Head Pose Estimation and Movement Analysis for Speech Scene

Rinko Komiya†, Takeshi Saitoh†, Miharu Fuyuno‡, Yuko Yamashita††, Yoshitaka Nakajima‡
†Kyushu Institute of Technology
680–4 Kawazu, Iizuka, 820–8502, Japan.
Email: saitoh@ces.kyutech.ac.jp
‡Kyushu University
4–9–1 Shiobaru, Minamiku, Fukuoka, 815–8540, Japan.
††Shibaura Institute of Technology
3–9–14 Shibaura, Minato-ku, Tokyo, 108-8548, Japan.

***Abstract*—Public speaking is an essential skill for a large variety of professions and in everyday life. However, it is difficult to master the skills. This paper focuses on the automatic assessment of nonverbal facial behavior of public speaking and proposes simple and efficient method of head pose estimation and motion analysis. We collected nine speech scenes of the recitation contest in a Japanese high school, and applied the proposed method to evaluate the performance. As for the head pose estimation, our method was obtained acceptable accuracy for the speech scene. Proposed motion analysis method can be calculated frequencies and moving ranges of head motion. As the result, it was found that there is correlation between the moving range and eye contact score.**

***Keywords*-Head pose estimation; movement analysis; speech scene;**

I. Introduction

The ability to communicate in social and public environments is influence a person's career development, help build relationships, resolve conflict. Public speaking performances are not only characterized by the presentation of the content, but also by the presenters' nonverbal behavior, such as gestures, tone of voice, vocal variety, and facial expressions. Nonverbal communication expressed through behaviors is a key aspect of successful public speaking and interpersonal communication. However, public speaking skills are difficult to master and require extensive training. Moreover, the evaluation of public speaking is occasional and heavily relies on human rating. The automatic assessment of public speaking is expected for training. Recently, interactive virtual audience system for public speaking training are proposed [1], [2], [3], however, these systems are required several special devices, such as head-mount display (HMD), Microsoft Kinect, physiological sensors. The efficient training system of simple constitution is expected.

Chen et al. proposed an automated scoring model for public speaking performance using multimodal cues [4]. In [4], two types of public speaking tasks: informative and impromptu presentations were collected by using Kinect. They calculated Kinect features, head pose and eye gaze, facial expression, lexical features, speech features as the multimodal features. These features were fed to three regression methods of support vector machine, glmnet, and random forest. Ramanarayanan et al. also presented similar approach [5]. From the standpoint of development of automatic scoring system, these are useful, however, from the standpoint of teaching, it is still difficult to instruct how to improve the speaking. The system that improvement points in detail can be instructed is expected to the learner.

This research focuses primarily on the automatic assessment of nonverbal facial behavior of public speaking. Related researches require Kinect, however, our method requires standard video camera. This paper proposes simple and efficient methods of head pose estimation and motion analysis for speech scene.

The remainder of this paper is organized as follows: Our speech dataset is described in Section **II**. Section **III** and **IV** describe the proposed head pose estimation and motion analysis method, respectively. Experimental result is described in Section **V**. This paper concludes in Section **VI**.

II. Dataset

We collected a dataset which took an official English recitation contest at high school by using digital video camera. In the contest, nine contestants went for the recitation contest. Three types of script $\{$A, B, and C$\}$ were prepared and each speaker chose one and recited it. Script A is excerpt from "The Principal's Address to the Graduates" by Tsuda Umeko, B is excerpt from "Acceptance Speech for the Jerusalem Award" by Haruki Murakami, and C is excerpt from "The Little Prince" by Antoine de Saint-Exupery. The word numbers of scripts A, B, and C are 355, 362, and 328, respectively. All contestants are Japanese high school students, and their major is English. Each speaker spoke in front of audience and five contest raters.

This contest was held in a class room and the background of speaker was a blackboard as shown in Fig. 1. The video camera was fixed on a tripod. The size of image is 854×480 pixels and the frame rate is 29.97 fps. The detail information of the dataset is shown in Table I. In the table, N_f is a number of frames at each scene.

In the contest, each rater judges each speaker. There are three types of criterion for recitation contest: four English

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

Fig. 1. Target speech scene.

TABLE I
DATASET.

speaker	script	N_f[frames]	duration[sec]	eye contact score
JF-R-1	A	5061	168.9	6.8
JF-R-2	C	5431	181.2	6.0
JF-R-3	A	5631	187.9	6.0
JF-R-4	C	4681	156.2	8.8
JF-R-5	A	5291	176.5	7.4
JF-R-6	B	5091	169.9	9.2
JF-R-7	A	4776	159.4	7.8
JF-R-8	C	5511	183.9	6.4
JF-R-9	A	5141	171.5	7.0
ave.		5179	172.8	7.3

TABLE II
EVALUATION ITEMS.

Item	Full score (each judge)	Description
① Pronunciation	10	pronunciation
② Intonation	10	intonation
③ Rhythm	10	speech rhythm
④ Speech delivery	10	delivery / flow / pace
⑤ Volume	10	volume of voice
⑥ Gestures	10	gestures
⑦ Eye contact	10	eye contact
⑧ Emotion	10	emotion / energy / passion
⑨ Memorization	20	memorization of asignment

criterion, four attitude criterion, and memorization as shown in Table II. This research is focused on facial movement, and the left column of Table I shows average eye contact score by five raters.

III. HEAD POSE ESTIMATION

Human head pose has three degrees of freedom (DOF) which can be characterized by pitch, roll, and yaw angles. The target person to estimate head pose of this research is a public speaking person, and there is hardly movement of pitch during speech. Thus, this research focuses on roll and yaw angles. There are a number of research for head pose estimation using computer vision technique. In this field, there are eight conceptual approaches to estimate head pose: appearance template methods, detector array methods, nonlinear regression methods, manifold embedding methods, flexible models, geometric methods, tracking methods, and hybrid methods [6]. However, the target head angles of this

Fig. 2. Feature points (left: 42 points by AAM, right: with four points for head pose estimation).

paper are yaw and roll, we adopt fast and simple approach of geometric method in these approaches. Geometric methods attempt to find local features, such as the eyes, mouth, and nose tip, and determine pose from the relative configuration of these features [7], [8]. In this paper, we propose a triangle model based method for estimating yaw and roll angles.

A. Feature-point detection

In general, the geometric method of head pose estimation is required some facial feature locations, and the proposed triangle model based method is required the eyes and nose point. To detect these points, active appearance model (AAM) [9] is applied to the speech scene for detecting facial feature points. AAM is a local search method that combines a full shape model and texture variation learnt from a training set. In this paper, a facial model consisting of eyes, eyebrows, a nose, an external lip contour, and a face outline is build, as shown in the left-side of Fig. 2. The number of feature points in this model is 42.

B. Head pose estimation

After detecting feature points, we estimate a head pose using three points: left eye point P_L, right eye point P_R, and nose point P_N. Here, P_L is a gravity point of four left eye points, P_R is a gravity point of four right eye points, and P_N is a middle-bottom point of detected nose point. The reason not to use a gravity point of detected nose points for P_N is to ignore the height of the nose which is difference by an individual. By using these three points, we consider a triangle $P_L P_R P_N$ as shown in the right-side of Fig. 2.

Roll angle can be calculated by

$$Roll = \arctan\left(\Delta y / \Delta x\right)$$

where $\Delta y = P_{R,y} - P_{L,y}$ is the vertical distance of two eye points, and $\Delta x = P_{R,x} - P_{L,x}$ is the horizontal distance.

It is difficult to directly calculate yaw angle, and the proposed method firstly calculate a ratio r. We consider $P_L P_R$ is a triangle's base, and P_H is a perpendicular foot as shown in the right-side of Fig. 2. Two distances $d_1 = |P_L - P_R|$ and $d_2 = |P_L - P_H|$ are calculated and r is calculated by $r = d_2/d_1$. Next, yaw angle is calculated by $yaw = f(r)$,

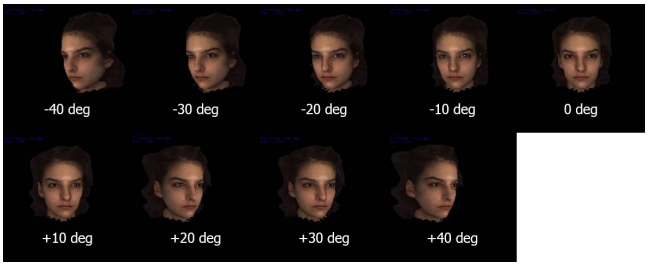

Fig. 3. Generated arbitary face images.

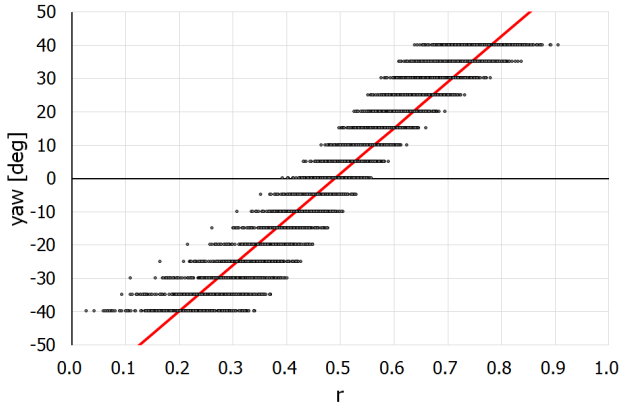

Fig. 4. $r - yaw$ scatter diagram for mapping function.

where f is a mapping function r to yaw. The detail of f is described in the next section.

During a speech, the speaker not only change expression and hand gesture, but also change the standing position. The standing position of the speaker is corresponded to the head position. Thus, the head position can be expect efficient information for analyzing speech scene. We already obtained facial feature points by applying AAM. Then, a gravity point (G_x, G_y) of all feature points is calculated as the head position.

From the above, four parameters $\{yaw, roll, G_x, G_y\}$ are calculated.

C. Mapping function

A mapping function f is outputted a yaw angle yaw as input a ratio r. Various relationships between r and yaw are required to make f. This is indicated that we need prepare various face images in which head pose at each face image is known. In this research, a high-resolution 3D dynamic facial expression database BU-4DFE [10] is used. This database contains 606 3D facial expression sequences captured from 101 subjects of various ethnic backgrounds. Each sequence has both 3D model, 2D texture, and annotated facial feature points. Since the facial data is a 3D, we can generate a face image of arbitrary head pose. Figure 3 shows generated face images in which yaw angle is changed from $-40°$ to $40°$, roll angle is set to $0°$, and pitch angle is set to $0°$.

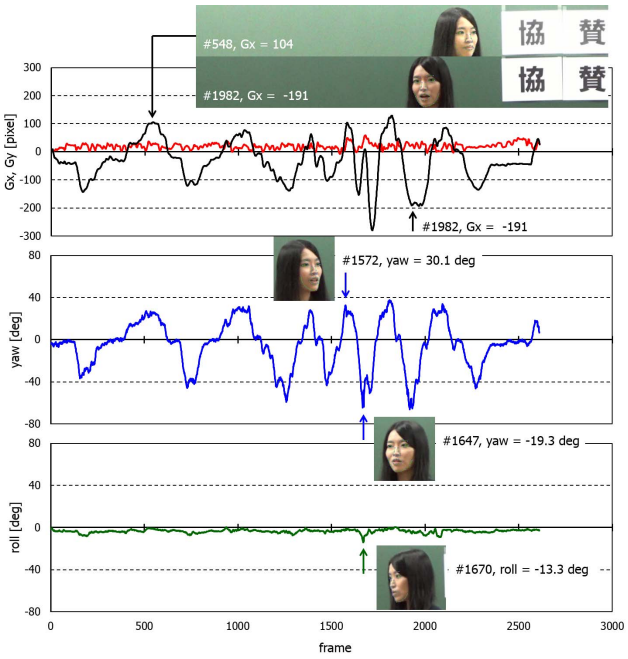

Fig. 5. Transition graphs of four parameters.

We generate 9,895 face images which are oriented to one of nine head poses $(0, \psi, 0)$, where the pitch and roll angles are set to 0, and the yaw angle ψ is set $\psi = \{0, \pm10, \pm20, \pm30, \pm40\}$ using BU-4DFE database. Next, a ratio r is calculated at each face image, and 9,895 pairs of r and yaw are collected. Then, a scatter diagram is made based on generated face images as shown in Fig. 4. In the graph, the horizontal axis is r and the vertical axis is yaw. Observing the graph, it is found that we can consider a linear approximation between r and yaw. We apply linear least-square method to obtain the approximate function f. A red line in Fig. 4 is obtained f.

IV. Motion Analysis

Figure 5 shows the transition of four parameters of yaw angle yaw, roll angle $roll$, and gravity point (G_x, G_y) during a speech. Based on these changing data, this paper defines two features: frequency and moving range. These features are corresponding to the eye contact of speaker.

It can be seen that the transition of parameter is one of the periodic signals. Then, Fourier transform is applied to the signal, and obtain a maximum frequency as the feature. Let x_i be the each parameter at frame i, and X_i is the Fourier transform. X_i is defined by

$$X_i = \Sigma_{i=0}^{N-1} x_i e^{-j2\pi\omega x_i/N},$$

where N is the number of frame. Next, maximum frequency m_X is calculated by

$$m_X = \max(X_i | 5 < i < 100) \times 29.97[\text{fps}] \times 60[\text{sec.}]/N[\text{frame}].$$

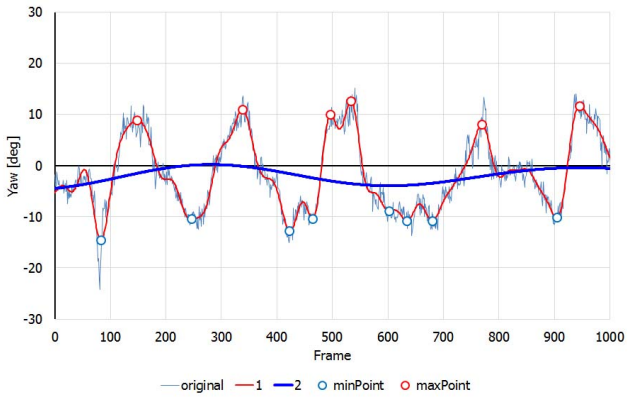

Fig. 6. Local maximum points and local minimum points detection.

As for yaw and $roll$, calculated m_X means a turning frequency of speaker's head per one minute.

It is difficult to measure how the speaker turns range by only the turning frequency. Next, we calculate the moving range. The concept of this value is a difference between average maximum head pose and minimum head pose. As for the yaw angle, the moving range R_m is defined by

$$R_m = \overline{x_{\max}} - \overline{x_{\min}},$$

where $\overline{x_{\max}}$ and $\overline{x_{\min}}$ are the average maximum of yaw and average minimum of yaw, respectively. The reason in consideration of the average is to reduce influence of the noise. The calculating process of the R_m is as follow: two smoothing signals of x_1 and x_2 are calculated by applying low pass filter. Here, we set the cut-off frequencies of x_1 and x_2 to 120Hz and 10Hz, respectively. x_1 is a signal for detecting local maximum points and local minimum points as shown in red signal of Fig. 6. x_2 is a signal for approximating the average value of the original signal x as shown in thick blue signal of Fig. 6. Local maximum points and local minimum points are detected on x_1, however, the point in a range $[x_2 - \alpha, x_2 + \alpha]$ is not detected as the point, where α is a threshold value. The detected local maximum points $x_{\max}$ and local minimum points $x_{\min}$ are shown in the plotted points of Fig. 6. Finally, $\overline{x_{\max}}$ and $\overline{x_{\min}}$ are calculated using detected local maximum points and local minimum points.

V. Experiments

A. Head pose estimation

Before applying the proposed method to speech scene, we evaluated the accuracy of proposed head pose estimation. We generated 5,826 face images using BU-4DFE [10], each face image is oriented to one of 63 head poses. The head pose is defined at $(\theta, \psi, 0)$, where θ is the pitch angle and its range is from $-15°$ to $15°$ with a step of $5°$ and ψ is the yaw angle and its range is from $-40°$ to $40°$ with a step of $10°$. Next we applied the proposed head pose estimation method to all generated face images, and calculated mean absolute error (MAE) between an actual yaw angle and an estimated yaw angle. As the result, we obtained MAE of $7.45°$ when considered all ranges. Here, when ψ is limited to $[-10°, 10°]$ in which this range is the same as the moving range described in the next experiment, MAE of $3.92°$ was obtained and this is sufficient value for speech scene.

TABLE III
Mutual correlation matrix.

	yaw[deg]	$roll$[deg]	G_x[pixel]	G_y[pixel]
yaw[deg]	1	-0.351	0.541	0.019
$roll$[deg]	—	1	-0.060	0.036
G_x[pixel]	—	—	1	-0.036
G_y[pixel]	—	—	—	1

B. Motion analysis

Head pose estimation described in **III** was applied to nine speech scenes, and obtained four parameters $\{yaw, roll, G_x, G_y\}$ at each frame. Before analysis the motion, mutual correlation matrix with all scenes was calculated, and its is shown in Table III. It was confirmed that yaw and G_x had strong correlation.

Next, the motion analysis described in **IV** was applied to nine speech scenes, and results are shown in Table IV. Table IV(a) shows calculated results of frequencies of four parameters. The standard deviations of these values are 5.6 at most, and all frequencies are almost the same. Table IV(b) shows calculated results of moving ranges of four parameters. It is obvious that yaw and G_x related to the horizontal movement have large ranges compared with $roll$ and G_y related to the vertical movement. The moving range is different depending on speakers.

Our dataset contains not only the video data but also scores by raters. Next, we calculated correlation coefficients between the calculated values and score. Here, our analysis method is focused on facial movement, and only eye contact score is used for evaluation. Calculated correlation coefficients are shown in Table IV(c). It was found that there is correlation between the moving range of G_y and eye contact score.

Since our features frequency and moving range are suitable for speech instruction. For example, moving ranges of yaw and roll angles of JF-R-5 whose eye contact score is 7.4 were smaller than moving ranges of other speakers. Especially, moving ranges of JF-R-4 whose score is 8.8 were largest among other speakers. By using these features, the teacher can be instructed JF-R-5 to see more wide ranges.

VI. Conclusion

Public speaking is an essential skill for a large variety of professions and in everyday life. However, it is difficult to master the skills. To solve this problem, recently, automated scoring methods are proposed. However, these methods are required special devices, and these methods are not suitable for instruction how to improve the speaking. Therefore, this paper presents a simple and efficient head pose estimation method and motion analysis method. As for the head pose estimation, our method was obtained MAE of $3.92°$ when the

TABLE IV
ANALYSIS RESULTS.

(a) Frequency.

speaker	yaw[Hz]	$roll$[Hz]	G_x[Hz]	G_y[Hz]
JF-R-1	22.4	22.4	22.0	14.9
JF-R-2	11.9	26.5	16.9	24.2
JF-R-3	12.8	14.4	13.4	20.8
JF-R-4	16.1	13.8	16.1	18.4
JF-R-5	15.3	15.3	14.3	32.6
JF-R-6	16.6	21.9	14.5	14.1
JF-R-7	15.1	15.1	13.9	18.4
JF-R-8	12.7	12.7	12.7	17.6
JF-R-9	15.7	14.3	15.7	22.0
ave.	15.4	17.4	15.5	20.3

(b) Moving range.

speaker	yaw[deg]	$roll$[deg]	G_x[pixel]	G_y[pixel]
JF-R-1	8.5	4.0	33.6	18.4
JF-R-2	14.6	9.0	53.7	47.8
JF-R-3	11.1	8.6	37.4	18.4
JF-R-4	20.9	13.4	70.5	63.0
JF-R-5	5.7	4.0	40.5	36.7
JF-R-6	9.4	9.2	41.3	40.5
JF-R-7	17.7	9.8	66.7	31.8
JF-R-8	17.7	12.6	82.3	41.2
JF-R-9	12.1	4.6	71.8	22.9
ave.	13.1	8.4	55.3	35.6

(c) Correlation coefficient with eye contact score.

	yaw[deg]	$roll$[deg]	G_x[pixel]	G_y[pixel]
Frequency	0.347	-0.076	-0.071	-0.295
Moving range	0.119	0.239	0.044	0.461

target angle is limited from -10° to 10°, and this value was acceptable accuracy for the speech scene. Proposed motion analysis method can be calculated frequencies and moving ranges of head motion. To show the effectivity, we collected a dataset which took on an official English recitation contest at high school, and applied the proposed method to the speech scenes. Moreover, calculated parameters were evaluated with eye contact score by raters. It was found that there is correlation between the moving range of G_y and eye contact score.

Related researches are evaluated not only head pose but also eye movement, emotion, body movement, and hand gesture. Future work will add other modalities to our method.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 15K12416.

REFERENCES

[1] D.-P. Pertaub, M. Slater, and C. Barker, "An experiment on public speaking anxiety in response to three different types of virtual audience," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 1, pp. 68–78, 2002.

[2] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer, "Cicero - towards a multimodal virtual audience platform for public speaking training," in *Intelligent Virtual Agents (IVA2013)*, 2013, pp. 116–128.

[3] M. Chollet, K. Stefanov, H. Prendinger, and S. Scherer, "Public speaking training with a multimodal interactive virtual audience framework," in *ACM on International Conference on Multimodal Interaction (ICMI2015)*, 2015, pp. 367–368.

[4] L. Chen, C. W. Leong, G. Feng, C. M. Lee, and S. Somasundaran, "Utilizing multimodal cues to automatically evaluate public speaking performance," in *International Conference on Affective Computing and Intelligent Interaction (ACII2015)*, 2015, pp. 394–400.

[5] V. Ramanarayanan, C. W. Leong, L. Chen, G. Feng, and D. Suendermann-Oeft, "Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring," in *ACM on International Conference on Multimodal Interaction (ICMI2015)*, 2015, pp. 23–30.

[6] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[7] T. Horprasert, Y. Yacoob, and L. S. Davis, "Computing 3-D head orientation from a monocular image sequence," in *International Conference on Automatic Face and Gesture Recognition (FG1996)*, 1996, pp. 242–247.

[8] J.-G. Wang and E. Sung, "EM enhancement of 3d head pose estimated by point at infinity," *Image and Vision Computing*, vol. 25, pp. 1864–1874, 2007.

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[10] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *8th International Conference on Automatic Face and Gesture Recognition (FG2008)*, 2008.