Photographing System Employing a Shoulder-mounted PTZ Camera for Capturing the Composition Designated by the User's Hand Gesture

Shunsuke Sugasawara and Yasuyuki Kono

Graduate School of Science and Technology, Kwansei Gakuin University
2-1 Gakuen, Sanda, 669-1137, Japan
{enm85486, kono}@kwansei.ac.jp

Abstract. We have developed the wearable system for photographing of the scenery / composition designated by the user's hand frame with a shoulder-mounted camera. The hand frame is the gesture of making a rectangular enclosure with both hands when the user considers the composition of a picture. The system detects the hand region of the user from an image of a head-mounted camera, and gets a "picking region image" by recognizing the hand gesture. The picking region is the region in the hand frame indicated by the user through the image of the head-mounted camera. It photographs high resolution image of the similar composition as the picking region image, called "target region image" by controlling PTZ (pan / tilt / zoom) of the shoulder-mounted camera. It performs robust control on noise such as the user's body sway.

Keywords: Shoulder-mounted camera, PTZ, Feature point tracking, Wearable system

1 Introduction

We have developed the wearable system for photographing of the scenery / composition designated by the user's hand frame with a shoulder-mounted camera. An example of a hand frame is shown in Figure 1. The hand frame is the action of making a rectangular enclosure with both hands when the people decides the composition of a picture. The hand frame is easy and intuitive method of decides composition. The photographing of the determined composition with the camera has two problems. One is that the scenery designated by the naked eye is different from the scenery through the camera lens. Another is that it takes some actions before transition from a hand frame to an attitude of photographing. In this research, we aim to create the wearable system that take the similar composition as the designated region by the user's hand frame with the shoulder-mounted PTZ (pan / tilt / zoom) camera.

Fig. 1. A hand frame

2 Related Work

Fuchi et al. [1] proposed a system that photographs the range designated by the user's hand gesture with three cameras installed on the ceiling. It's usable environment is limited, because their system requires to install cameras preliminarily. Furthermore, the taken pictures are not from user's view. Chu et al. [2] proposed an interaction technique enabling users to control digital camera's functions such as pan, tilt, and shutter using hand gestures when taking self-portrait pictures. Our proposed wearable system saves an arbitrary area in the user's viewpoint as a still image. Sakata et al. [3] proposed a system that supports remote collaborative work with shoulder-mounted camera and laser pointer. It always photographs and projects the remote collaborator's instruction as a laser on fixed point by estimating the operator's motion and controlling the panning and tilting of the camera. Their research doesn't control the zoom function of the camera. In this research, the system controls pan / tilt robustly against the user's body sway to hold a specific region even after zooming the camera.

3 System Configuration

The system configuration is shown in Figure 2. The user wears an HMD (Head-Mounted Display) with a camera for checking the composition of the picture and the PTZ camera on the right shoulder for taking high resolution picture of the similar composition as the picking region image. 1) The picking region is the region in the hand frame indicated by the user through the image of the head-mounted camera. 2) The system searches for the region equivalent to the picking region image from the image of the shoulder-mounted camera, and determines the discovered region as the target region. 3) The system controls the PTZ of the shoulder-mounted camera and saves the camera image as a still image.

Fig. 2. System configuration

4 Getting a Picking Region Image

In this research, the system gets a picking region image designated by the user to perform gesture determination process on user's hand region detected from the image of the head-mounted camera. The rough flow of user operations is shown in Figure 3. II) To detect skin color region from an image of a head-mounted camera, we apply the expression proposed by Song et al. [4] to each pixel and the image is binarized in skin color region. III) Since the hand region and the background region with similar RGB values may overlap, the edges obtained by applying the Sobel filter to the image of the head-mounted camera are set as boundary lines and these are separated. IV) It detects two regions from the largest region after separation. It decides that the one located on the lower left is the left-hand and the one located on the upper right is the right-hand region.

The system recognizes hand gestures from hand region detection image obtained above. V) This system recognizes three hand gestures: designation (of a picking region), waiting (for photographing), performing (for photographing). User performs these hand gestures successively on this system.

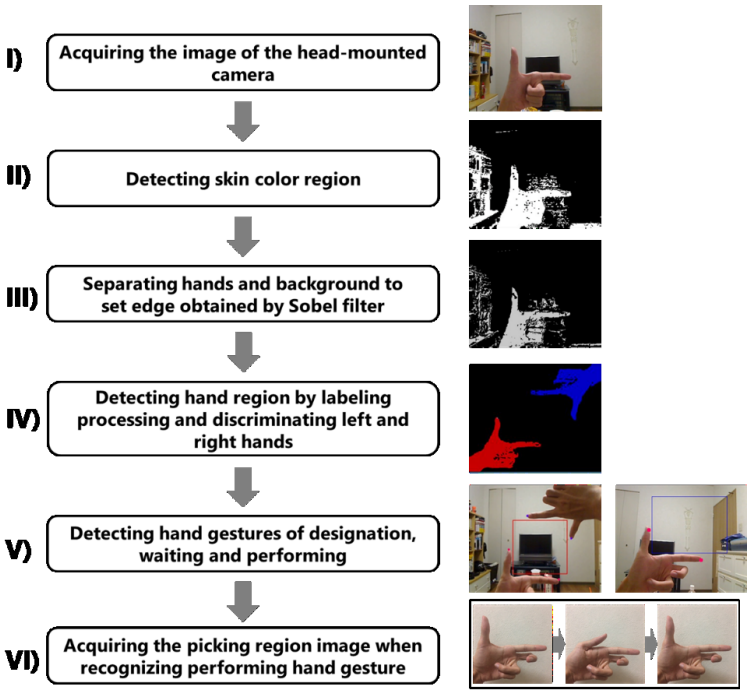

Fig. 3. User operations for getting a picking region image

The "designation" hand gesture determines the composition shown by user's hand frame as the picking region. User's view in the HMD while recognizing the hand gesture is shown in Figure 4, where the red rectangle is designated region and has the base of the fingers of the right hand and the left hand as the diagonal. User can freely modify the size and the position of the picking region by moving the position of hands. In order to recognize this hand gesture, the system detects the hand frame. We define the hand frame as a state where the left and right hands are in proper positional relationship and the thumb and forefinger are open at right angles. First, it detects thumb tip and forefinger tip. Next, it searches a pixel of hand region so as if drawing a circle toward the forefinger with the thumb tip as the center (a round search), and assumes the found point is the base of fingers. It calculates the closing angle of each line connecting the thumb tip and the forefinger tip, the forefinger tip and the base. A similar search is performed centered on the found point, and the crossing angle is determined again using for new found point. These processes are continued until the crossing angle begins to increase. If the minimum crossing angle until the end is approximate to a right angle, the system decides the point that creates the crossing as the base and the hand region forms a hand frame.

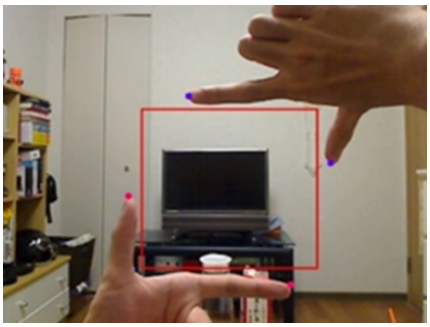

Fig. 4. Recognition of designation hand gesture

The "waiting" hand gesture is recognized when a part of hand region is detected under the left forefinger while detecting a hand frame. This hand gesture is able to be performed with less hand movements such as stretching the middle finger or loosening the fist, and a transition from the "designation" hand gesture is smooth. After recognizing the hand gesture, it records and fix the picking region size and stop detecting the right hand. It records the relative position of the base to the forefinger tip. This allow the user to utilize the system with only the left hand, prevents the right hand from entering the image of the shoulder-mounted camera and becoming occlusion of a picture.

The "performing" hand gesture is recognized when the system detects the action of inclining the thumb to the forefinger side to a certain angle and standing it at right angle again. This is similar to the shutter release operation. In order to calculate the inclination angle of thumb, it calculates the angle of three points the thumb tip, the forefinger tip and the provisional base at the recorded relative position. By using for the provisional root, it is able to prevent situations where the base position is shifted or becoming difficult to calculate the inclination angle. VI) The system gets a picking region image and starts PTZ control of a shoulder-mounted camera.

5 PTZ Control Method

The proposed system searches for the region equivalent to the picking region image designated by the user from the image of the shoulder-mounted camera, and controls the PTZ of the shoulder-mounted camera so that the discovered region is shown on the its whole screen. The region in the camera image may be moved due to the influence of such as the user's body sway. In order to track this, we implement robust control focusing on the positional difference of feature points. The rough flow is shown in Figure 5. A) The system searches for the region equivalent to the picking region image from the image of the shoulder-mounted camera, and determines the discovered region as the target region. B) It pans and tilts the camera so that the target region moves to the center of

the camera image. It estimates the positional difference of feature points in the camera image per pan / tilt degree. C) If the target region is in the center, the system performs to zoom so that the target region is shown on the entire screen of the camera image. D) It detects the positional gap between the camera image and the target region. E) It pans / tilts so that the positional gap is eliminated. If the positional gap is equal to or less than the threshold value, the camera image is saved as a still image.

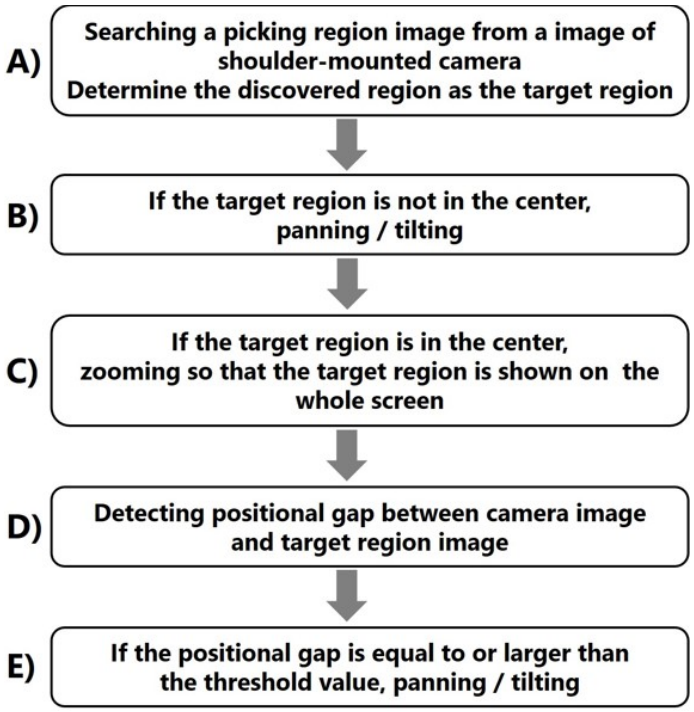

Fig. 5. Flow of PTZ control

5.1 Estimation of the Positional Difference of Feature Points

The system requires to figure out the positional difference in the x direction and the y direction of feature points in the camera image to eliminate the positional gap between the target region and the camera image. It calculates the positional difference of feature points to track feature points by comparing the frames just before and after panning / tilting. An example of observation is shown in Figure 6. While capturing an object A in which the position / shape is invariant, if the camera is panned to the right, A appearing in the camera image moves to the left. It can calculate positional difference by panning from the difference in coordinates of A. However, the positional difference of feature points is affected

by the positional relationship between the user and the object in real space. The system doesn't limit the usable environment, therefore, it observes the actual positional difference on the environment to acquire images and control pan / tilt continuously. It uses AGAST [5] for feature points detection and Lucas-Kanade method [6] for feature points tracking.

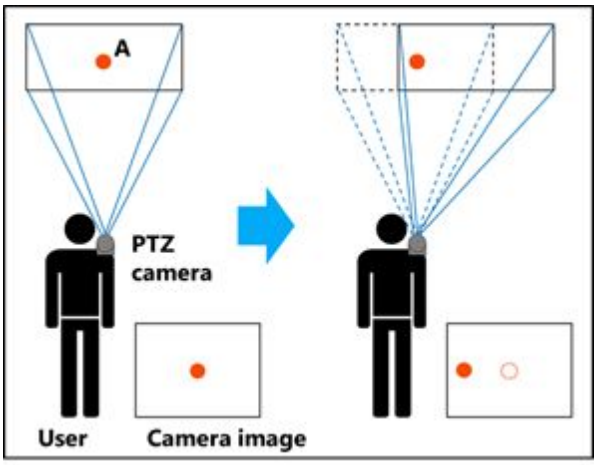

Fig. 6. An example of observation for positional difference of feature points

The observed positional difference contains errors due to the noise such as the user's body sway or changing in the feature points group detected. We implement a method to estimate the positional difference without errors by observation applying Kalman filter [6]. Kalman filter is a process for estimating and controlling an appropriate state of a certain linear dynamic system from observation values including errors. In this filter, two phases are executed for each one step of time. One is the prediction phase that estimates next state of the system using for the information learned in the past and another is the correction phase that acquires the current observation value and adjusts the estimation method model. Since the estimation method is improved by comparing the predicted value and the observation value at each time, it is possible to reduce the error from the prediction value according to accumulation of data. In this research, we prepare for two Kalman filters and estimate the positional difference separately for panning and tilting. It estimates the positional difference before panning or tilting is executed, and after execution, acquires the actual the positional difference and adjusts the estimation method model.

5.2 Tracking the Target Region

The target region in the camera image may move due to the user's body sway. The system controls PTZ to track the target region. After zooming, when the

entire target region is contained in the camera image, the camera image is saved as a still image. PTZ control in the system is divided into two phases. The former phase (phase 1), PTZ control for "approaching" the target region, is a process from the initial state of shoulder-mounted camera operation until the target region is centered by panning / tilting and adjusted the zoom magnification. It performs template matching to track the target region from the camera image. The latter phase (phase 2), PTZ control for "staying" in the target region, is a process of controlling the panning / tilting in order to eliminate the positional gap between the camera image and the target region after zooming. It performs feature points tracking to track the target region from the camera images.

When the "performing" hand gesture is detected from the images of the head-mounted camera, the target region image is acquired and the phase 1 is launched. It searches the target region from the image of the shoulder-mounted camera. It utilizes for template matching of a picking region image as a template image for the search method. It obtains the coordinates of discovered target region, and if the target region isn't in the center, controls the panning / tilting of the camera so that the image to move to the center. It performs observes and estimates the positional difference of feature points described in Section 3.1. When the target region is detected at the center of the camera image, the system controls zooming so that the target region is equal to the size of the camera image.

After Zooming, the phase 2 is launched. It detects feature points from the camera image and the target region image, and performs feature point tracking between these images. The system calculates the positional gap between the camera image and the target region image from feature points that can be tracked. It pans and tilts the camera to the direction for decreasing the positional gap. By estimating the positional difference of feature points per panning / tilting from the estimation method described in Section 3.1, the rotation angle necessary for eliminating the positional gap is obtained. However, position and posture of the shoulder-mounted camera are always fluctuating due to the user's body sway, so it is difficult to obtain the image as expected. The system again calculates the positional gap after the panning / tilting and continues control. If the positional gap is equal to or less than the threshold value, it determines that the photographing is successful and saves the camera image as a still image. We set the threshold to be half of the positional difference of the feature points in the x direction 1-degree panning and in the y direction 1-degree tilting. This is to prevent situations where the positional gap increases after panning / tilting.

6 Experiment

6.1 Experimental Method

We conducted an experiment to evaluate the effectiveness of the PTZ control method in this system. A subject wears the cameras on the head and the right shoulder. The camera mounted on the right shoulder is a PTZ camera. A subject look see the image of the head-mounted camera in an upright state and designates five regions as picking regions. The five regions are, in order, upper

left, upper right, lower left, lower right, and center of the camera image. It is shown in Figure 7. In the eight different environments, we determine the similarity between the image of the picking region specified from the head-mounted camera image and the image saved as a still image from the shoulder-mounted PTZ camera. The eight environments are defined as Environment 1 to Environment 8, and shown in Figure 8. The similarity calculation obtains the value of zero mean normalized cross correlation [6] of between the saved image and the picking region image expanded to the same size. The value when two images are perfectly matched is 1, and the value when the two images are not correlated at all is 0. The higher the similarity between two images, the more positive the value. In this experiment, the head-mounted camera and the shoulder-mounted camera acquire the image of 640 pixels horizontal and 480 pixels vertical. Regarding the picking region, the horizontal size was fixed at 400 pixels and the vertical size was fixed at 300 pixels. At the time of executing the PTZ control, experimenter is always in a state in which the right arm is being lowered, and it is assumed that the experimenter doesn't intentionally move the body.

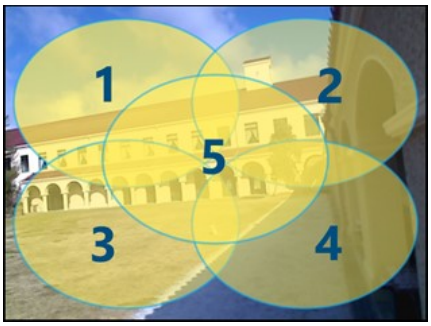

Fig. 7. Designated five regions

Fig. 8. Environment 1 to Environment 8

6.2 Result

Results of experiments in eight environments are shown in Table 1 to Table 8. The rows 1, 2, 3, 4, 5 of the table are regions in the image of the head-mounted camera shown in Figure 7. The table shows the success or failure of photographing and the similarity when the region is designated as the picking region. The success or failure of photographing is judged to be failed in the case where the target region is missed for some reason. Furthermore, if a shoulder-mounted camera image is saved as a still image however an erroneous region is, it is judged as a failure. In the case where the correct region is saved it is judged to be successful.

Table 1. Result of experiment in Environment 1

	1	2	3	4	5
Success or failure of photographing	Success	Success	Success	Success	Success
Similarity	0.528	0.420	0.584	0.446	0.393

Table 2. Result of experiment in Environment 2

	1	2	3	4	5
Success or failure of photographing	Failure	Success	Success	Success	Failure
Similarity	-	0.733	0.822	0.712	-

Table 3. Result of experiment in Environment 3

	1	2	3	4	5
Success or failure of photographing	Success	Success	Success	Failure	Success
Similarity	0.496	0.681	0.809	-	0.762

Table 4. Result of experiment in Environment 4

	1	2	3	4	5
Success or failure of photographing	Success	Success	Success	Failure	Success
Similarity	0.496	0.681	0.809	-	0.762

Table 5. Result of experiment in Environment 5

	1	2	3	4	5
Success or failure of photographing	Failure	Success	Success	Failure	Success
Similarity	-	0.850	0.709	0.662	0.831

Table 6. Result of experiment in Environment 6

	1	2	3	4	5
Success or failure of photographing	Success	Success	Success	Failure	Success
Similarity	0.662	0.681	0.535	-	0.586

Table 7. Result of experiment in Environment 7

	1	2	3	4	5
Success or failure of photographing	Success	Failure	Success	Success	Success
Similarity	0.642	-	0.735	0.368	0.574

Table 8. Result of experiment in Environment 8

	1	2	3	4	5
Success or failure of photographing	Success	Failure	Success	Failure	Success
Similarity	0.621	-	0.474	-	0.563

7 Discussion

7.1 Limitation

We obtained the experimental result photographing was successful 30 out of 40 regions. The system has some limitations.

In Environment 4, the system failed on the lower right region. The causes include halation of an image. Halation is a phenomenon in which an especially highlighted part of the image blurs white. We think that the system executed to pan / tilt with an erroneous rotation angle and the positional gap increased, because of erroneous feature point tracking in strong light hit parts. The rotation angle and the execution time of the next panning / tilting increase in proportion to the length of the positional gap. Thus, the system erroneously obtains a frame in the middle of panning / tilting. The system loses track of the target region and judges as a failure, because feature points are hardly detected from the frame in the middle. We think that it improves the PTZ control accuracy by removing points having outliers from feature point groups. It is required to set a standby time for frame acquisition after execution in proportion to the rotation angle of panning / tilting.

In the photographing of the lower right region of the environment 5, still images were saved for the wrong region. The designated picking region image and the saved image is shown in Figure 9. The reason is that the distribution of pixel values is similar between the picking region and the wrong region, and it is thought that the template matching was erroneous. We think that it is possible to reduce errors by limiting the search range using for the positional relationship between the two cameras.

Fig. 9. The picking region image (left) and saved image (right) of the Region 4 of Environment 5

Although the photographing was successful, the similarity decreased in several cases due to the roll of the saved image being shifted to left compared with the picking region image, because the shoulder-mounted camera fell off the top of the shoulder. As an example, the picking region image and saved image of the upper left region of the Environment 4 are shown in Figure 10. Implementing adjustment of the zoom magnification and two-dimensional affine transformation

enables the system to save the image in which the gap of the rolling direction is corrected with less decreasing the image quality so much. Feature point matching of ORB features is robust to rotation and varying illumination. The system performs ORB feature point matching after zooming, and calculates the rolling gap using for the positional relationship of the matching points between the picking region image and PTZ camera image. Since the error in the initial setting matching is large, it reduces erroneous matching to restrict the positional relationship.

Fig. 10. The picking region image (left) and saved image (right) of the Region 5 of Environment 4

7.2 Usability

We mainly implemented two interactive technics, hand detection and shoulder-mounted PTZ camera control. These technics have the potential to be applicable to various applications.

The hand detection method can be implemented employing only a monocular camera regardless of indoor or outdoor environment. The interface based on this method is intuitive, since the user can operate a system just by holding their hand in front of the camera. Besides activating each function of a system, it can be applied for indicating a specific region on computer vision.

We implemented PTZ camera control robust to the user's body sway as a wearable system. This camera control matches compositions from two different viewpoints and maintains them by removing the influence of external factors. We think that this method can be applied in the field of personal robot or remote annotation. Sharing objects to be gazed with others will promote mutual understanding.

8 Concluding Remarks and Future Work

In this research, we have developed the system for photographing of the same composition designated by the user's hand frame with a shoulder-mounted PTZ

camera. The system gets the picking region image to detect a particular hand gesture from the image of the head-mounted camera. It searches for a region equivalent to the picking region image from the image of the shoulder-mounted camera, and decides the discovered region as the target region. It controls panning / tilting / zooming so that the target region is shown on the whole screen of the shoulder-mounted camera. When the shoulder-mounted camera puts the entire target region in the camera image, the camera image is saved as a still image.

As a result of the experiments, we successfully photographed 30 PTZ camera images out of 40 regions, and showed the effectiveness of this system. However, the system has some limitations. I would like to respond to these, and revalidate by taking into account the size change of the picking region and the user's body sway.

In the future, we have plan to develop the photographing system employing a small Unmanned Aerial Vehicle (UAV), a drone. This system requires a smartphone instead of HMD, and a drone instead of a shoulder-mounted PTZ Camera. It is difficult to track a target region with a shoulder-mounted PTZ Camera when trying to take a very small far region. We think that this problem can be solved by flying a drone equipped with a camera to the position where it can photograph. Furthermore, it is inconvenient for the user to wear video see-through HMD on a daily basis. We think that setting the pinch-out operation on the touch panel of the smartphone to the designated operation of the picking region enables a burden on the user to be reduced. We want to develop an interface to automatically generate the flight plan of the drone based on computer vision and implement it in a wearable system.

References

1. Kazuma Fuchi, Shin Takahashi, Zirou Tanaka.: A system for taking a picture by hand gesture. In: Proceedings of the 70th National Convention of IPSJ, Japan (2008).
2. Shaowei Chu, Jiro Tanaka.: Hand Gesture for Taking Self Portrait. In: Human-Computer Interaction. Interaction Techniques and Environments (HCI 2011), pp. 238-247, USA (2011).
3. Nobuchika Sakata, Takeshi Kurata, Masakatsu Kourogi, Hideki Kuzuoka, Mark Billinghurst.: Remote Collaboration using a Shoulder-Worn Active Camera/Laser. In: Multimedia, Distributed, Cooperative, and Mobile Symposium (Dicomo 2004), pp. 377380, Japan (2004).
4. Jie Song, Gbor Srs, Fabrizio Pece, Sean Ryan Fanello, Shahram Izadi, Cem Keskin, Otmar Hilliges.: In-air Gestures Around Unmodified Mobile Devices. In: ACM User Interface Software and Technology Symposium (UIST 2014), pp. 319-329, USA (2014).
5. Elmar Mair, Gregory D. Hager, Darius Burschka, Michael Suppa, Gerhard Hirzinger.: Adaptive and generic corner detection based on the accelerated segment test. European Conference on Computer Vision (ECCV'10), Greece (2010).
6. Gary Bradski, Adrian Kaehler.: Learning OpenCV ——Computer Vision with the OpenCV Library. O'Reilly Media, USA (2008).