Measuring the Arrangement of Multiple Information Devices by Observing Their User's Face

Saori Kikutani[1], Koh Kakusho[1(✉)], Takeshi Okadome[1], Masaaki Iiyama[2], and Satoshi Nishiguchi[3]

[1] School of Science and Technology, Kwansei Gakuin University, Sanda, Japan
{kiku, kakusho, houmi}@kwansei.ac.jp
[2] Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan
iiyama@mm.media.kyoto-u.ac.jp
[3] Faculty of Information Science and Technology, Osaka Institute of Technology, Hirakata, Japan
satoshi.nishiguchi@oit.ac.jp

Abstract. We propose to measure the 3D arrangement of multiple portable information devices operated by a single user from his/her facial images captured by the cameras installed on those devices. Since it becomes quite usual for us to use multiple information devices at the same time, previous works have proposed various styles of cooperation among the devices for data transmission and so on. Other previous works propose to coordinate the screens so that they share the role of displaying contents larger than each screen. Those previous works obtain the 2D tiled arrangement of the screens by detecting their contacts with each other using sensing hardware equipped on their edges. Our method estimates the arrangement among the devices in various 3D positions and orientations in relation to the user's face from its appearance in the image captured by the camera on each device.

Keywords: Multiple portable devices · Device coordination · Screen arrangement · Facial image processing · Camera calibration

1 Introduction

It becomes quite usual for us to use multiple information devices such as mobile or tablet PCs, smartphones, PDAs at the same time. Aiming to take full advantage of those devices, many previous works have proposed various styles of cooperation among those devices for transmission and sharing of selected data among the devices [1–4], operation of the contents displayed on the screens [5, 6], and so on. Some other previous works propose to coordinate the screens so that they share the role of displaying contents, which, for example, are larger each screen.

In order to make several screens coordinated with each other for this purpose, we need to measure their arrangement in advance. The previous works described above obtain the tiled 2D arrangement among the screens based on their adjacency detected

N. Streitz and P. Markopoulos (Eds.): DAPI 2015, LNCS 9189, pp. 296–304, 2015.
DOI: 10.1007/978-3-319-20804-6_27

by their physical contact with each other using sensing hardware equipped on their edges [7, 8]. However, as we experience in setting the screen arrangement of a PC and its external display manually, we often prefer more various arrangements such as those of screens placed a little bit apart or in contact with each other just at their corners. It is also useful to display a 3D virtual space by specifying a 3D screen arrangement where we are surrounded by the screens.

In this article, we discuss how to measure those various arrangements of multiple screens. Since recent information devices are usually equipped with cameras, we measure the screen arrangement of the devices using the images of the user captured by those cameras on the devices. In the field of computer vision, the 3D geometric arrangement of multiple cameras is conventionally measured by the method for so-called *strong camera calibration*, in which the same set of markers whose 3D positions have already known is observed by each camera. We employ the feature points on the user's face for those markers. When a user is operating some information devices, the user should keep gazing at their screens and thus his/her face can be observed from the camera on each device. The facial feature points of the user appearing in the camera image can be extracted by facial image processing, although those facial feature points may sometimes fail to be extracted depending on their appearance in the image. The positions of the facial feature points on the face are approximately available because they are similar for any persons, although some amount of personal differences are included in those positions.

By considering these properties of facial feature points, we discuss how to measure the arrangement of information devices with their cameras from the images of the user's face. In the discussion, we also try to cope with the failure in extracting facial feature points by introducing the continuity of the change in the arrangement of the devices as a geometric constraint.

2 Measuring Arrangement of Devices

Measuring Geometric Arrangement of Each Device and the User's Face. As we described above, we measure the arrangement of portable information devices at each moment of their operation by the same user at the same time by employing the feature points of the user's face as the markers for the strong calibration of the cameras on the devices. The 3D position of the k-th feature point on the face is denoted by $\boldsymbol{p}_k (k = 1, \cdots, K)$, where K is the number of the feature points that we employ for the calibration. These 3D positions of the facial feature points are represented by the face-centered coordinate system with its origin at the center between the two eyes, the x axis passing through the eyes from left to right on the face, the y axis directed downward on the face and the z axis set forward from the face. The 2D positions where the k-th facial feature point appears in the image captured by the camera on the i-th information device is denoted by $\boldsymbol{q}_k^i (i = 1, \cdots, N)$, where N is the number of the devices. This 2D position is represented by the camera-centered coordinate system with its origin at the optical center of the camera, the x, y axes set rightward and upward to the image plane, and the z axis forward along the optical axis of the camera. The geometric relation among these face-centered coordinate system and camera-centered

coordinate systems for the *i*-th device and the *j*-th device in the situation where the user is gazing at the screens of those devices is illustrated in Fig. 1.

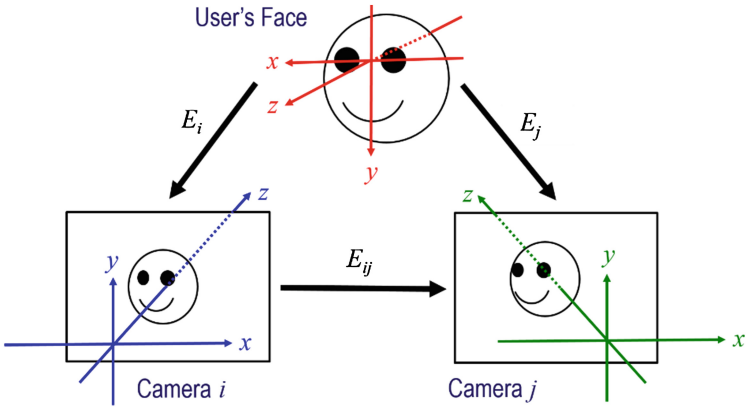


Fig. 1. Geometric relation among the cameras on different devices and the user's face

The geometric relation between $\boldsymbol{p}_k$ and $\boldsymbol{q}_k^i$ can be described as follows:

$$\boldsymbol{q}_k^i = \lambda A_i E_i \boldsymbol{p}_k \tag{1}$$

Where A_i is the matrix of the internal parameters of the camera on the *i*-th device. This matrix mathematically represents the process of optical projection from each point in the 3D space onto the 2D image plane of the camera with scaling parameter λ. The matrix E_i includes the external camera parameters that consist of rotation matrix R_i and translation vector $\boldsymbol{t_i}$, which represent the orientation and the position of the user's face relative to the camera on the *i*-th device, as follows:

$$E_i = \left[\begin{array}{ccc|c} & \boldsymbol{R}_i & & \boldsymbol{t}_i \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \tag{2}$$

In Eq. (1), $\boldsymbol{q}_k^i$ is given as the 2D position of the *k*-th facial feature point extracted from the image obtained by the camera on the *i*-th device by facial feature processing, whereas $\boldsymbol{p}_k$ can be specified from the 3D position of the *k*-th facial feature point of the standard human face if we ignore its personal difference. Matrix A_i of the camera on the *i*-th device can be obtained by preliminary internal camera calibration using usual markers.

Since E_i includes 12 variables, this matrix can be determined by solving the equations above, if $\boldsymbol{p}_k$ and $\boldsymbol{q}_k^i$ are obtained for not less than six facial feature points out of *K*. When the user is gazing at the screen of the *i*-th device, it is expected to be possible to obtain $\boldsymbol{q}_k^i$ for more facial feature points from his/her facial image captured

by the camera on the device by facial image processing. Thus, by solving Eq. (1) for all the variables in E_i, we can measure the geometric arrangement of the user's face and the camera on the i-th device. Since the camera installed on an information device is usually fixed around the edge of the screen with its optical axis perpendicular to the screen, we can estimate the geometric arrangement of the user's face and the screen of the device by measuring the position of the camera on the i-th device relative to its screen in advance.

After $E_1, \cdots, E_N$ for all the cameras on all the devices are obtained, the geometric arrangement of the i-th device and the j-th device $(i \neq j)$ can be calculated from E_i and E_j as follows:

$$E_{ij} = E_j E_i^{-1} \tag{3}$$

Where E_{ij} is the matrix that represents the orientation and the position of the camera on the i-th device relative to that on the j-th device $(j = 1, \cdots, N;\ j \neq i)$.

Continuity in the Geometric Arrangements. When a user is operating multiple information devices, the user is usually not gazing at the screens of all the devices at the same time, but gazing at only one of those screens depending on his/her interest. In that situation, facial feature point extraction is successful for the image obtained by the camera on the device gazed by the user because his/her face is captured by the camera from the very front, whereas extraction of facial feature points often fails for the images obtained by the cameras on the devices that are not currently gazed by the user, depending on his/her facial orientation to the cameras. Failure in extracting facial feature points is also caused by occlusion of those points due to the user's unconscious behavior such as putting a hand over the mouth and so on.

Since the geometric arrangement of different devices are indirectly obtained from the arrangement of each device and the user's face, the information about the arrangement related to the devices with the camera images at the moments when the failure in facial feature point extraction occurs for those images is completely lost during the user's operation of the devices. Moreover, even when the geometric arrangement is estimated for a device and the user's face after facial feature point extraction is successful for the camera image of the device; the estimated geometric arrangement inevitably includes a certain amount of error.

In order to cope with the problems above, we introduce the continuity in the estimated arrangement. The geometric arrangement of the devices and the user's face does not change drastically as far as the user keeps operating those devices at the same time in a similar manner. By considering it, we estimate the geometric arrangement under the constraint that the difference between the estimated arrangements at adjacent moments should be small as much as possible. This constraint smooth's the geometric arrangements estimated for each device and the user at different moments when those arrangements can be measured after successful feature point extraction, as well as extrapolates the arrangements at the moments when those arrangements cannot be measured due to the failure in facial feature point extraction by simply duplicating the arrangements estimated at previous moments.

Let us represent E_i at any moment τ of the user's operation of multiple devices by $E_i(\tau)$. When the facial feature points are successfully extracted from the image obtained by the camera on the i-th device, $E_i(\tau)$ can be directly measured by the procedure described in 2.1, but includes a certain amount of error. When those facial feature points cannot be extracted from the camera image, $E_i(\tau)$ cannot be measured. Thus, we estimate the correct value of $E_i(\tau)$, regardless of the possibility of its measurement. The estimated value of $E_i(\tau)$ is denoted by $\hat{E}_i(\tau)$. In order to make $\hat{E}_i(\tau)$ coincide with $E_i(\tau)$ when it is measure from the camera image, and to make $\hat{E}_i(\tau)$ close to $\hat{E}_i(\tau-1)$ at the previous moment $\tau-1$, $\hat{E}_i(\tau)$ is determined so that the following function $\mathcal{E}$ is minimized at each possible moment τ:

$$\mathcal{E}\left(\hat{E}_i(\tau)\right) = \sum_{\tau}\left\{f_i(\tau)\left\|\hat{E}_i(\tau) - E_i(\tau)\right\|_F^2 + \left\|\hat{E}_i(\tau) - \hat{E}_i(\tau-1)\right\|_F^2\right\} \tag{4}$$

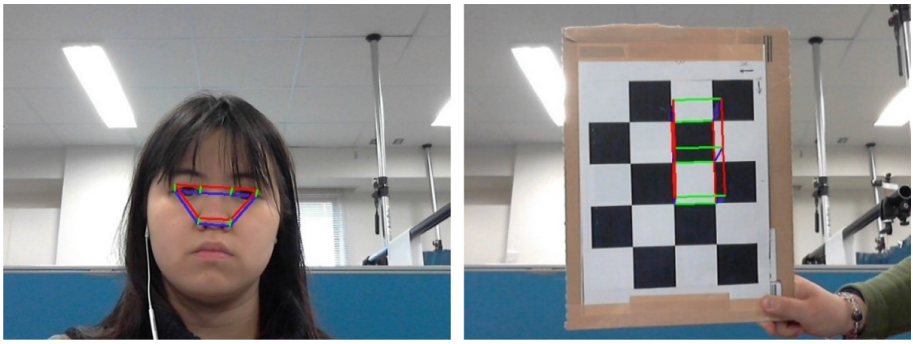
Where $f_i(\tau)$ is the variable that takes the value 1 at the moment τ when facial feature point extraction is successful for the camera image of the i-th device, and becomes zero otherwise. This function evaluates the difference of $\hat{E}_i(\tau)$ from $E_i(\tau)$ and $\hat{E}_i(\tau-1)$ using Frobenius norm denoted by $\|.\|_F$.

3 Experimental Results

Comparing Results from Facial Feature Points and Corners of a Grid. We have implemented the procedure described above by employing OKAO Vision of OMRON Corporation for facial image processing and OpenCV for other image processing including camera calibration.

We first compared the results of estimating the geometric arrangement between a camera and the user's face from facial feature points and markers used by the traditional camera calibration method. We employed six facial feature points, which include the inner corners and the outer corners of the eyes and the wings of the nose because they are comparatively stable for extraction by facial image processing and distributed widely over the face. We specified the positions of these facial feature points on the face-centered coordinate system based on the data of the standard human face [10], neglecting the personal difference. For the markers used for the traditional camera calibration method for comparison, we employed six corners around the center of a grid pattern on a checkerboard.

The results are shown in Fig. 2. Virtual 3D objects are drawn on the face and the checkerboard using the estimated arrangement. In spite that the actual positions of the facial feature points in the face-centered coordinate system are neither the same as the positions specified from the standard human face nor located on a flat plane such as a chessboard, the error in the appearance of the virtual object drawn based on the arrangement estimated from those facial feature points is comparable to the result from the corners of the grid patterns. This result shows that the facial feature points are sufficiently available for estimating the arrangement of an information device and its user's face.



(a) Result from the six facial feature points. (b) Results from six corers of a grid.

Fig. 2. Comparison between the results of camera calibration using the facial feature points and corners of a grid pattern.

Estimating the Arrangement of Two Devices and the User's Face. For evaluating the error in estimating the geometric arrangement of the information devices and the user's face by our method, we compared the estimated arrangement of the devices with their actual arrangement observed by a camera at the viewpoint of the user. In the experiment, two tablet PCs were gazed by the same user. At the top of each PC, we installed a web camera. The user wore eyeglasses equipped with a camera at their bridge as shown in Fig. 3, in order to obtain the images of the devices from the viewpoint of the user. The thick black frame of the eyeglasses was covered by a fabric tape in a skin color in case facial feature point extraction was interfered by the frame.

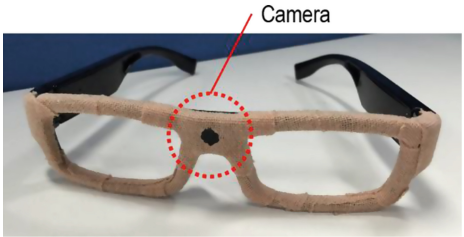
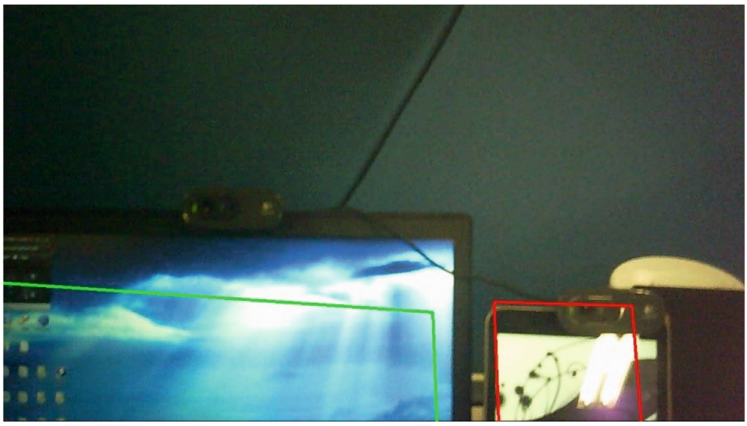


Fig. 3. Eyeglasses with a camera at their bridge

Figures 4(a) and (b) are the images obtained by the cameras on the two devices, where the points in each image are the facial feature points extracted by facial image processing. Figure 4(c) illustrates the result of drawing the appearances of the devices from the viewpoint of the user based on the geometric arrangement estimated by our method on the image observed by the camera on the eyeglasses. The difference of the estimated arrangement from the actual arrangement seems to be acceptable for the purpose of coordinating the contents to be displayed on the screens of the devices.

(a) User image by the left camera.

(b) User image by the right camera.

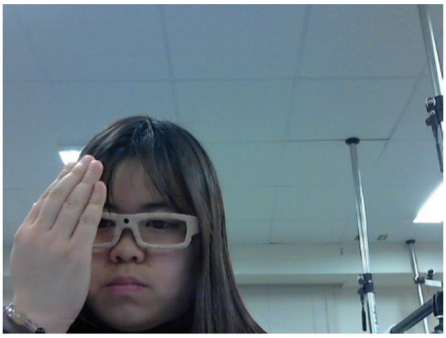


(c) Actual image taken by the camera on the eyeglass superimposed by the estimated appearances of the devices.

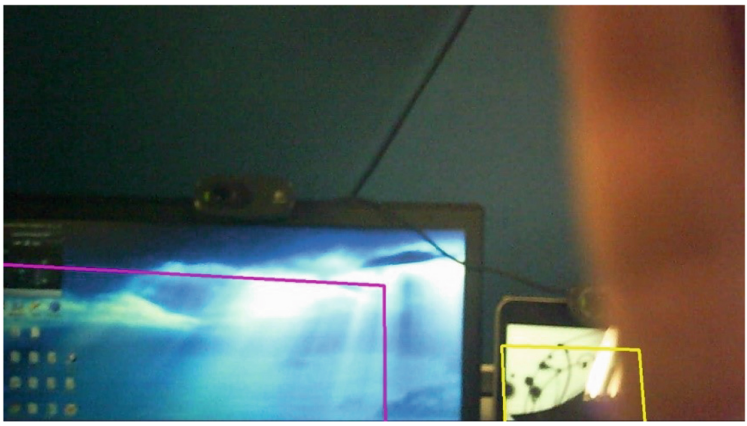
Fig. 4. Resultant images without failures of facial feature point extraction occlusion

We also estimated the arrangement of the same devices in the situation where the facial feature points fail to be extracted. Figures 5(a) and (b) are the images obtained by the cameras on the two devices in the situation. Since the user places a hand on the face, no facial feature point is extracted. Figure 4(c) illustrates the result of drawing the appearances of the devices from the viewpoint of the user based on the geometric arrangement estimated by our method in this situation. By extrapolation of the geometric arrangement at this moment from that estimated at the previous moment, the appearances of the devices are still be able to be obtained. Although the geometric arrangement of the devices and the user's face are slightly changed due to the motion of the user for placing the hand on the face, the amount of the error between the estimated appearance and the actual one is similar to that in the result of Fig. 3(c). This result shows that the extrapolation of the arrangement in our method is effective when the motion of the user is small. However, we need more sophisticated method for the extrapolation to cope with more various situations with the failure of facial feature point extraction.

(a) User image by the left camera.



(b) User image by the right camera.



(c) Actual image taken by the camera on the eyeglass superimposed by the estimated appearances of the devices.

Fig. 5. Resultant images with occlusion of the face

4 Conclusions

We proposed to measure the 3D geometric arrangement of multiple portable information devices using the facial images of the user captured by the cameras on the devices. In our method, facial feature points are extracted from the images taken by the camera on the devices and the geometric arrangement of each device and the user's face by using the camera calibration technique. To cope with the errors of estimating the geometric arrangement from facial feature points as well as the failure in extracting those facial feature points, we introduce the process of smoothing and extrapolation for the geometric arrangements estimated at different moments. From some experimental results, we confirmed that our method can estimate the arrangement of the devices and the user's face with the amount of error acceptable for coordination of contents to be displayed on the screens of those devices even at the moment when the facial feature points fail to be extracted due to occlusions as far as the change in the arrangement is small.

For one of the future steps, we need to introduce more sophisticated methods for extrapolating the geometric arrangement of the devices and the user's face to cope with more various and large amounts of changes in the arrangement by analyzing the patterns of the change during operation of multiple information devices by users. It is also useful to employ the data of the orientations, accelerations and so on of the devices obtained by various kinds of sensors usually installed in recent information devices to further reduce the error in estimating of the geometric arrangement.

References

1. Yatani, K., Tamura, K., Hiroki, K., Sugimoto, M., Hashizume, H.: Toss-it: intuitive information transfer techniques for mobile devices using toss and swing actions. IEICE Trans. Inf. Syst. **E89-D**(1), 150–157 (2006)
2. Dippon, A., Widermann, N., Klinker, G.: Seamless integration of mobile devices into interactive surface environments. In: ACM International Conference on Interactive Tabletops and Surfaces (ITS 2012), pp. 331–334 (2012)
3. Schmidt, D., Seifert, J., Rukzio, E., Gellersen, H.: A cross-device interaction style for mobiles and surfaces. In: Designing Inteactive Systems Conferece (DIS 2012), pp. 318–327 (2012)
4. Seifert, J., Dobbelstein, D., Schmidt, D., Holleis, P., Rukzio, E.: From the private into the public: privacy-respecting mobile interaction techniques for sharing data on surfaces. Pers. Ubiquit. Comput. **18**(4), 1013–1026 (2013)
5. Hahne, J., Schild, J., Elstner, S., Alexa, M.: Multi-touch focus+context sketch-based interaction. In: EUROGRAPHICS Symposium on Sketch-Based Interfaces and Modeling, pp. 77–83 (2009)
6. Baur, D., Boring, S., Feinter, S.: Virtual projection: exploring optical projection as a metaphor for multi-device interaction. In: ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2012), pp. 1693–1702 (2012)
7. Hinckley, K.: Synchronous gestures for multiple persons and computers. In: ACM Symposium on User Interface Software and Technology (UIST 2003), pp. 149–158 (2003)
8. Schneider, D., Rukzio, J.J.E.: MobIES: extending mobile interfaces using external screens. In: International Conference on Mobile and Ubiquitous Multimedia (MUM 2012), pp. 59:1–59:2 (2012)
9. https://www.dh.aist.go.jp/database/91-92/data/list.html