Detection of a Hand Holding a Cellular Phone Using Multiple Image Features

Hiroto Nagayoshi[1], Takashi Watanabe[1], Tatsuhiko Kagehiro[1], Hisao Ogata[2], Tsukasa Yasue[2], and Hiroshi Sako[1]

[1] Central Research Laboratory, Hitachi Ltd., Tokyo, Japan
{hiroto.nagayoshi.wy,takashi.watanabe.dh,tatsuhiko.kagehiro.tx, hiroshi.sako.ug}hitachi.com
[2] Hitachi-Omron Terminal Solutions, Corp., Aichi, Japan
{hisao_ogata,tsukasa_yasue}@hitachi-omron-ts.com

Abstract. Detection of a hand holding a cellular phone was developed to recognize whether someone is using a cellular phone while operating an automated teller machine (ATM). The purpose is to prevent money transfer fraud. Since a victim is told a bogus reason to transfer money and how to operate the machine through a cellular phone, detecting a working cellular phone is necessary.

However, cellular phone detection was not realistic due to variable colors and shapes. We assumed that a user's hand beside the face was holding a cellular phone and decided to detect it.

The proposed method utilizes color, shape, and motion. Color and motion were used to compare the input to the face. Shape was used to compare the input to the standard hand pattern. The experimental result was a detection rate of 90.0% and a false detection rate of 3.2%, where 7,324 and 20,708 images were used respectively.

Keywords: hand detection, multiple features, color, shape, motion, HOG, optical flow, face detection.

1 Introduction

"Money transfer fraud" that targets ATM users is a social problem in Japan. The number of cases increased substantially in 2005. The number of cases is now more than 10,000, and the financial loss is more than 25 billion yen per year.

In that type of fraud, criminals use a bogus reason to transfer money. For instance, they tell the victims that they can get a tax refund and ask them to go to an ATM. Then, the criminals show the victims how to operate the ATM through a cellular phone. Usually, the victims are those who are not familiar with that kind of electronic device. Therefore, the victims obey the instructions and operate the ATM without knowing that they are sending money to the criminals.

As described above, one of the specific features is that the victim uses a cellular phone. A system that can detect a working cellular phone and warn the victims is needed to prevent such a crime.

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 151–160, 2009.

We developed a system for that purpose using image recognition techniques. Other methods, such as detecting radio waves from the phone and voice recognition are also effective. However, the fact that many ATMs now have cameras makes it reasonable to use an image recognition based system.

One of the difficulties is that the colors and shapes of cellular phones are variable. In addition to that, the cellular phone is concealed by the hand. To avoid these problems, we focused on detecting the hand holding the cellular phone, rather than detecting the cellular phone itself. The variety of colors and shapes of hands is less than that of cellular phones.

One of the major applications of hand detection techniques is gesture recognition [1]. In that field, the environment sometimes can be modified to be appropriate for detection. For instance, a plain white or black background can be used. However, that is frequently impossible for a place such as an ATM booth. Therefore, a robust method against the variety of backgrounds is required.

Some robust methods were proposed which use color [2], hand contours [3,4], and spacial frequency of intensity [5]. Usually, shape detection is more robust against lighting variation. In particular, it gives the best performance when it captures the image of the fingers. However, in our application, the hand was in the shape of a fist, and sometimes fingers were concealed. Therefore, using complementary multiple images of different features was necessary.

This paper firstly describes the method to detect a hand utilizing multiple features such as color, shape, and motion. Then, the experimental results using the images that simulate real scenes are presented to show the validity of the proposed method.

2 Detection of a Hand Holding a Cellular Phone

2.1 Basic Idea

Fig.1 shows a person using a cellular phone. As described above, it is difficult to detect a cellular phone because of concealment by the hand and the variety of shapes and colors. Therefore, we decided to detect the hand beside the face, which can be recognized as a person using a cellular phone. Although there are other reasons for this hand position, such as scratching a cheek, it was possible to distinguish this because the duration was short.

The flow is shown in Fig.2. First, face detection is executed. We applied a face detector by Sochman et al., that was an improved version of [6]. Their method is characterized by WaldBoost classifier that gives optimal time and error rate trade-off. Then the feature value using color was calculated, and the positions of the hand candidate regions were revised using that value. The other feature values using shape and motion were calculated inside those candidate regions. The final decision was made using intermediate decisions based on each feature value.

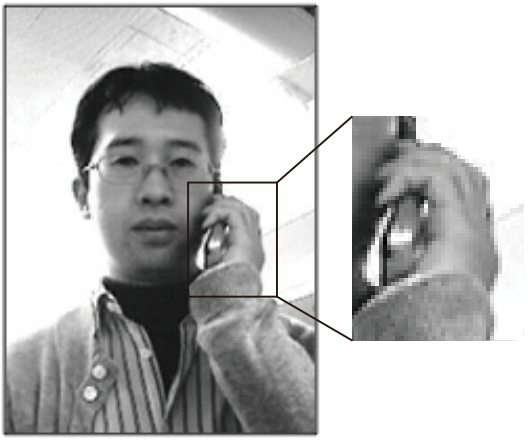

Fig. 1. A person using a cellular phone

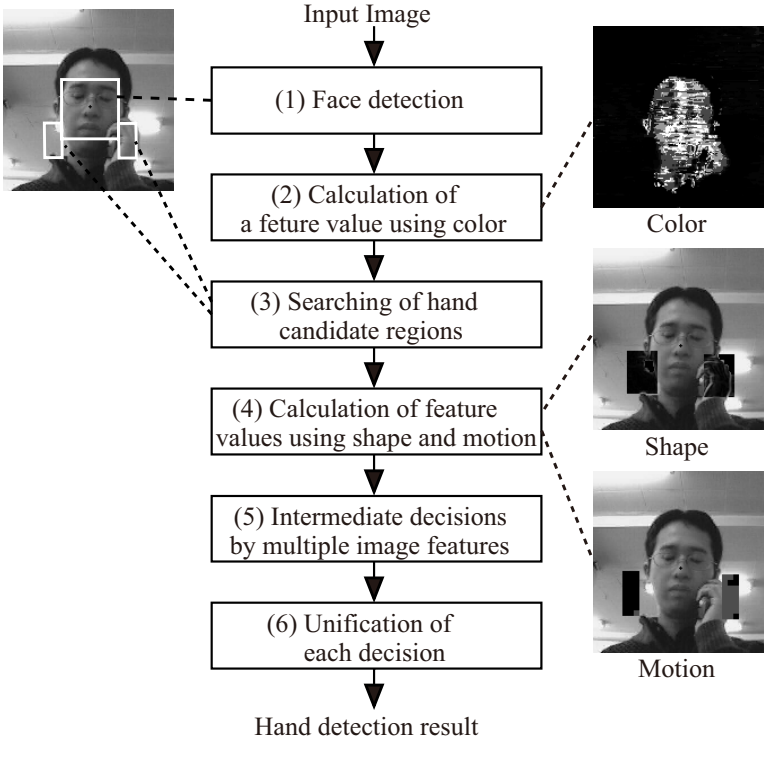

Fig. 2. Flow of the system

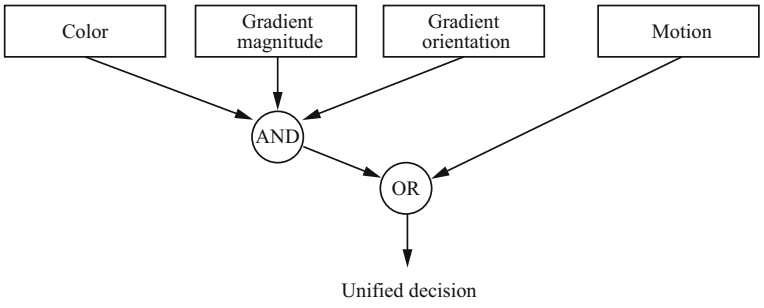

Fig. 3. Unification of decision for each feature

2.2 Image Features

To examine effective image features, we summarized the characteristics of a hand from the point of view of image processing.

1. The color of a hand is similar to that of the face [3].
2. The complexity of hand shapes will cause many thick edges.
3. Fingers not concealed make parallel edges. This means that the edges of the specific orientation are dominant.
4. Since the hand is beside the face, its motion is similar to that of the face.

For characteristic 1, similar color pixels to that of the face need to be detected. For 2 and 3, it was important to choose a method that could detect shapes. We decided to use a gradient based method, which is reported to be suitable for object detection [7]. One factor was gradient magnitude, and the other was gradient orientation. For 4, an optical flow was used to detect an object that had a similar motion to that of the face.

Using each feature, intermediate decisions showed whether a hand was detected. After that, those results were unified as shown in Fig.3. The results for color and shape were unified by the AND operator to reduce false detection. On the other hand, motion was independent of them. For instance, even if the results for color and shape were "not detected", an objects that moves the same way as the face should be recognised as a hand. This was why the result for motion was unified using the OR operator.

2.3 Color

The color similarity to the face was one of the important features. This feature had the advantage that it was robust to the variation in color of the environment because it used the face color as the standard. Even if the environment had changed, the hand and the face would still be very similar in color. However, the intensity between the face and the hand may be different because of shading. To avoid this problem, only hue and chroma components in the HSI color system were used.

The details of the method are as follows. First, the probabilistic distribution of the face color was calculated. This was implemented by hue-chroma histogram. Then, we calculated the likelihood that showed how similar each pixel was to that of the face (Fig.2(2)). The position of the candidate region was revised to where it had the maximum likelihood (Fig.2(3)), according to following the equation:

$$y_r = \arg\max_{y_s} \sum_{y=y_0+y_s}^{y_0+y_s+height} \sum_{x=x_0}^{x_0+width} L(x,y) \tag{1}$$

$L(x,y)$: Likelihood that shows how similar the pixel was to that of the face at position(x,y).

x_0, y_0 : The top-left position of a hand candidate region, which was determined by the size and the position of the detected face.

width, height : Width and height of candidate regions - they are proportional to those of the detected face.

Once we had hand candidate regions on both sides of a face, they were utilized in calculating feature values of shape and motion too.

To get the intermediate decision whether a hand was placed beside the face at this step (Fig.2(5)), the number of pixels whose likelihood is larger than a threshold was calculated first. Then, the decision "a hand is detected" was reached when the number exceeded a threshold.

2.4 Shape

One of the most effective methods to evaluate shape is HOG descriptors [7]. We used a more simple method for stability and computational cost. When using HOG descriptors, the target region is divided into many cells; here there was only one cell. As a hand holding a cellular phone is considered to be a solid object, one cell was enough.

The purpose of the shape feature was to detect the target that had thick edges and a specific bias of edge orientations. For the former, we simply calculated the average gradient magnitudes. For the latter, the orientation of each gradient was accumulated in a histogram weighted by its magnitude. The histogram had 8 bins for 8 orientations. Then, the feature vector of 8 dimensions was evaluated using Mahalanobis distance.

Two values were calculated. One was the average magnitude of the gradients. The other was a Mahalanobis distance that expressed the dissimilarity from the standard hand pattern. Intermediate decisions (Fig.2(5)) were made by applying the thresholding to each of them.

2.5 Motion

The hand that holds a cellular phone will be beside the face. This means that the face and the hand usually moved in the same direction. Therefore, detecting an object whose motion was similar to that of the face is the purpose of this section.

We applied the Lucas-Kanade method [8] to calculate the motion because its low computational cost was suitable for real time processing. When a pixel whose intensity value was $I(\boldsymbol{x}, t)$, was moving with velocity expressed as motion vector $\boldsymbol{v}$, the constraint equation is given as follows:

$$\nabla I(\boldsymbol{x}, t) \cdot \boldsymbol{v} = -I_t(\boldsymbol{x}, t) \tag{2}$$

where $\boldsymbol{x}$ denotes the position and t denotes the time. The assumption that the motion vector $\boldsymbol{v}$ is constant in the local area is introduced, and the following equation is given by using the least square method.

$$\boldsymbol{v} = \arg\min_{\tilde{\boldsymbol{v}}} \sum_{\boldsymbol{x}} [\nabla I(\boldsymbol{x}, t) \cdot \tilde{\boldsymbol{v}} + I_t(\boldsymbol{x}, t)]^2 \tag{3}$$

Then, the motion vector is given as follows:

$$\boldsymbol{v} = (A^T A)^{-1} A^T \boldsymbol{b} \tag{4}$$

$$A = [\nabla I(\boldsymbol{x}_1), \ldots, \nabla I(\boldsymbol{x}_n)]^T \tag{5}$$

$$\boldsymbol{b} = [I_t(\boldsymbol{x}_1), \ldots, I_t(\boldsymbol{x}_n)]^T \tag{6}$$

First, the motion vector was calculated inside a face region where the motion vector could be considered to be constant. The hand detection occurred only if the magnitude of a face motion vector exceeded a threshold. This is because when the face stopped, the hand would also stop, and it was impossible to distinguish it from the background.

The candidate region was divided into small cells. The motion vector in each cell was compared to that of the face using a similarity measure defined as follows:

$$s = \frac{\boldsymbol{v}_h(i) \cdot \boldsymbol{v}_f}{|\boldsymbol{v}_h(i)||\boldsymbol{v}_f|} \tag{7}$$

When the number of cells that exceeded a threshold was more than another threshold, the intermediate result in Fig.2(5) would be “hand detected”.

3 Experiments

For an evaluation, 24 bit color images, which were compressed using motion JPEG format, were used. Those images were captured by USB cameras. The resolution was 320×240 pixels, and the frame rate was 15 frames per second. The number of frames are listed in Table 1. The A1 data is for evaluating the detection rate, and the B1-B3 data is for the false detection rate. The former data was collected when the person had a cellular phone while the latter data was collected when the person did not. We assumed that the camera was set at a position and an angle where it could capture the face of an ATM user head-on. In A1 and B1, the person looked straight at the camera and moved slightly. B2 and B3 include more variations. In B2, the person was standing in front of the camera but facing various directions. In B3, there were 3 people behind the

Table 1. Number of images for experiments

Database name	Situation	Number of frames	Number of frames where a face is detected
A1	facing the camera	10,789	7,324
B1	facing the camera	11,392	7,879
B2	facing various directions	3,954	1,874
B3	many people	14,805	10,955

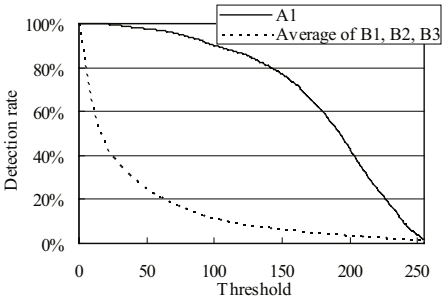

Fig. 4. Detection rate vs threshold (color)

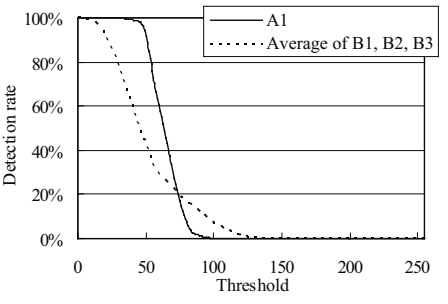

Fig. 5. Detection rate vs threshold (gradient magnitude)

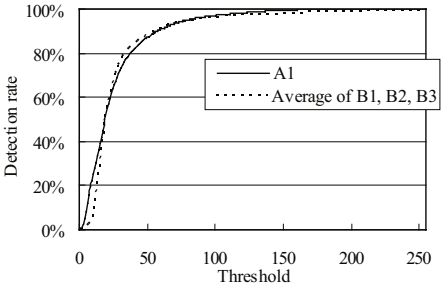

Fig. 6. Detection rate vs threshold (gradient orientation)

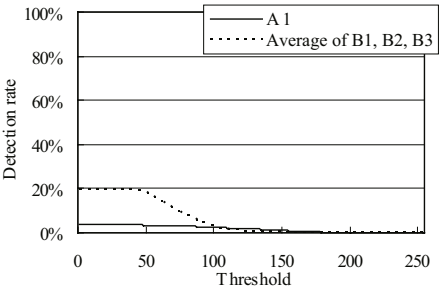

Fig. 7. Detection rate vs threshold (motion)

main person who stood in front of the camera. A1 and B1 consisted of 3 different sequences that captured 3 different people. B2 and B3 consisted of 5 sequences as well.

Using these data, the detection rate for each feature was evaluated. The results are shown in Fig.4-7. Solid lines indicate the detection rate evaluated using A1. Broken lines indicate the average rate of false detection evaluated using B1-B3. Comparing each result, the color feature gave the best result.

Table 2. Rate of correct and false detection using combinations of features

Data	C	C+Gm	C+Go	C+M	All features
A1	88.2%	87.3%	87.9%	88.4%	87.1%
average of B1, B2, B3	9.0%	4.3%	8.6%	10.0%	5.0%
B1	1.0%	0.2%	1.0%	1.3%	0.5%
B2	14.1%	3.5%	13.6%	16.3%	5.2%
B3	11.8%	9.2%	11.3%	12.4%	9.3%

*C: Color, Gm: Gradient magnitude, Go: Gradient orientation, M: Motion

Table 3. Rate of correct and false detection using multiple frames decisions

Data	C	C+Gm	C+Go	C+M	All features
A1	90.6%	90.0%	90.6%	90.8%	90.0%
average of B1, B2, B3	8.0%	3.1%	7.7%	8.5%	3.2%
B1	0.3%	0.1%	0.3%	0.4%	0.1%
B2	12.5%	0.7%	12.1%	13.6%	1.1%
B3	11.3%	8.7%	10.7%	11.6%	8.4%

*C: Color, Gm: Gradient magnitude, Go: Gradient orientation, M: Motion

Next, detection rates using combinations of features were evaluated. The thresholds were determined empirically. The results are shown in Table 2. The evaluated combinations of features are color, color and gradient magnitude, color and gradient orientation, color and motion, and all four features. The intermediate decisions for each feature were unified according to Fig.3.

From Table 2, gradient magnitude was the second best feature. It reduced false detections by 10.6 points in the B2 database while the correct detection rate went down by 0.9 points. Other features were less effective. Gradient orientation could reduce false detection in B2 and B4 by 0.5 points, while it reduced the correct detection rate by 0.3 points. Motion did not work well in this evaluation. However, motion can be used as a complementary feature for color and shape. For instance, when the user wore black gloves, it was difficult to detect the hand by color or shape. The reason for the worse performance may have been the instability of the motion vector. Especially, the small area size which was used to calculate the motion of hand would cause the unstable motion vector. More accurate techniques, such as introducing weight to each pixel according to its confidence are required.

The total accuracy was improved by using multiple frames. We introduced a decision rule that the final result was "detected" when a hand was detected in more than 4 of 10 successive frames. The result is shown in Table 3. The lowest average false detection was 3.1% using color and gradient magnitude while the correct detection rate was 90.0%. Using all four features gave a 90.0% detection rate and 3.2% false detection rate.

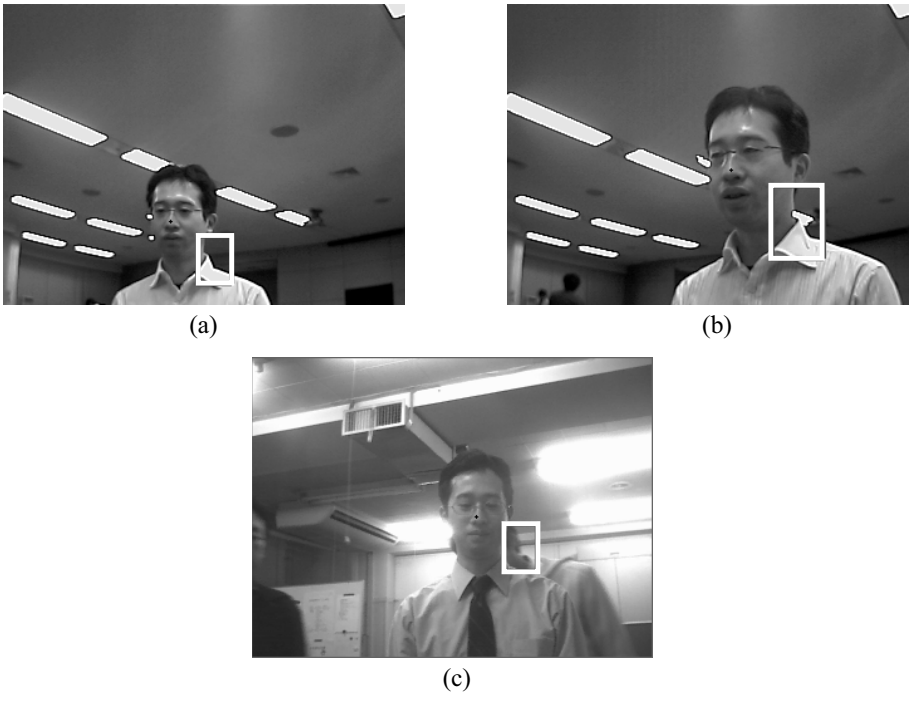

Fig. 8. Samples of false detection: (a),(b) samples from B2; (c) sample from B3

Some samples of false detection are shown in Fig.8. The falsely detected region included the cheek, neck, or other person's face. All of them are similar in color to that of the user's face. To reduce these false detections, shape and motion features need to be improved.

The computational time was fast enough for real time processing, 11.7 ms for each frame. The specification of the PC were Core 2 Duo 1.8 GHz, 3 GB RAM, where only one processor was used.

4 Conclusion

Detection of a hand that holds a cellular phone was developed to protect the ATM user from money transfer fraud. The reason for this is the victims usually indicated how to operate the ATM to a criminal through a cellular phone.

The detection method was based on image recognition techniques. The reason we focused on the detection of the hand is that the detection of cellular phones is very difficult due to the variety of colors and shapes.

Multiple features such as color, shape, and motion were used complementarily. To evaluate the detection rate, 7,324 images where a person faces the camera

were used. To evaluate the false detection rate, 20,708 images where a person faced various directions were used. The results were a 90.0% correct detection rate and a 3.2% false detection rate.

One of our future tasks is to reduce false detection. The major reason was that the color of the cheek, neck, and other person's face is similar to that of the face. To resolve this problem, improving detection by shape will be effective. Another future task concerns the database. The proposed method was evaluated using limited databases. Preparing more databases from real situations is necessary.

References

1. Utsumi, A., Tetsutani, N., Igi, S.: Hand detection and tracking using pixel value distribution model for multiple-camera-based gesture interactions. In: Proceedings of IEEE Workshop on Knowledge Media Networking, pp. 31–36 (2002)
2. Girondel, V., Bonnaud, L., Caplier, A.: Hands detection and tracking for interactive multimedia applications. In: Proceedings of International Conference on Computer Vision and Graphics, pp. 282–287 (2002)
3. Ong, E.-J., Bowden, R.: A boosted classifier tree for hand shape detection. In: Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2004), pp. 889–894 (2004)
4. Caglar, M.B., Lobo, N.: Open hand detection in a cluttered single image using finger primitives. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2006), p. 148 (2006)
5. Kolsch, M., Turk, M.: Robust hand detection. In: Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2004), pp. 614–619 (2004)
6. Sochman, J., Matas, J.: Waldboost-learning for time constrained sequential detection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2 (2005)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 886–893 (2005)
8. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of 7th International Joint Conference on Artificial Intelligence, vol. 81, pp. 674–679 (1981)