop a tool to assist users taking a photo of ideal composition, several recent studies used data driven approach for post-processing photographs to have optimal compositions ([2]–[4]). Nishiyama et al. [2], for example, designed a technique for learning photo quality and then determining. They prepared a large database of photos with manually provided aesthetic quality scores, created an evaluation function for classifying photo quality via learning processes, trimmed the input image at different scales and sizes, and used the evaluation function to assign a quality score to each sampling image. The system then returned the sampling image with the optimal value as the photo with a good composition. Yan et al. [3] also proposed a method for learning optimized photo cropping position with a good composition using large photo database. The learning process is able to understand a scene of photo and their approach can apply to a new photo by extracting and combining foreground information, intensity difference, texture difference, isolation of foreground and so on. Liu et al. [4] also proposed a method that searched for optimal compositions via an evaluation method combining visual balance, the rule of thirds, and diagonal dominance. The Liu et al. method used saliency map to estimate the main subject of the target image and, based on the saliency results, divided the image into regions to obtain straight-line segments of regional boundaries. To create a distribution map of these important components, the researchers used the three aesthetic guidelines to determine evaluation values for the composition of the image, trim the image at random locations and different scales, and repeat the trimming process until arriving at the optimal composition. Fang et al. [5] also uses saliency map to evaluate a visual appearance. In their approach, to integrate saliency map and edge information with manually weight can find good composition. The saliency map technique allows users to estimate physical locations in video footage and still imagery that attract the viewer's interest. Guo et al. [6] proposed a method for recomposing a photo to align the main subject with the rule of thirds. The approach involved detecting the subject in an image via saliency map and then estimating the optimal position of the subject using a rule of thirds based scale. The researchers found inconspicuous seams at the edges of the target subject and applied several iterations of the Seam Carving process [7] to move the subject to its optimal location.

In these past studies, the process of detecting the subject in a given photo occurred via image processing techniques after the corresponding user took the input photos; the disconnect between the subject-detection step and the actual shooting process made it quite possible that the system's subject-detection "result" would not match the user's actual subjective target. The proposed method operates on a different definition of the subject. Instead of assuming the subject of a photo to be the salient region, our process focuses on estimating the photographer's subjective target subject during the shooting process. Therefore, the proposed method has two key components. First, operating on the assumption that the user's camera movement provides valuable clues for identifying the target subject, the method uses video of camera movement to estimate the user's intended subject—what exactly the photographer is trying to capture in his or her image. Second, the method assesses the aesthetic merit of the subject's position within the complete photo, using the objective heuristic of the rule of thirds.

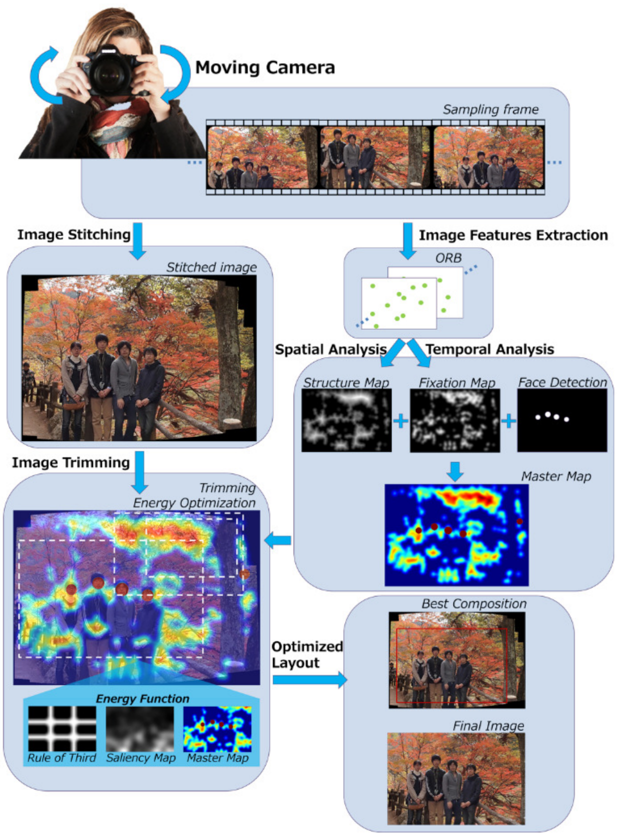


Figure 2: An overview of the proposed method

3 OVERVIEW OF THE PROPOSED METHOD

Composition plays a crucial role in conveying the content and message of a photo, and images that present subjects effectively have a good composition [8]. Figure 2 provides an overview of the proposed method. First, we use the movement of the user's camera as the informational basis for our auto-framing approach. To gather that information, we record video of what the user sees as he or she holds the camera, moves the device around, and composes the shot until eventually releasing the shutter. We then sample the video information at regular intervals to create a set of sampling images. Next, we stitch the sampling images together using feature-matching and projective transformation techniques. This translates the input video into a large synthesis image, which recreates the user's perspective from behind the lens. We also extract ORB (Oriented FAST and Rotated BRIEF) features [9] from each sampling image and perform spatiotemporal analyses on the ORB features to create a Structure Map, which provides a distribution profile of local image features in space, and a Fixation Map, which presents a concentration distribution of local image features in a temporal sequence. To also consider the possible attention on faces, a type of higher-order semantic information, we create a Master Map by integrating an image that has undergone face-recognition [10] processing with the Fixation Map and Structure Map. Based on the Master Map, we then search the stitched composite image to enable trimming at the optimal scale and in the optimal location. When identifying the ideal composition, our system searches for candidates at different scales and different locations in order to find a composition that adheres to the rule of thirds. By also incorporating saliency map [11] to estimate locations in the trimmed image that might draw the viewer's interest, the system searches for locations with objective aesthetic merit. After finding the optimal value, we crop the section out of the stitched composite image and present the result to the user as the best possible composition.

4 ALGORITHM

4.1 Generating the Master Map

One can assume that a photographer's target subject is likely to appear in a vast majority of the corresponding input video frames. In obtaining time-series imagery from the input video, we thus sample the footage at regular intervals. The regular-interval approach not only allows us to cover the full scope of the user's camera movement information but also makes it possible to incorporate local image features in a temporal fashion, thereby enabling estimations of the subject region present throughout the duration of the video.

The photographer's intended subject area appears in most of the frames of the camera-movement video. Drawing on that subject area hypothesis, we create a Master Map that represents the probability of containing the user's subjective target and the distribution of the subject's structural information. The Master Map uses information from two sources: a Fixation Map, which uses a temporal analysis of image features to specify the subject region of the user's intent, and a Structure Map, which uses a spatial analysis of image features to provide structural information on the target subject. Here, the sizes of both the Fixation Map and the Structure Map are the same as the size of the stitched image. Each frame of the video footage contains local image features. To match in-image locations on a frame-to-frame basis and thereby facilitate the image synthesis process, we used ORB features [9]. Using the video of the user's camera movement, we match the descriptors on a frame-by-frame basis and then analyze the descriptors temporally to create a corresponding Fixation Map. We define Fixation Map M^t_{ORB} at frame *t* as follows.

$$M^t_{ORB} = f^t_{ORB} \otimes G\left(0, \sigma_f^2\right), \tag{1}$$

$$f^t_{ORB}(x, y) = \begin{cases} 1 & if\,(x, y)\,is\ ORB\ feature \\ 0 & if\,(x, y)\,is\ not\ ORB\ feature \end{cases}$$

Here, $G\left(0, \sigma_f^2\right)$ corresponds to a Gaussian kernel with mean 0 and variance σ_f^2 and ⊗ indicates a convolution operation. (x, y), meanwhile, represent a pixel of the sampling image. Converting f^t_{ORB} (a local image feature) into a map estimating the probability of information-containing makes it possible to estimate the important location in a given frame. Integrating that process along a temporal path produces final Fixation Map $M_{Fixation}$.

$$M_{Fixation} = \sum_{i=1}^{n} M^i_{ORB} \tag{2}$$

Here, *n* represents the number of sampled frames. As the user's subjective target subject naturally appears in numerous sampling frames, adding together the local image features on a temporal basis makes Fixation Map $M_{Fixation}$ a reliable indicator of the region containing the user's intended subject.

Drawing on the spatial distribution of the local image features, one can also infer structural information characterizing the subject. The proposed method thus detects the target subject (a subjective element) and evaluates the aesthetic merit of the photo's composition (from an objective perspective), giving it the ability to identify ideal scenes reflecting both user intent and compositional beauty. The process of evaluating a photo's composition requires structural information in the distribution of local image features. The rule of thirds holds that overall compositional quality improves when the photographer places the subject of the photo on or close to dividing lines or the intersections of the dividing lines. In order to align important subjects with these dividing lines and thereby achieve a better overall composition, one needs structural information on the subject region, the skeleton of the target subject. The Structure Map, which is for gathering structural information on the region containing the user's intended subject, is generated from the stitched master image in the following 4 steps: 1) Extract ORB keypoints from master image. 2) Create a black image of the same size as the master image. 3) For each ORB keypoint, add a white disk to the image with the radius of the disk proportional to the scale of the ORB feature. 4) Apply distance transform to the binary image to obtain a skeleton representing the structure of feature distribution.

The final Master Map is obtained as the weighted average of Fixation Map $M_{Fixation}$, Structure Map $M_{Structure}$, and the face-recognition results M_{Face}:

$$M'_{Master} = xM_{Fixation} + yM_{Structure} + zM_{Face} \tag{3}$$

Here, x, y and z are user controllable constant and are empirically set to 0.4, 0.3 and 0.3, respectively, in current implementation.

4.2 Automatically determining the optimal composition via objective aesthetic evaluation

4.2.1 The rule of thirds. The rule of thirds is a standard measure for photo composition. The rule is a compositional guideline that divides an image into nine equal parts via "power lines" (two horizontal and two vertical lines). The rule encourages photographers to align their subjects with the power lines or their intersection points ("power points"). According to the rule, placing points of interest on the key line segments and intersections gives the composition a better balance. The rule of thirds also plays into spatial factors. For example, assigning 1/3 of an image's space to the background and the remaining 2/3 to the foreground (or vice versa) divides the total space into thirds and thereby stabilizes the composition. After estimating the photographer's subjective target subject, our proposed method satisfies the general composition rules of photographic science by applying a model that recognizes the importance of aligning subject regions with the line segments and intersections of the rule-of-thirds guideline.

4.2.2 Creating an evaluation function. Give an aspect ratio, to determine the optimal composition locations, we use a random-sampling approach to search the Master Map having the same size as the stitched composite image. By varying the scale with each successive trial and modifying the sampling locations, our system works to determine the optimal composition.

In our method, we define the following energy function to estimate aesthetically pleasing locations in the composite image. First, E_{Line} is an energy function for evaluating how well the given location complies with the central tenet of the rule of thirds: that placing points of interest on the key line segments and intersections gives the composition a better balance.

$$\begin{aligned} E_{Line} &= \gamma \sum_{(x,y)} M'_{Master}(x,y)^{\alpha} \cdot M_{RoT}(x,y)^{1-\alpha} \\ &+ (1-\gamma) \sum_{(x,y)} M_{Saliency}(x,y)^{\beta} \cdot M_{RoT}(x,y)^{1-\beta} \end{aligned} \tag{4}$$

Here, α , β, and γ are parameters for adjusting the weights of the three integrated map, M'_{Master} , $M_{Saliency}$ and M_{RoT} . (x,y) represents a pixel in each map. M'_{Master} is the Master Map trimmed with the given aspect ratio and location. $M_{Saliency}$ is a saliency map computed for the trimmed stitched image. We employ a frequency analysis-based approach (Hou et al. [9]) for the fast computing of saliency map. M_{RoT} is obtained by applying a Gaussian filter to the binary image with "1" represents the dividing lines of one third rule.

$$M_{RoT} = f_{RoT} \otimes G\left(0, \sigma_f^2\right), \tag{5}$$

$$f_{RoT}\left(u,v\right) = \begin{cases} 1 & if\left(u,v\right) is\ on\ LINE\ /\ POINT \\ 0 & if\left(u,v\right) is\ not\ on\ LINE\ /\ POINT \end{cases}$$

Formula (4) first uses map multiplication operations to determine the degree to which the Master Map (reflecting the user's intention) conforms to the rule of thirds. The formula then combines that result with the degree to which the trimmed saliency map meets the rule of thirds standards, providing a comprehensive result. When both the locations with high values in the Master Map and the salient locations after a trimming trial are also prevalent on the M_{RoT}, then, the formula produces a high evaluation value.

Our method also needs to account for the spatial element of the rule of thirds, which holds that assigning 1/3 of an image's space to the background and the remaining 2/3 to the foreground (or vice versa) divides the total space into thirds and stabilizes the composition. We define the final energy function E_{Space} as follows.

$$E_{Space} = \frac{1}{\mathrm{X\ mod\ 3} + 1} \tag{6}$$

Here, mod represents a remainder operator. X represents the number of informative blocks among the 9 blocks divided via power line (maximum number is 9). When the ratio of informative blocks and non-informative blocks is a 1/3 or 2/3 (or vice versa), formula (6) reaches the highest value 1. We use M'_{Master} and $M_{Saliency}$ to estimate the information volume in each divided block and then add together those two maps' values to find the number of blocks exceeding a given threshold. The final cost function is defined as follow:

$$E = E_{Line} \times E_{Space} \tag{7}$$

The final cost is able to take into account the two phases of the rule of the thirds. In the composition evaluation model of the formula (7), a higher value signifies a better composition. In the formula (7), the evaluation function provides an objective assessment of how well the post-trimming information (the Saliency Map) and the user intention-reflecting information (the Master Map) conform to the two main principles of the rule of thirds.

5 EXPERIMENT AND RESULTS

For our experiment, we used a digital single-lens reflex camera (OLYMPUS E-PL2) and a lens (M. Zuiko 17 mm F2.8) with a fixed focal length (equivalent to 34 mm on a 35-mm camera). After having the photographers (nine graduate students, all in their 20s) adjust the focus manually, we started recording video when the user began the shooting process and stopped the recording when the user released the shutter. Each video had a recording resolution of 1280 x 720 and a frame rate of

30 fps. We had the experiment participants shoot freely, taking time to compose their images carefully. Over the course of the experiment, the participants took photos of natural subjects, people, landscapes, night views, food, buildings, and other artificial objects for a total of 34 scenes in normal, everyday situations. In landscape photos (see Figure 1, for example), the aspect ratio of the output has a considerable effect on the overall impression of the image. The results also suggest that subjects in flowing motion (waterfalls, etc.) have little effect on the image synthesis results. During the experiment, the proposed method's process of searching for optimal compositions was not always successful in detecting important locations. Night scenes were problematic, for example. The pervasive darkness in a night scene severely limits the range of feature points to locations where light is present; the system thus has minimal feature input, making it much more difficult to create maps for defining feature importance.

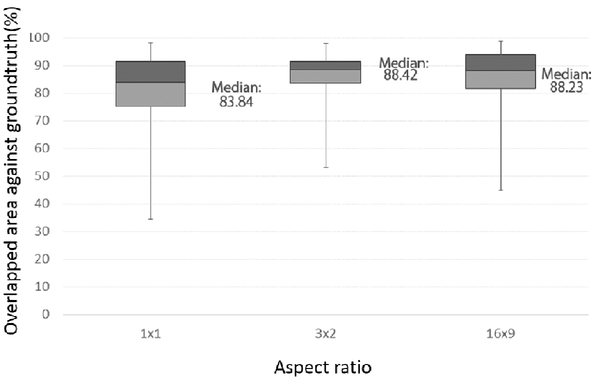


Figure 3: System evaluation result

We also had the photographers manually extract their target subjects from the stitched images at three different aspect ratios. Using that input, we then performed area calculations to determine the accuracy rate of our method by aspect ratio. Figure 3 is a boxplot of the distribution of the 34 scenes across the three aspect ratios. The high median values and small distribution widths in all the aspect ratios suggest that the method delivers high-accuracy results.

6 CONCLUSION

Our study proposed an auto-framing photography technique that 1) detects the target subject based on the photographer's subjective interest and 2) evaluates the composition of the photo from an objective, aesthetic perspective. By basing our technique around the fact that a user naturally moves the camera when composing an image, a process that reflects his or her aesthetic intent, we successfully developed a system that allows the user to take optimally composed photos without having to take any additional, extraneous action. Moving forward, we hope to embed our technology in an actual camera to enable both automatic detection of optimal compositions and real-time guidance functionality.

ACKNOWLEDGEMENTS

This study was funded by JSPS Grants-in-Aid for Scientific Research (No. 16k12459).

REFERENCES

[1] F. Nielsen, S. Owada, and Y. Hasegawa, "Autoframing: A Recommendation System for Detecting Undesirable Elements and Cropping Automatically Photos," in *Proceedings of the international conference on Multimedia and Expo*, pp.417-420, 2006.

[2] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based Photo Cropping," in *Proceedings of the international conference on Multimedia (MM 2009)*, pp.669-672, 2009.

[3] J. Yan, S. Lin, S-B. Kang, X. Tang, "Learning the change for Automatic Image Cropping," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2013)*, pp. 971-978, 2013.

[4] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing Photo Composition," *Computer Graphic Forum (Proceedings of Eurographics 2010)*, vol.29, no.2, pp.469-478, 2010.

[5] C. Fang, Z. Lin, R. Mech, X. Shen, "Automatic Image Cropping using Visual Composition, Boundary Simplicity and Content Preservation Models,"in *Proceedings of the 22nd ACM international conference on Multimedia(MM)*, pp. 1105-1180, 2014.

[6] Y.W. Guo, M. Liu, T.T. Gu, and W.P. Wang, "Improving Photo Composition Elegantly: Considering Image similarity During Composition Optimization," *Computer Graphic Forum (Proceedings of Pacific Graphics 2012)*, vol.31, no.7, pp.2193-2202, 2012.

[7] S. Avidan, and A. Shamir, "Seam Carving for Content-aware Image Resizing," *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2007)*, vol.26, no.3, 2007.

[8] B. Krages, "Photography: The Art of Composition," Allworth Press, 2005.

[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," *IEEE International Conference on Computer Vision (ICCV2011)*, pp. 2564-2571, 2011.

[10] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol.24, no.7, pp.971-987, 2002.

[11] X. Hou, and L. Zhang, "Saliency Detection: A Spectral Residual Approach," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, pp.1-8, 2007.