A Selfie System Adapted to Background Environment using an Omnidirectional Camera

Kanta Nakayama and Norihiko Kawai

Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata, Osaka, Japan

ABSTRACT

The widespread use of smartphones and social networking services has increased the popularity of selfie photos. Many users seek images that are visually appealing not only in terms of facial appearance but also in terms of important background objects and overall composition. However, capturing both the user's face and major background landmarks within a single smartphone photograph is often difficult. This paper proposes a method that automatically generates well-composed selfie images using an omnidirectional camera. The proposed system detects the user's face and identifies important background objects through face detection and semantic segmentation, then selects the optimal composition from multiple perspective-projection candidate images. We evaluated the proposed method using images captured at several tourist spots and compared the results with subjective evaluations.

Keywords: selfie system, omnidirectional camera, face recognition, semantic segmentation

1. INTRODUCTION

The widespread use of smartphones and social networking services has increased opportunities for selfie photography. Many users prioritize not only their face but also the balance between themselves and the background environment. Especially at tourist spots, it is important to include popular landmarks within a photograph. However, when using smartphone cameras, achieving appropriate framing and composition depends heavily on the user's arm length and shooting posture, making it difficult to capture all desired objects within the frame.

Previous studies have proposed methods for evaluating and improving photographic composition. Nishiyama et al.[1] and Tang et al.[2] proposed methods for objectively evaluating photo quality by learning aesthetic principles. Some research extends beyond evaluation to provide real-time photography assistance. Li et al.[3] proposed a system that guides users in positioning subjects according to composition rules in real time. Other studies focus on improving composition through post-capture processing. For example, Li et al.[4] proposed an automatic cropping method to determine optimal framing. However, all these approaches assume the use of conventional cameras and therefore cannot fundamentally solve the problem that important background objects may not fit within the frame at the moment of capture.

Several studies have also explored the use of omnidirectional cameras and 360-degree video to address such limitations. Pano2Vid[5] automatically selects appropriate viewpoints from 360-degree videos, while Deep360Pilot[6] optimizes video viewpoints through subject tracking. However, these methods focus on enhancing 360-degree video viewing experiences rather than selfie photography. In the context of selfies, Kawai et al.[7,8] proposed a method that selects a frame with optimal facial expressions from short 360-degree video segments and generates a perspective-projection image. However, their purpose is to ensure that all subjects fit within the frame, rather than to optimize composition that accounts for both faces and important background objects.

In this research, we address the limited viewing angles and compositional constraints encountered during selfie photography at tourist spots. We propose a selfie system that uses an omnidirectional camera to generate images in which the user's face and the majority of important background objects fit within the frame with well-balanced composition. The proposed system detects the user's face, generates multiple perspective-projection candidates with different compositions, and then applies semantic segmentation to identify important background objects. Based on the extracted object regions, the system selects the optimal perspective-projection image.

Further author information:
Kanta Nakayama: E-mail: m1m24a34@oit.ac.jp
Norihiko Kawai: E-mail: norihiko.kawai@oit.ac.jp

2. PROPOSED METHOD

2.1 Outline

The proposed method consists of five steps. First, the user takes an image at a tourist spot using an omnidirectional camera (i). Next, the photographer's face position is estimated using face detection (ii). Based on the detected facial position, multiple perspective-projection images are generated using different composition styles and aspect ratios (iii). Semantic segmentation is then applied to extract important background object regions (iv). Finally, the system selects the image in which these background objects occupy the largest area (v).

2.2 Omnidirectional Image Acquisition at Tourist Sites

In step (i), the user captures an image using an omnidirectional camera, which records the entire surrounding environment in a single shot. This allows important background objects to be included in subsequent composition generation without being limited by camera orientation. During capture, the user holds the camera extended forward at shoulder height. This positioning helps prevent face-detection failures caused by image distortion. Additionally, the user positions themselves so that important background objects are located behind them, ensuring that these objects appear in the image during subsequent compositional analysis.

2.3 Estimation of Photographer and Facial Position

In step (ii), faces are detected in the omnidirectional image using MTCNN (Multi-task Cascaded Convolutional Networks).[9] To avoid false detections of surrounding people, the method assumes that the photographer's face corresponds to the largest detected bounding box in the image.

2.4 Generation of Rule-of-Thirds and Centered Composition Images

In step (iii), the system generates ten different perspective-projection images based on the facial position obtained in step (ii). We employ two common composition styles: the rule-of-thirds composition and centered composition. Assuming that most users will view images on smartphones, we create both portrait (9:16) and landscape (16:9) aspect-ratio versions. As shown in Fig. 1, five different projection planes are set for both vertical and horizontal orientations. The user's face is positioned either at an intersection point of the rule-of-thirds grid or at the center for centered composition. Perspective projection is then applied to each plane, generating ten compositional candidates in total.

2.5 Extraction of Major Background Object Regions

In step (iv), semantic segmentation is applied to each generated image using Mask2Former[10] to extract major background object regions that determine the visual appeal of a selfie. The perspective-projection images generated in step (iii) include many elements irrelevant to compositional evaluation, such as sky, ground, vegetation, and other people. Therefore, these non-target classes are first identified and removed by a logical negation operation, leaving only the important background object regions.

To obtain stable and consistent regions for area measurement, simple morphological processing is applied as a post-processing step. Specifically, small noise regions are removed and fragmented areas are smoothed using opening and closing operations. This refinement improves the robustness of subsequent compositional evaluation.

(a) Rule-of-Thirds composition (b) Centered composition

Figure 1: Examples of composition styles used in this study.

2.6 Comparison of Extracted Region Size

In step (v), we count the number of pixels corresponding to important background objects in each image obtained in step (iv). Assuming that images in which important tourist attractions appear larger generally have greater visual appeal, the system selects the image with the largest background-object area as the final output presented to the user. By automatically selecting the composition containing the most prominent background elements, this method enables objective compositional evaluation without relying on subjective judgment.

3. EXPERIMENTS

3.1 Overview

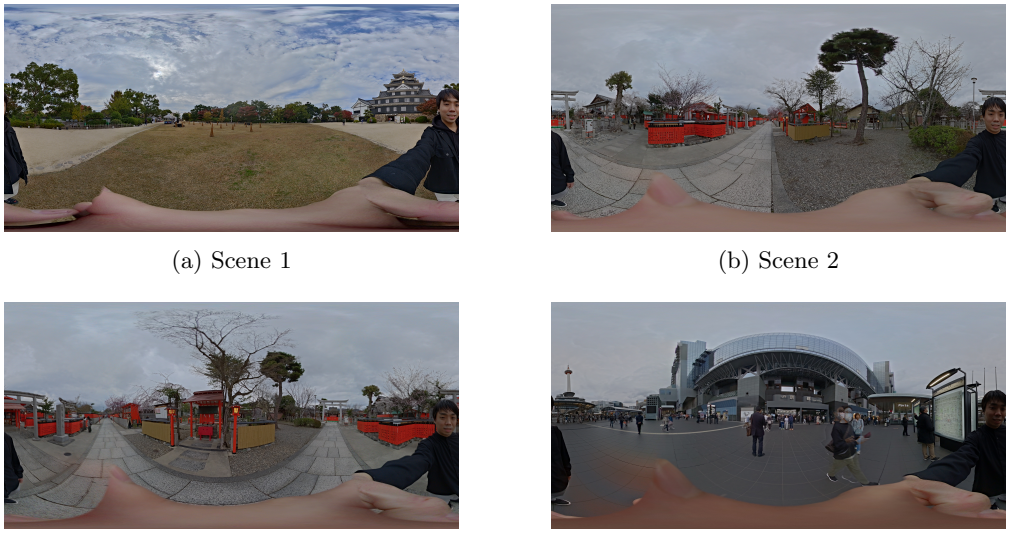
We evaluated the proposed method using the omnidirectional images shown in Fig. 2 as input. During the generation of the perspective-projection images, the viewing angle was fixed at 80 degrees (vertical direction for portrait orientation and horizontal direction for landscape orientation). Ten compositional candidate images were generated for each scene, as shown in Fig. 3. The ten images for each scene were ranked through subjective evaluation, and Fig. 4 presents the images that received the highest ratings for each scene. The effectiveness of the proposed method can be assessed by comparing the images selected by the system with those preferred in the subjective evaluation. For Scenes 1 to 4, we used images captured at four tourist spots: Okayama Castle, Kurumazaki Shrine, Geino Shrine, and Kyoto Tower. The results for each scene are presented below.

3.2 Experimental Results

In Scene 1, the system successfully identified important background objects (Fig. 5a), and the image selected by the system (Fig. 5b) matched the image judged most visually appealing in the subjective evaluation (Fig. 4a). However, since the entire castle structure did not fit completely within the frame, we concluded that adjusting the viewing angle based on the building size is necessary.

In Scene 2, the system correctly detected important background objects (Fig. 6a), and the image selected by the system (Fig. 6b) matched the result of the subjective evaluation (Fig. 4b).

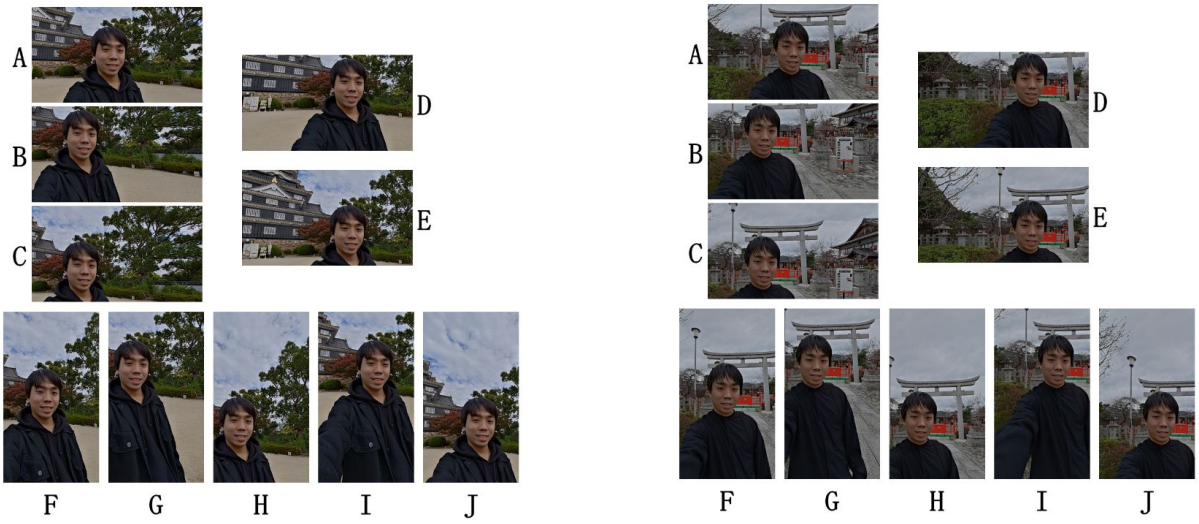
In Scene 3, we assumed that the torii gate and main hall would constitute the important background objects. However, as shown in Fig. 7a, the system also detected the *tamagaki* fence as an important object. Additionally, part of the sky remained in the segmentation results. Consequently, the system selected the image ranked third in the subjective evaluation (Fig. 7b). Fig. 7c shows the mask corresponding to the image judged most visually appealing. In Fig. 7c, the torii gate and main hall occupy a larger area compared with Fig. 7a, indicating



(a) Scene 1 (b) Scene 2

(c) Scene 3 (d) Scene 4

Figure 2: Input images for each scene.



(a) Scene 1

(b) Scene 2

(c) Scene 3

(d) Scene 4

Figure 3: Ten generated perspective-projection images.

that important background objects are more strongly emphasized. This demonstrates that errors caused by the detection of unintended objects and remaining sky regions led to a mismatch with the image judged most visually appealing in the subjective evaluation (Fig. 4c).

Similarly, in Scene 4, a nearby sign was incorrectly detected as an important object, resulting in a larger detected area than Kyoto Tower, which should have been the primary focus (Fig. 8a). As a result, the system selected an image that was ranked third in the subjective evaluation (Fig. 8b). Fig. 8c shows the mask for the highest-rated image, in which Kyoto Tower is clearly detected. These comparisons demonstrate that the detection of unintended objects inflated the apparent area of background objects, causing the system to select a different image from the one preferred in the subjective evaluation (Fig. 4d).

3.3 Discussion

We discuss both the strengths and limitations of the proposed method. Regarding effectiveness, using the background-object area as an evaluation index enabled consistent selection of compositions in which tourist


(a) Scene 1


(b) Scene 2


(c) Scene 3


(d) Scene 4

Figure 4: Images that received the highest ratings.


(a) Extracted object areas


(b) Selected image

Figure 5: Experimental Results of Scene 1.


(a) Extracted object areas

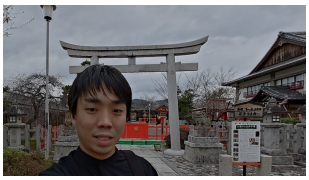

(b) Selected image

Figure 6: Experimental Results of Scene 2.


(a) Selected object areas


(b) Selected image

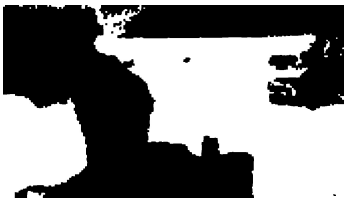

(c) Important background object areas of the image with the highest ratings

Figure 7: Experimental Results of Scene 3.


(a) Extracted object areas


(b) Selected image


(c) Important background object areas of the image with the highest ratings

Figure 8: Experimental Results of Scene 4.

landmarks such as castles and gates appear prominently. Particularly in Scenes 1 and 2, where the images contain fewer elements and the primary background object is clearly defined, the images selected by the system matched those preferred in the subjective evaluation. This suggests that the proposed method performed effectively when the important background object is clearly defined and visual noise is minimal.

However, in Scenes 3 and 4, where images contain many structures in addition to the intended important objects, or where objects have slender and complex shapes, segmentation accuracy decreased, making area measurement unreliable. As a result, even when a composition appeared visually appealing in the subjective evaluation, the system sometimes underestimated the background-object area and selected an incorrect image. Since the current system employs a fixed viewing angle, it cannot flexibly adjust the angle according to the size or distance of the background object, which constitutes another limitation. Moreover, when objects belonging to excluded categories (sky, ground, vegetation, people) themselves become the important objects—such as the Dr. Clark Statue in Hokkaido or *Yakusugi* cedar trees in Kagoshima—the method struggles to detect them accurately, resulting in unstable compositional evaluation.

4. CONCLUSION

This study proposed a selfie system that automatically generates a well-balanced compositional image from a single omnidirectional image. Our experiments demonstrated that the method can correctly identify important background objects in certain scenes and select compositions that align with subjective aesthetic preferences. However, we also identified several limitations: the detection of unintended objects and the use of object area as the evaluation metric can result in ambiguous judgments. Future work includes improving the segmentation process to suppress unintended detections and reconsidering the criteria for identifying important background objects.

REFERENCES

[1] Nishiyama, M., Okabe, T., Sato, I., and Sato, Y., "Aesthetic quality classification of photographs based on color harmony," in [*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 33–40 (2011).

[2] Tang, X., Luo, W., and Wang, X., "Content-based photo quality assessment," *IEEE Transactions on Multimedia* **15**(8), 1930–1943 (2013).

[3] Li, Q. and Vogel, D., "Guided selfies using models of portrait aesthetics," in [*Proceedings of Conference on Designing Interactive Systems*], 179–190 (2017).

[4] Li, D., Wu, H., Zhang, J., and Huang, K., "A2-rl: Aesthetics aware reinforcement learning for image cropping," in [*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 8193–8201 (2018).

[5] Su, Y.-C., Jayaraman, D., and Grauman, K., "Pano2vid: Automatic cinematography for watching 360 videos," in [*Proceedings of Asian Conference on Computer Vision (ACCV)*], (2016).

[6] Hu, H., Lin, Y., Liu, M., Cheng, H., Chang, Y., and Sun, M., "Deep 360 pilot: Learning a deep agent for piloting through 360° sports video," in [*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2017).

[7] Kawai, N., Kiuchi, K., and Imamura, S., "Selfie system using an omnidirectional camera considering facial expressions," *Journal of the Institute of Image Electronics Engineers of Japan* **54**(1), 138–146 (2025). (in Japanese).

[8] Kiuchi, K., Imamura, S., and Kawai, N., "Selfie taking with facial expression recognition using omni - directional camera," in [*Proceedings of International Workshop on Frontiers of Computer Vision (IW-FCV)*], (2024).

[9] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y., "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters* **23**(10), 1499–1503 (2016).

[10] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girshick, R., "Masked-attention mask transformer for universal image segmentation," in [*Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 1280–1289 (2022).