Estimating the Position of a Student's Face from Classroom Video in a Low-Quality Environment

Haruya Murasaki, Akira Suganuma

National Institute of Technology, Ariake College,
150, Higashihagio-machi, Ohmuta-shi, Kumamoto-ken 836-8585, Japan

Corresponding Author: suga@ariake-nct.ac.jp

Abstract

Currently, we are developing a system to take pictures of a class and estimate the atmosphere of the class from the images. We are developing a system to determine whether the students are taking the class seriously or not. In order to determine if a participant is taking the class seriously, the position of the participant's face needs to be determined from a low-quality image. To achieve this, we used two special integrals and a combination of them. We also developed the system with perspective weighting and exclusion of non-human areas in mind. As a result, the system was able to output 80% of the expected value correctly when the appropriate photographic environment was set up. In addition, non-skin areas such as the floor and coat are no longer considered to be faces. To improve the accuracy further, it is necessary to narrow down the area where the face can exist and to take measures against wearing a mask.

Keywords: Low quality image, Two special Integrals

1. Introduction

We are developing a system to detect the student's attitude by image processing. In order to make the system affordable, it is necessary to determine whether a student is sleeping or not and so on, by capturing the position of the student's face in a low-quality classroom scene. We have proposed a system detecting dozing students⁽¹⁾. However, the user (the teacher) needs to input the position of the student's face in this system. The purpose of this research is to incorporate a function to automatically detect the position of a student's face into this system.

Face detection is also implemented as a library in

Fig. 1. Image of sample data for develop.

OpenCV⁽²⁾. However, this method targets the case where the front of the face is reflected in the image. Considering the behavior of students in class, some participants will take the note frequently. In that situation, their faces face down and do not appear in the image. It is, furthermore, difficult to detect facial features such as eyes in a low-quality environment. Although it is possible to detect flesh color pixels, they do not always be flesh color. To solve this problem, we made two special integrals. We call them "Existence Integral" and "Continuous Integral." We have developed a system detecting students' face with them.

2. Solution

2.1 Existence Integral

It is difficult to determine if a person's skin is present simply by the number of times the skin tone appears. The most of participant often look down to take notes. It is not

(a) Output of Existence Integral ($N = 8, a = 8$).

(b) Output of Existence Integral ($N = 2, a = 4$).

(c) Image subtracted Figure b from Figure a.

Fig. 2. Result of Existence Integral.

uncommon for them to look at the blackboard for just a moment and then repeat the process of taking notes for a

while. In this case, the skin tone appears will be drastically reduced.

So, we need an evaluation method that ensures a certain amount of evaluation points for what the skin (we named the points *skin point*) even in such cases.

In this system we have implemented a method to calculate the *skin point* by the following formula. In this equation, $S_{x,y}(t)$ denotes the *skin point* in t -th frame at the coordinates of (x,y) . And $p_{x,y}(t)$ denotes the result estimated whether the pixel of the coordinate (x,y) is flesh color.

$$S_{x,y}(0) = 0 \quad (1)$$

$$S_{x,y}(t) = \begin{cases} S_{x,y}(t-1) + a & (b_{x,y}(t-1) > 0) \\ S_{x,y}(t-1) & (otherwise) \end{cases} \quad (2)$$

$$b_{x,y}(t) = \begin{cases} N & (p_{x,y} = true) \\ b_{x,y}(t-1) - 1 & (p_{x,y} = false) \end{cases} \quad (3)$$

In the above equation, we further use the element $b_{x,y}(t)$. It is appropriately described as a “grace of scoring.” The *skin point* will be increased while $b_{x,y}(t)$ is non-zero and it will be set to N if skin tone is detected, otherwise it will be reduced by one from its value in the previous frame.

This makes the *skin point* to the face of a participant who frequently looks down and one who does it less often almost equal. We call it “Existence Integral” in this study because it calculates the existence of flesh colored objects by performing an accumulation such as finding the temporal range of the appearance of flesh color.

Also, by varying the parameter a in equation 2 and the parameter N in equation 3, it was possible to make the target not a human face. By reducing the value of N to a very small value, only the pixels corresponding to a flesh colored still object will have a high *skin point*. Figure 2 shows the output for a group of test images for varying values of a and N , respectively.

In Figure 2a, the entire range of skin tones detected, while in Figure 2b, only stationary objects, such as floors and hanging coats, are highly rated by the Existence Integral. Considering the subtraction of Figure 2b from Figure 2a to further limit the area considered to be the skin from these two, we get Figure 2c.

2.2 Continuous Integral

The integration method described in the previous

section makes it possible to detect almost certainly the areas where flesh color is present. However, the output white area where the face is present is larger than expected with this method alone.

Therefore, we considered an integral to find the area in which a flesh-colored object stays longer. This allows us to find the center of motion of a flesh colored object, because as the object moves, the flesh color will disappear in some areas near the contour, but the flesh color will continue to appear near the center.

While the method guarantees the continuous addition of *skin points*, this method assumes that the amount of addition increases at an accelerated rate with continuous detection. We call it “Continuous Integral” in this study because it is an integration that emphasizes the output of continuous flesh color.

That is, if flesh color is detected continuously in a specific area, the area is presumed to be a place where some flesh colored object exists in principle. To implement this, we considered a calculation method expressed in the following formula.

$$S_{x,y}(0) = 0 \quad (4)$$

$$S_{x,y}(t) = S_{x,y}(t-1) + k_{x,y}(t-1) \quad (5)$$

$$k_{x,y}(t) = \begin{cases} k_{x,y}(t-1) + f \sqrt{S_{x,y}(t)} + c & (true) \\ 0 & (false) \end{cases} \quad (6)$$

In this formula, f denotes a multiple of the acceleration amount obtained from the current *skin point* and c denotes the minimum guaranteed acceleration amount. Based on Continuous Integral, the result of calculating *skin points* by changing the parameters is shown in Figure 3. It was confirmed that if the values of f and c are excessively large, it deviates from the goal of taking the core of the object. Therefore, the most appropriate data to use in this system is the pair of $f = 0.1$ and $c = 3$.

2.3 Extraction of facial parts

Two special integrals made it possible to estimate the areas where flesh colored objects exist. On this basis, it is necessary to evaluate which areas are the human face.

Based on the *skin points*, the areas where flesh colored objects are considered to exist are colored in Figure 4. In order to eliminate areas that are not considered to be skin,

(a) Output of *skin point* ($f = 0.1$, $c = 3$).

(b) Output of *skin point* ($f = 0.02$, $c = 10$).

Fig. 3. Result of Continuous Integral.

we considered basing it on the number of pixels in the area. Considering that the images are subject to perspective, and that the face area of the back row of participants is narrower, and that they are located at the top of the screen, you can see that the weight of the area varies with the coordinates.

Considered the weights of the fractional function distribution with the y-axis coordinates of the image as a variable, differing by a factor of 6 at the bottom and top of the screen. After taking this into account, the areas that do not meet the criteria, (*area pixel count* $\times$ *weight*) is less than 1300, are excluded in Figure 5.

Furthermore, the arm area would be output in this state, so we thought to exclude it. Since most of the face areas are equal in height and width, and areas such as arms are not, we excluded areas with extreme differences in height and width as not being faces. The result of this process is shown in Figure 6. At this point, only the face area can be

Fig. 4. Candidates where the objects exist.

Fig. 6. Candidate facial areas considering the aspect.

Fig. 5. Candidate facial areas considering the depth

Fig. 7. Final output from develop data.

extracted from the development data.

3. Result

3.1 For images used in development

The output result obtained as the final image is shown in Figure 7. Table 1 evaluates the acceptability of these results. The "verifiable (visually)" is the number of human faces that author can visually identify in the image. Among them, "expected in system" is the number of faces, excluding those that are absolutely impossible to detect, such as those near the edge of the image. The "right output" is the number of things that truly capture a face in the area the system estimates to be a face, and "not face output" is the number of outputs from non-face areas, such as the floor.

Although we were unable to detect the participant on the left of Figure 1, we accepted it this time because the

Table 1. Number of detections by data for development.

Items	Numbers
verifiable (visually)	11
expected in system	9
right output	8
not face output	0

number of frames in which he was facing forward were extremely a few.

3.2 For data taken previously

First, the system was used to judge the participants wearing summer clothes. This data was taken by the seniors of our lab. Figure 8 shows the operating conditions and Table 2 shows the output area. In this figure, we can see that the *skin point* is low on the human face and high on the contour. This problem was caused by student who raised his

(a) Image of sample data.

(b) Output of *skin point*.

(c) Output of face position.

Fig. 8. Results on previously taken data.

or her face long enough to be judged as a still object. Visual inspection of the data showed that the subject raised his

(a) Image of sample data.

(b) Output of *skin point*.

(c) Output of face position.

Fig. 9. Results from the optimal state data.

face for more than 200 consecutive frames.

We can also confirm that the two participants' faces in

Table 2. Number of detections by previous data.

Items	Numbers
verifiable (visually)	14
expected in system	13
right output	7
not face output	2

Table 3. Number of detections by optimal data.

Items	Numbers
verifiable (visually)	20
expected in system	16
right output	13
not face output	0

the left part of the image are judged as the same area. This is considered to be a flaw in the camera's shooting angle. Our system originally assumed that the faces of the last row of students are at the top of the image, but the target is in the center of the image. It was confirmed that the deviation from the expected input greatly reduced the accuracy of the output.

3.3 For data acquired in the correct environments

In the previous section we mentioned that the angle of capture of the image to be input is important. With this in mind, we set up the camera and acquired the data ourselves. Fig. 8 and Table 3 show the results of judging this data by our system.

The data here confirms that the output is generally as desired. In the case of the summer wear, the extra noise is often caused by the exposed arms and other parts of the body, but in the case of the winter wear, it is thought to be due to the fact that there was less of it. On the other hand, in the winter data, some participants wore masks, but this cannot be judged due to the current system. Countermeasures for the wearing of masks are an issue for the future.

4. Conclusion

In order to determine if a participant is taking the class seriously, the position of the participant's face needs to be determined from a low-quality image. To achieve this, we used two special integrals and a combination of them. We also developed the system with perspective weighting and

exclusion of non-human areas in mind.

It is now possible to estimate the position of the face without detecting facial features. Its detection rate is currently 80%. In addition, we have been able to significantly reduce the number of non-face areas that the system determines to be faces.

On the other hand, challenges include increasing accuracy and reviewing detection methods. In order to increase accuracy, we are attempting to limit the area where human faces are likely to be present by using the positional relationship of linear alignment of participants. The challenge is to devise a system to understand the positional relationship of the participants. The detection method should be reviewed to accommodate the wearing of masks, for example. Since the majority of the participants wear masks, in now, it is not appropriate to detect only flesh color. There are various types of masks, both in color and shape, and countermeasures for these are considered essential.

References

- (1) Taichi Kuriyama and Akira Suganuma : "Detection of Dozing Students from an Ordinary Classroom Scene with CV Technique," Proceedings of the 7th IIAE International Conference on Industrial Application Engineering, pp. 186-190, 2019.
- (2) Gray Bradski and Adrian Kaehler : "Learning OpenCV: Computer Vision with the OpenCV Library," O'Reilly Media, 2008.