Estimation of Facial Motions in Lectures from Degraded Video Considering Privacy

Takashi Ozeki
Fac. of Engineering
Fukuyama University
Fukuyama, 729-0292, JAPAN
Email:ozeki@fuip.fukuyama-u.ac.jp

Eiji Watanabe
Fac. of Intelligence and Informatics
Konan University
Kobe, 658-8501, JAPAN
Email:e_wata@konan-u.ac.jp

***Abstract*—Many researchers are trying to estimate the degree of concentration of students on lectures by analyzing the movement of their faces from the video taken of the attendance status. However, taking pictures with a video camera in places like classrooms where attendees are limited has a problem of privacy protection. So, it is impossible to take videos unless all students accept it. If we can analyze the movement of faces from degraded videos that cannot be identified individually, it will be easy for students to accept that they will be taken with a video camera. Therefore, in this paper, we made several low resolution videos from an original video taken of the attendance status and examined how difficult it is to estimate the movement of their faces for these degraded videos. According to some experiments, face detection became difficult gradually due to the degree of smoothing. However, it was showed that if the area of each face can be correctly detected in smoothed videos, we can sufficiently estimate the movement of faces by examining the number of skin color pixel in the area.**

Keywords—Privacy; Facial movement; Lecture; Education; Degraded video analysis

I. INTRODUCTION

Technology using IoT has been spreading in many fields. In particular, many security cameras are installed in public places such as railway stations and airports. They are used as a means to solve problems when incidents occur. For this purpose, personal identification is performed in video surveillance [1]. However, when videos of high resolution are taken in public places, there is a possibility of a video leakage accident that leads to infringement of privacy of individuals. Therefore, many methods of controlling access to high resolution videos [2, 3, 4] or automatically applying smoothing, mosaic and masking to the head and human body appearing in videos [5, 6] have been proposed. However, even if such video managements are performed, as long as high resolution original videos are present, the possibility of a video leakage accident still remains. Also, in places where specific people gather like classrooms, it is impossible to record videos without all participants' consent. For this reason, studies of personal identification using image analysis sometimes have been stagnant.

On the other hand, surveillance cameras are also used for human behavior analysis of a group in addition to the purpose of crime prevention. For example, by knowing the flow of people and the route of movement, it can be used to improve customer service by better arranging supermarket products, or used to guide safe evacuation in a disaster. Also, authors are attempting to estimate the degree of concentration of students on lectures from the movement of their faces in the classroom [7]. For this purpose, high resolution videos as used for personal identification are not necessary. Rather, low resolution videos are preferable for obtaining students' permission to take their attendance status.

In this paper, we will create videos with several stages of low resolution from an original video taken of the students' attendance status. Next, in comparison with the original video, we examine what happens to the estimation of the movement of faces in degraded videos.

II. ESTIMATION OF FACIAL MOTIONS

To create low resolution videos for an original video, we prepare eight types of smoothing filters. The sizes of the smoothing filters are S x S: S= 1, 3, 5, 7, 9, 11, 13 and 15. In particular, when S=1, it makes the original video itself. Then, as the method of detecting faces in those videos, the method used in the paper [7] is used. That is, first, the area including each student's face is detected using the Haar-like feature and its classifier [8, 9] and skin color information. Next, by examining the increase or decrease of the number of skin color pixel in the detected area, it is judged whether or not each student is looking forward in the classroom.

III. EXPERIMENTAL RESULTS

A video taken of students' status in a lecture used in our experiments is for 6 minutes. That is 360 seconds. The video rate is 30 fps, and the size of the image is 1280 × 720. The Haar-like classifier used for students' face detection is implemented with OpenCV [10]. It is used for all degraded videos. Also, face detection is done every 30 frames. Therefore, face detection is performed at 360 frames in the entire frame 10800. Rectangles A to F in Fig. 1 are including six faces detected in a certain frame of the original video of S=1. Students A, B, C and D are seated in the first row, student E is in the second row and student F is seated in the fourth row. Also, each size of the obtained rectangles including the face of students A, B, C, D, E and F is 55 x 55, 55 x 55, 55 x 55, 69 x 69, 45 x 45 and 27 x 27, respectively.

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

Table 1 shows the first sampling frame number of detecting these faces from the video degraded by each smoothing filter. Here, the notation "-" means not detected at all. In some low resolution videos like s=7, 9, 11, 13 and 15, the detection of small faces in the rear row such as students E and F becomes extremely difficult.

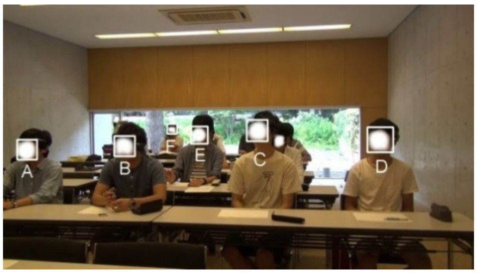

Fig. 1 Six facial rectangles detected from an original frame.

Table1: First sampling frame number detected for each degraded video.

S	1	3	5	7	9	11	13	15
A	1	1	1	1	2	2	4	4
B	1	1	1	1	1	1	5	41
C	1	1	1	1	1	2	2	10
D	1	1	1	1	1	1	1	1
E	2	5	7	14	16	16	-	48
F	2	2	7	231	261	334	-	-

Table 2 shows the number of times of each face detected in 360 sampling frames for each degraded video. As the smoothing filter becomes larger, it becomes more difficult to detect every faces gradually. Also, it has been happening that the number of detection of students C, D and E increase slightly in the case of S=3. The reason is that the each number of skin color pixel in these facial rectangles increases due to smoothing.

Table 2: Number of frames successfully detected for each degraded video.

S	1	3	5	7	9	11	13	15
A	213	201	181	169	156	118	39	1
B	150	143	124	96	51	43	25	2
C	200	201	190	175	152	118	86	11
D	317	308	297	265	235	222	180	116
E	179	199	181	132	44	7	-	2
F	62	15	13	2	2	1	-	-

In the following, we will analysis how much influence the degradation of the video by each smoothing filter affects the change in the skin color pixel number for six students.

A. Student A

Fig. 2 shows graphs of the change in skin color pixel number of student A. Even with the smoothing video of S=15, there is not much change as compared with the original video of S=1. It is possible to sufficiently estimate the motion of the face up and down from the degraded video. Fig. 3 shows magnified faces of student A in the case of S=7, 13 and 15. By visual inspection, personal identification is gradually difficult.

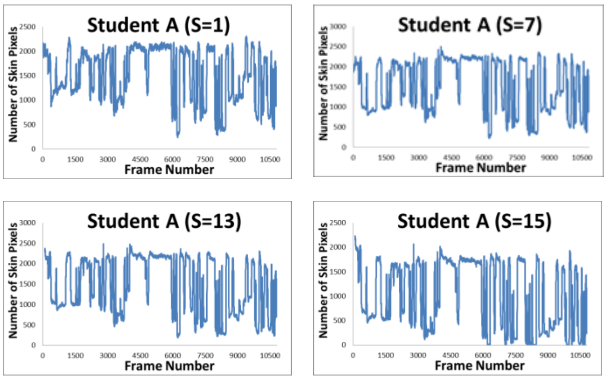

Fig. 2 Change in skin color pixel of student A for each smoothing filter.

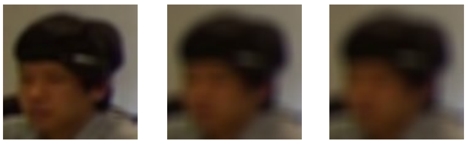

Fig. 3 Faces of student A for S = 7, 13 and 15.

B. Student B

Fig. 4 shows graphs of the change in the skin color pixel number of student B. Even with smoothing with S=7, there is not much change in the graph. However, in the smoothed video with S=15, it differs greatly from the case of the original video of S=1. Fig. 5 shows enlarged faces of student B in the case of S=7, 13 and 15. Individual identification also becomes difficult gradually.

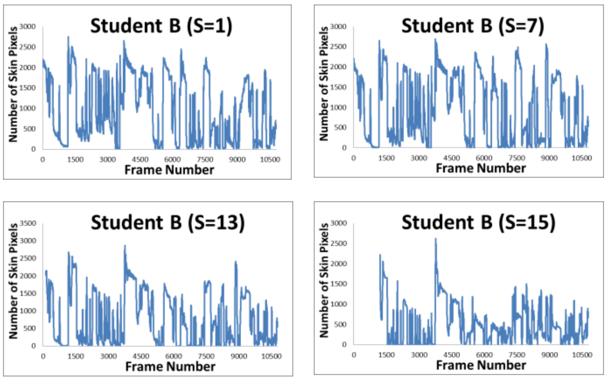

Fig. 4 Change in skin color pixel of student B for each smoothing filter.

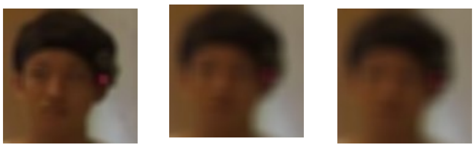

Fig. 5 Faces of student B for S = 7, 13 and 15.

C. Student C

Fig. 6 shows graphs of the change in skin color pixel number of student C. Like student A, there is no big change even in the case of the video with S=15. Fig. 7 shows enlarged faces of student C in the case of S=7, 13 and 15. At S=13, it is difficult to visually identify individuals.

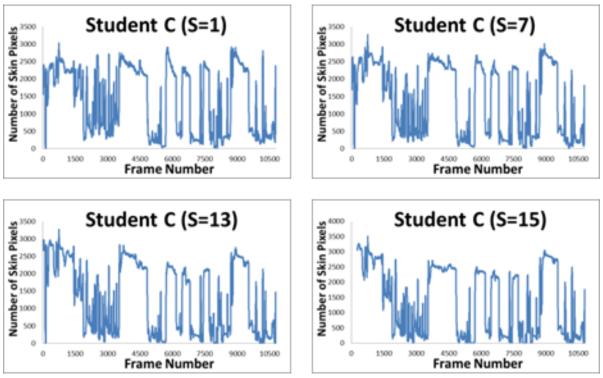

Fig. 6 Change in skin color pixel of student C for each smoothing filter.

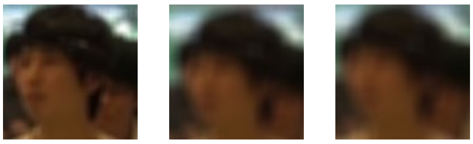

Fig. 7 Faces of student C for S = 7, 13 and 15.

D. *Student D*

Fig. 8 shows graphs of the change in skin color pixel number of student D. Like student A, there is no big change even in the case of the degraded video with S=15. Fig. 9 shows enlarged faces of student D in the case of S=7, 13 and 15. Again, at S=13, it becomes difficult to visually identify individuals.

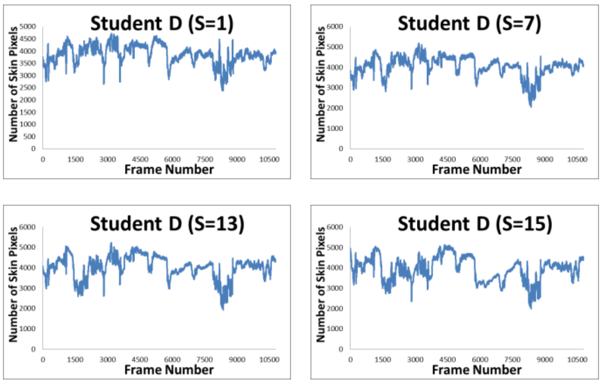

Fig. 8 Change in skin color pixel of student D for each smoothing filter.

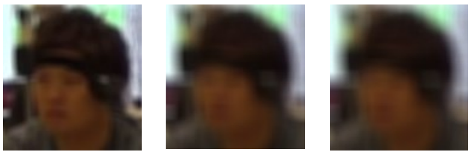

Fig. 9 Faces of student D for S = 7, 13 and 15.

E. *Student E*

Fig. 10 shows graphs of the change in the skin color pixel number of student E. Until the case of degraded video with S=11, no significant change is seen with the original video. However, at S=13 it is impossible to detect the face even once. So, it is impossible to know the skin color pixel number. However, at S=15, the face detection frame is only two, but the number of skin color pixel can be estimated from the way almost the same as the original video of S =1 in a long section.

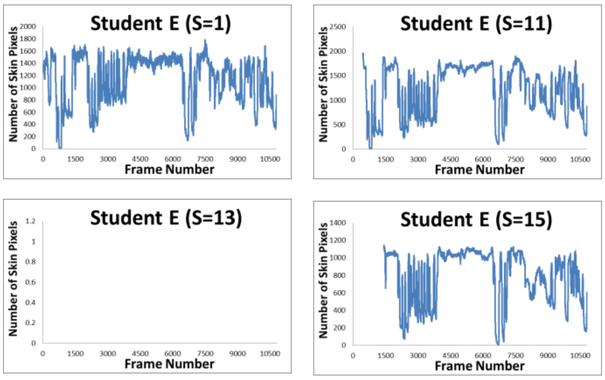

Fig. 10 Change in skin color pixel of student E for each smoothing filter.

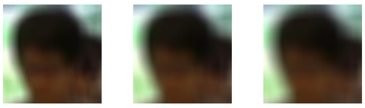

Fig. 11 Faces of student E for S = 11, 13 and 15.

F. *Student F*

Fig. 12 shows graphs of the change in the skin color pixel number of student F. Until the case of the degraded video with S=5, a large change is not seen compared with the original video of S=1 except for some sections. However, after S=7 and thereafter, face detection is almost impossible. So, it is impossible to get the number. Moreover, the graph of skin color pixel number is greatly different from the case of the original video of S=1. This reason comes from the fact that student F placed the chin on the hand. So, the hand was erroneously recognized as a part of the face. Hence, the position of the detected facial rectangle was misaligned.

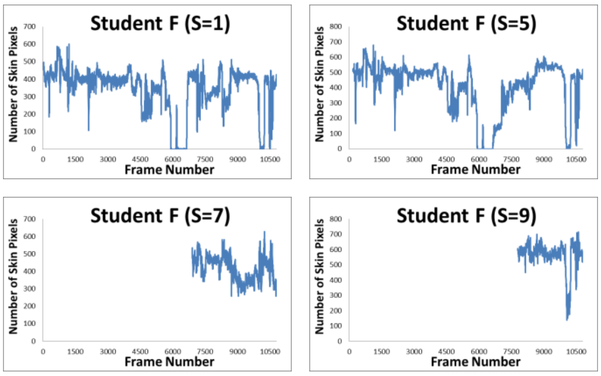

Fig. 12 Change in skin color pixel of student F for each smoothing filter.

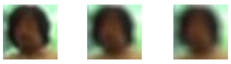

Fig. 13 Faces of student F' for S = 5, 7 and 9.

Fig. 13 shows enlarged faces of student F in the case of S=5, 7 and 9. Even at the video with s=7, it is difficult to visually identify individuals. Since the size of the face of the student sitting in the last row is less than half the one of the students in the front row, a small smoothing has a large effect and it makes the detection of the face more difficult.

G. When the position of students' faces are known

We consider the case where every positions of the detected faces in the case of the original video with S =1 are known. Then, from the smoothed video with S=15, we obtain the change in the number of skin color pixel for the above six students. Red lines in Fig. 14 show the results. They are almost the same as blue lines in the case of S=1.

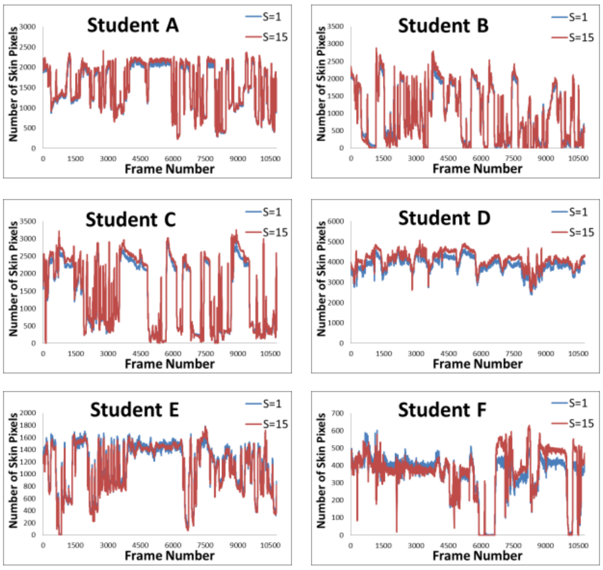

Fig14: Changes in the number of skin color pixel of six students in the smoothed video with S = 15 when the position of the faces are given.

Next, we examine how the estimation of face orientation changes between the original video of S=1 and the smoothed video with S=15. For this purpose, we make a histogram of skin color pixel number for each student and obtain the threshold by Otsu method [11]. Here, we determine the face in every frame is looking forward when the skin color number is more than the threshold. Table 3 shows the thresholds of each student and the error rates of face orientation. The error rate for each student is not extremely large except student D. This is due to the distribution of the histogram. We calculate the binary degree of each histogram [12]. The result is the binary degree in Table 3. When it is close to 0, the distribution has a unimodal property. Student D does not move his face too much. So, the histogram has a unimodal characteristic. Therefore, the error rate became large due to the change of the threshold.

Table 3: Thresholds of the histogram of skin color pixel number and error rates.

Threshold	S=1	S=15	Error Rate	Binary Degree
A	1462	1513	1.5%	0.81
B	1045	1079	2.3%	0.83
C	1341	1420	0.3%	0.90
D	3776	4018	7.6%	0.61
E	1044	1031	2.2%	0.79
F	230	254	3.5%	0.72

As a result, it became clear that we can sufficiently estimate the movement of the faces even from the video degraded with S=15 if we can correctly know the position of the faces.

IV. Conclusion

In this paper, we examined the possibility of the estimation of facial motions in videos degraded by some smoothing filter. By smoothing, it became difficult to obtain accurately skin color pixel number gradually. Especially, the tendency was higher for students with smaller faces seated in the back row. This reason is that the detection of the position of the face gradually becomes more difficult as the smoothing progresses in the case of smaller faces. However, if the correct position of the face of the student was given, there was no big difference in the number of skin color pixel obtained even from smoothed videos with a severe deterioration. Therefore, in such a case, it is possible to estimate the movement of faces even from degraded videos.

The future work is to establish a method to accurately detect the position of faces even from the smoothed video that cannot identify individuals

Acknowledgment

This work has been supported by JSPS KAKENHI Grant Number JP16K00499.

References

[1] M. Cristani, R. Raghavendra, A. D. Bue and V. Murino, “Human behavior analysis in video surveillance,” A Social Signal Processing perspective, Neurocomputing archive Volume 100, pp. 86-97, 2013.

[2] W. Zang, S. S. Cheung M. Chen, “Hiding Privacy Information in Video Surveillance System,” Proc. ICIP, pp. 868-871, 2005.

[3] T.Sekiguchi and H. Kato, “Proposal and Evaluation of Video-based Privacy Assuring System Based on the Relationship between Observers and Subjects,” IPSJ, Vol. 47, No. 8, pp. 2660-2668, 2006.

[4] K. Chinomi, G. Li, D. Nakshima, N. Nitta, Y. Ito and N, Babaguchi, “PriSurv: Privacy Protected Video Surveillance System,” IPSP CVIM, Vol. 1, No. 2, pp. 152-162, 2008.

[5] E.M. Newton, L. Sweeny, B. Malin, “Preserving Privacy by De-Identifying Face Images,” IEEE Trans. Knowledge and Data Engineering 17(2), pp. 232–243, 2005.

[6] A. Cavallaro, O. Steiger, T. Ebrahimi, “Semantic Video Analysis for Adaptive Content Delivery and Automatic Description,” IEEE Trans. Circuits and Systems Video Technology 15(10), pp. 1200–1209, 2005.

[7] T. Ozeki, E. Watanabe and T. Kohama, “A Measurement Method of Students' Facial Movements in Lectures Using a Haar-like Classifier,” Proc. of IWAIT 2017, in USB(4 pages),.Penang, Malaysia, 2017.

[8] P. Viola and M. J. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 511-518, 2001.

[9] R. Lienhart and J. Maydt, “An Extended Set of Haar-like Features for Rapid Object Detection,” IEEE International Conference on Image Processing, Vol. 1, pp. 900-903, 2002.

[10] G. Bradski and A. Kaebler, Learning OpenCV, O'REILLY, 2008.

[11] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Trans. Syst. Man Cybern., Vol. SMC-9, pp. 62-66, 1970.

[12] T. Ozeki and F. Kobayashi, “Restoration of Blurred Image Using Otsu's Method,” IIEEJ, Vol. 28, No. 5, pp. 651-660, 1999.