A Walkthrough System to Display Video Corresponding to the Viewer’s Face Orientation

T. Watanabe¹, C. Liu², and S. Shibusawa²
¹Graduate School of Science and Engineering, Ibaraki University, Hitachi, Ibaraki, Japan
²Department of Computer and Information Sciences, Ibaraki University, Hitachi, Ibaraki, Japan

Abstract - Walkthrough systems have hitherto been developed with the ability to produce a virtual display of a person’s field of view. If the actual position and orientation of a person’s face can be ascertained, then walkthrough images could be created to match this person’s face direction, thereby recreating video corresponding to the view from this person’s face direction. A possible application of this system is in store surveillance systems, where the direction in which someone is looking can be estimated to visualize what they are looking at in order to identify suspicious behavior. Therefore, we aim to produce a walkthrough system that uses image processing to acquire a person’s position and face direction, and displays video corresponding to this person’s face direction. In this system, multiple omnidirectional cameras are used to estimate the positions of people based on the bearings of moving bodies acquired by each camera. People’s face orientations are estimated from the skin regions of their faces. We have created a system that uses a person’s position and face orientation obtained in this way to provide an observer with walkthrough video corresponding to the person’s face orientation. To evaluate this system, we conducted experiments on the accuracy of human positions and face orientations and the accuracy of the output video, and we confirmed that the system can display video images corresponding to a person’s face direction.

Keywords: face direction, walkthrough system, omnidirectional camera, surveillance, skin area

1 Introduction

In recent years, surveillance systems have been used for purposes such as monitoring and recording the status of facilities. Due to improvements in surveillance technology and reductions in the price of sensor devices, it is now possible to use large numbers of sensors in surveillance systems. This has enabled the construction of systems that can monitor wide areas [1]. Situation recognition systems have also been studied for use in monitoring applications [2]. By learning the traffic lines of people acquired by this system, it is possible to predict the direction in which people will move based on their previous traffic lines. Recently, omnidirectional cameras [3],[4] have also been used in monitoring systems. These devices consist of a camera and a hyperboloidal mirror, and are capable of capturing an entire 360° field of view. With an omnidirectional camera, it is easy to provide wide coverage within which it is possible to track targets reliably.

Researchers have also been studying how to estimate the orientation of human faces [5],[6]. Most of these studies use the face orientation direction as the direction in which the person is interested. For example, in store surveillance applications, this technique can be used for marketing by identifying which product shelves the customers are looking at based on their face directions. It is also expected that the face direction can be used to identify suspicious behavior in surveillance applications. Normally, when people are committing crimes, they often look in different directions to the people around them, and their gaze point also changes in a different way. Someone who is about to commit a crime will often perform actions such as checking their surroundings, glancing at surveillance cameras and mirrors to search for blind spots, and loitering for prolonged periods [7],[8]. If video of the face direction can be displayed so that an observer can check for unusual behavior, this might be a useful way of preventing crimes before they take place.

Currently, studies are underway to identify abnormal behavior based on the traffic lines of people acquired from video images [9],[10]. In these studies, although it is possible to detect people who are acting strangely, it is not possible to figure out which way they are looking, or what they are looking at. For this reason, it is difficult for an observer to predict suspicious behavior based on positional information alone, such as traffic lines.

A number of walkthrough systems have been developed that can produce a virtual representation of a person’s field of view [11],[12]. So far, walkthrough systems can only operate in virtual spaces. But if a person’s actual position and face direction can be ascertained, then it should be possible to generate walkthrough video corresponding to this person’s face direction.

In this study, we create a walkthrough system that uses image processing to acquire a person’s position and face direction, and display video corresponding to this person’s face direction. This system first acquires a person’s position and face direction as time-series data based on the video from multiple omnidirectional cameras. At the same time, a walkthrough space is constructed from the acquired video. The person’s position and face direction are associated with a position and field of view inside this walkthrough space, and a walkthrough video oriented in this direction is output as a representation of this person’s view. An observer can use this
system to check the video corresponding to the face direction of the person being monitored. This gives the observer a better understanding of the objects and areas that the person is looking at, which aids in the identification of suspicious behavior.

To evaluate the system, we performed experiments to measure the face direction estimation accuracy, and to compare the system’s video output with ordinary camera images captured from the same position with the same orientation. Our experiments showed that accurate estimation is possible in the central region covered by multiple cameras, but that the estimation accuracy declines at positions close to a camera and at positions that are distant from all the cameras. As for the estimation of face directions, it was found that accurate estimation is possible for people looking towards the camera, but the accuracy drops off for faces looking in the opposite direction. In the results of an experiment to compare the output of the system with the output of an ordinary camera, it was found that this system’s output is similar to the camera image in the horizontal direction, and can display video of the face direction to some extent.

2 Related research

2.1 Monitoring systems

The number of cameras and other sensors used in monitoring systems has been increasing. This has also led to the construction of large-scale systems [1],[13]. Recently, monitoring systems have been constructed with the ability to make decisions differently in different situations through additional functions of situation recognition [2]. For example, the speed at which vehicles are judged to be driving dangerously can be changed by recognizing the distance between vehicles and the traffic levels for this time of day. In such a system, it is possible to predict the direction in which the target objects move and how they will behave.

However, these monitoring systems do not obtain face direction information, which would be useful for recognizing suspicious behavior. They also lack functions for specifically displaying video corresponding to the face direction of a surveillance target.

2.2 Face direction estimation

In previous studies to estimate the direction of human faces, face directions are mainly estimated either by a skin region method, or by a machine learning method. In the skin region method [14], the head position is identified in moving images obtained by a technique such as background subtraction, and an attempt is made to estimate the face direction from the skin regions identified as parts of the face. In machine learning methods [15], learning is performed by using images of different human heads in different orientations as input for machine learning, and by matching the current acquired frame to the machine learning results. There have also been many studies aimed at estimating the face direction of humans from the arrangement of parts of the face [16].

Another study, currently in progress, aims to estimate the actual gaze direction from detailed face images [17]. In this study, a person’s position is acquired from multiple fixed cameras, and detailed face image data is then acquired by pointing a PTZ camera in this direction in order to estimate the person’s field of sight. These methods require high-resolution face images, and thus have a smaller effective range and are less suited to surveillance applications.

2.3 Identifying suspicious behavior

Various sensors are used to identify suspicious behavior from behavior patterns [9]. In this study, a data set of human traffic lines obtained from devices such as cameras, GPS, and laser radars is used to detect people who are acting outside the normal range of human behavior patterns, and to classify behavior patterns. So far, no system has been able to reproduce video images in the direction of a person’s face to allow an observer to check what the person is looking at.

3 Estimation of head position and face direction

3.1 Head position

Fig. 1(a) shows the appearance of an omnidirectional camera. An omnidirectional camera can capture images over an entire 360° field of view. It captures omnidirectional video images by using a camera to capture a view of its surroundings reflected in a hyperboloidal mirror. Fig. 1(b) shows the sort of image that can be captured with an omnidirectional camera.

(a) Omnidirectional camera   (b) Circular image captured by this camera

Figure 1.  Omnidirectional camera

In this sort of omnidirectional image, the head of a person standing nearby will be positioned near the periphery of the image, as shown in Fig. 2. In this case, the head position can be estimated by using an image obtained by applying a distance transformation to the omnidirectional image. Distance transformation is a transformation to replace pixel values that have a value of 1 in a binary image with values corresponding to their distance from the closest pixel with a value of 0. In addition, the pixel value at the center of the head in a distance transform image is at least as large as some fixed value, that is, threshold value. It can therefore be assumed that the pixels furthest from the center of the image whose pixel value is greater than some fixed quantity correspond to the person’s head. For example, Fig. 3(b) shows the image obtained from the distance transformation of Fig. 3(a). When the head position is obtained from this image, the results are as shown in Fig. 3(c).
3.2 Estimation of face direction

In an omnidirectional image containing a person, as shown in Fig. 4(a), we first find the following coordinates:

- Center of the head
- Center of the skin region of the head

This is shown in Fig. 4(b). Then, we find the vector from the central coordinates of the head in the omnidirectional image to the centroid coordinates of the skin region of the head. This is shown in Fig. 4(c). Since the face appears tilted as shown in Fig. 4(a) when using an omnidirectional camera, the transformation shown in Fig. 4(c) is used. In Fig. 4(c), \( r \) is the distance between the center of the image and the centroid of the skin region, \( R \) is the distance between the center of the image and the center of the head, \( \theta \) is the difference between the angles subtended by the centroid of the skin region and the center of the head relative to the center of the image, and \( \Phi \) is the angle used for estimation. This technique makes use of the characteristic that, when concentrating on parts of the head that are covered in hair and on the skin region of the face, the skin region lies at under the center of the head when the face is pointing straight at the camera, and moves to the left and right when the head is turned to the left or right [19]. In this study, it is thought that the face direction can be estimated by associating this angle \( \Phi \) with the actual angle of the face. It should be noted that this method is difficult to apply to people wearing masks or with close-cropped haircuts.

4 Proposed system

4.1 System overview

Fig. 5 shows an overview of the system. From images obtained from the omnidirectional camera, the system creates an image close to the current view of the person being monitored. The system comprises a processing unit that estimates a person’s position and performs face direction based on image processing of the camera output, and a walkthrough display unit. Video of the person’s face direction is produced based on the person’s position and face direction as acquired by the processing unit, which are converted into a walkthrough by the walkthrough display unit.

4.2 Estimating a person’s position

A person’s position is estimated from multiple omnidirectional camera images. Fig. 6 illustrates the position estimation method. First, as shown in Fig. 6(a), the person’s outline is acquired from the region of movement obtained by background subtraction. This outline is scanned to determine the angular range of the region occupied by the moving object from the center of the image. Next, as shown in Fig. 6(b), the angular ranges of moving objects acquired by multiple omnidirectional cameras are displayed in a planar space, and the region where the angular ranges of all cameras overlap is acquired as the region where a person exists. The centroid of this region is taken to be the person’s position. Fig. 6 shows how this is done.
4.3 Estimation of face direction

The face direction is estimated according to the method of Section 3.2. First, background subtraction is used to extract a moving object. The head position of this moving body is then obtained from the method of Section 3.1. The skin region of the moving object is also extracted, and its centroid is calculated.

Fig. 7 shows the head position and the centroid of the skin region plotted on the moving object image. The circle indicates the head position, and the cross indicates the centroid of the skin region. Fig. 7 shows the results obtained for people facing in different directions: (a) straight towards the camera, (b) turned at 45°, (c) facing sideways, (d) facing directly away from the camera. Between images (a), (b) and (e), the angle $\Phi$ increases. In this way, the angle $\Phi$ in Fig. 4(b) can be obtained according to the method of Section 3.2. The face direction is estimated from the correspondence between this angle and face direction. In Fig. 4(d), the person is facing in the opposite direction to the camera so the face parts are not captured and the area of the skin region is very small. Therefore, a threshold value is set for the area of the skin region. A moving object that falls below this threshold is judged to be facing away from the camera. These processes are performed for all the moving object regions detected by background subtraction.

![Figure 6. Acquiring a person’s position](image)

4.4 Correspondence between person’s position and face direction

The people detected according to Section 4.3 are correlated with the results of face direction estimation for each camera. This process yields face direction information from a number of cameras for each person. These results are weighted and added together, and the result is treated as the final face direction. If there are $n$ cameras within range, the weighting $w_i$ of camera $i$ is defined by the following formula:

$$w_i = \frac{S_i / M_i}{S_1 / M_1 + S_2 / M_2 + \cdots + S_n / M_n}$$

(1)

In (1), $M_i$ and $S_i$ represent the area of the entire moving object region and the area of the parts corresponding to the skin region in image from camera $i$. In the method of Section 3.2, the face direction is estimated with the highest accuracy when the face is oriented straight towards the camera. Therefore, greater weighting is applied to faces that are turned more towards the camera.

4.5 Creating the video walkthrough

A walkthrough space can be created relatively easily from the omnidirectional images. Fig. 8 shows an overview of a walkthrough based on an omnidirectional image. In this method, the circular image captured by the omnidirectional camera is first projected into a panoramic image. This image is associated with the walkthrough space, and its visible range is set. Movement of the walkthrough view and operations to move forwards and backwards are implemented by panning this range to the left and right and zooming in and out.

In the walkthrough system produced in this way, output video is produced using the positions and face directions of people obtained as described in Sections 4.3 and 4.4. First, the camera closest to the person’s position in the walkthrough space is ascertained, and the video from this camera is used in the walkthrough. Next, the range of the image to be displayed in the walkthrough from this person’s position and face direction is determined. This range is extracted from the video and displayed by enlarging or reducing it to match the window size.

![Figure 8. Walkthrough based on an omnidirectional image](image)

5 System implementation

We implemented the system according to the design of Section 4. Fig. 9 shows the system in operation. The system displays separate windows for the video acquired from each camera, the results of background subtraction to extract moving objects, the head positions of these moving objects, the centroid positions of the facial skin regions, the coordinates and estimated face directions of people relative to the cameras, the traffic lines of these people, and the calculated walkthrough video.
We calculated the errors between the actual positional coordinates and the positional coordinates output by the system, and from these we calculated the square root of the sum of squares of the errors along the X and Y axes. Fig. 12 shows the variation of the system’s output error when changing the Y-axis distance from the positions of each camera A, B and C. The errors became large at positions close to the camera and far away from the camera. One reason for this is the fact that the moving object appeared larger than necessary at positions close to the camera. Conversely, the increased errors at locations far away from the camera are thought to be due to the inability to acquire a sufficient number of pixels over the moving object. For these reasons, the most accurate estimations were made at positions close to the center of the region surrounded by the cameras. To improve the precision of position estimation, it is necessary to extract only the region where a person’s feet touch the floor in order to narrow down the person’s position more accurately.

6.2 Estimation of face direction

6.2.1 Experimental environment

In the layout of Fig. 13, we estimated face directions using two cameras. In this experiment, the target person was first placed at a distance of 100 cm from the camera. At this position, the person’s orientation was changed from 0 to 360° in 45° increments while estimating the face direction. This was repeated at distances of 100–500 cm, increasing in steps of 100 cm. In this test, the direction straight towards the camera was taken as 90°. The results were calculated from the average estimated value over 20 frames.
6.2.2 Experimental results and discussion

Fig. 14 shows the target person’s actual face direction and the errors of the system’s output. In the overall test, the errors were small when the face direction was between 0 and 180°, but were large between 225 and 315°.

![Figure 14. Errors in the estimation of face direction](image)

Although there were errors in the results, we were able to estimate face directions to some extent. The errors were particularly small when the target was facing the camera. For face directions where there was a large number of face pixels in the image, the estimation accuracy was relatively high. There is also a tendency for the face direction estimation error to increase as the distance from the camera increases. This is because people occupy a smaller number of pixels as they move away from the camera, resulting in lower accuracy. Also, the lack of symmetry in the errors recorded at face directions of 0° and 180° is thought to be affected by the lighting direction in the test environment.

6.3 Comparison with photographic images

6.3.1 Experimental environment

To evaluate this system as a whole, we performed an experiment to compare the walkthrough video with pictures captured by an ordinary camera. The experiment was performed with four different layouts as shown in Fig. 15. In this experiment, we used two cameras installed 2 meters apart. First, a target object was placed in the test space. An ordinary camera was then used to photograph the target object from the person’s location. We also saved the system’s output video produced when oriented from this position in the direction of the target object. We then compared the saved images with the images captured by the ordinary camera.

![Figure 15. Experimental layouts for comparison with photographic images](image)

6.3.2 Experimental results and discussion

Fig. 16(a) and (b) show the output of the system and an image taken with a camera in layout 1. Similarly, the results for layouts 2–4 are shown in Figs. 16(c) through (h) respectively. In Fig. 16, the position of the target object is enclosed by a circle. Also, Table 1 shows the pixel offset percentage between the photographic image and the system output at the target object. From the results of layouts 1 and 2, it can be seen that the image offset percentage is less than 10% in the X direction and less than 20% in the Y direction, and that in the horizontal plane it was possible to display video in the face direction. In layout 3, the target object was not captured and it was not possible to reproduce video in the face direction. Also, in layout 4, the offset in the Y direction of the image became larger.

In layout 3 as shown in Fig. 15(c), this system was unable to produce video in the face direction for a subject at an intermediate position with respect to the camera. To make it possible to do so, it would be necessary to increase the range over which video can be produced by increasing the number of cameras. Also, as shown in Figs. 16(g) and (h), this system produces large errors when reproducing directions that are tilted up or down. To improve the result, it would need to be capable of estimating the vertical component of the face direction.

![Figure 16. Results of comparison experiment](image)
Table 1. Offset of system output from photographic image

<table>
<thead>
<tr>
<th>Layout</th>
<th>Offset of X coordinates [%]</th>
<th>Offset of Y coordinates [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layout1</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Layout2</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Layout3</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Layout4</td>
<td>16</td>
<td>37</td>
</tr>
</tbody>
</table>

7 Conclusions

In this study, we have produced and evaluated a walkthrough system that uses image processing to acquire a person’s position and face direction, and displays video corresponding to this person’s face direction. In this system, multiple omnidirectional cameras are used to determine a person’s position based on regions of movement. The face direction is estimated from the head position and skin region. We have also produced a walk-through system that uses an omnidirectional image to create a walkthrough video in the person’s face direction.

In evaluation experiments, we examined the system’s accuracy in the estimation of human positions and face directions, and we compared the system’s output with ordinary camera images captured from the same position and orientation. The positions of humans can be estimated accurately in regions surrounded by multiple cameras, but the accuracy decreases with increasing distance from the cameras. Face directions can be estimated accurately with low errors for faces that are pointing towards the camera. Also, in an experiment to compare camera images with the system output, we were able to acquire video close to the camera images in the system output for most positions.

In future studies, it will first be necessary to improve the face direction estimation method. In the current method, the face direction is estimated for each individual camera, but what is needed is a mechanism whereby the face direction can be estimated by the system as a whole. Specifically, it will be necessary to judge the face direction in three dimensions by associating 3D spatial coordinates with the image of the moving region in each camera used for face direction estimation. Furthermore, it will be necessary to introduce distributed processing on multiple PCs in order to reduce the processing load so that high-resolution images can be used.

8 References