

Initial Assessment of Job Interview Training System using Multimodal Behavior Analysis

The COVID-19 pandemic has had a significant socio-economic impact on the world. Specifically, social distancing has impacted many activities that were previously conducted face-to-face. One of these is the training that students receive for job interviews. Thus, we introduce a job interview training system that will give students the ability to continue receiving this type of training. Our system recognizes the nonverbal behaviors of an interviewee, namely gaze, facial expression, and posture using a Tobii eye tracker and cameras. The system compares the recognition results with those of models of exemplary nonverbal behaviors of an interviewee and highlights the behaviors that need improvement while playing back the interview recording. Most current interview training systems require high-end Hardware and Software and are not designed for general users, and there are few systems using CG agents to give feedback. The development goal for our system was to construct an inexpensive and easy-to-use system using commercially available HWs, open-source code, and a CG agent that would provide feedback to the interviewee. The results of the initial evaluation of the system indicate that improvements in the recognition accuracy of nonverbal behaviors and the quality of the interaction with the CG agent are needed.

CCS CONCEPTS • Human-centered computing~Interaction design~Empirical studies in interaction design; • Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI; • Human-centered computing~Human computer interaction (HCI)~HCI design and evaluation methods~User models; • Human-centered computing~Interaction design~Interaction design process and methods~User interface design;

Additional Keywords and Phrases: gaze recognition, posture recognition, facial expression recognition, job interview training, nonverbal behaviors, CG agent, multi-modal interaction

ACM Reference Format:

1 INTRODUCTION

From its onset in late 2019, COVID-19 has had a significant social and economic impact, worldwide, and people have been asked to avoid human contact as much as possible. As a result, the face-to-face training of interviewees, which could be useful in the employment search process, is now avoided. Interview training can help students acquire skills by experiencing the content and flow of job interviews and can increase their confidence in their search for employment. However, interview training has been limited due to the number of interviewers and time available for interviewing [1]. Moreover, the impact of COVID-19 has made interview training more difficult to conduct. This suggests that there is an increasing need for a system that allows students to train for job interviews independently.

There is a growing body of research demonstrating the power of the social signals that people consciously or unconsciously exhibit in a variety of situations, such as job interviews and group discussions. Visual nonverbal behavior during a dialog accounts for 55% of all the information conveyed [2]. Washburn et al. pointed out that the outcome of an interview is affected more by the nonverbal behaviors of an interviewee than their verbal behaviors [3]. Moreover, Arvey et al. noted that nonverbal behaviors such as gaze, body movements, and tone of voice greatly influence the interviewee's evaluation [4]. These studies show that the use of nonverbal behavior and its impact on job interview success has been a major focus in research.

In recent years, social signal processing techniques using multimodal information have been used for dialog analysis [5] and have been applied to AI-based interview recruitment systems [6, 7, 8, 9] and interview training systems [10, 11, 12, 13]. Specifically, there are those that visualize the information of the nonverbal behavior and provide feedback during or after the interview [14, 15, 16, 17], and those that change the behavior of the interviewer, i.e., the CG agent [18, 19, 20]. However, most of the research conducted in the field of social signal processing focused on the recognition of emotions based on speech and facial expressions and paid less attention to posture recognition. In addition, some studies [6, 7, 9] proposed high-end computer systems that are not affordable for general users.

It has also been reported that practicing interviewing with a CG agents is more effective in improving skills compared to using books and videos on job interviews [15, 17]. Other interview practice systems using CG agents have been found to elicit self-disclosure [21, 22]. [22] has shown using a CG agents increase user's self-disclosure and feelings of rapport, self-efficacy, and trust.

Consequently, the purpose of this study was to develop an interview training system with affordable HW and SW, specializing in the recognition of three types of nonverbal behaviors, namely, gaze, facial expression, and posture, and to utilize a CG agent that would provide feedback on the appropriateness of the nonverbal behaviors of interviewees. We expect this system allows people to practice their interview skills by themselves.

2 JOB INTERVIEW TRAINING SYSTEM

2.1 System Overview

The system was developed using Unity, FaceAPI [23], OpenPose [24], TobiiEyeTracker4C [25], and a webcam. This system consists of three phases: a demographic input phase, a mock interview phase, and a feedback phase. The demographic input phase (Figure 1) was used to input the number of users and gender. During the mock interview phase (Figure 2) the interview was video-recorded from a front-left angle in order to obtain the nonverbal behaviors of the interviewee, including gaze, facial expression, and posture. The

captured video was analyzed by the following procedures (see 2.5) and played back in the feedback phase. The system paused the video where feedbacks were needed and the CG agent provided feedbacks on any points for improvement(Figure 3). Figure 4 shows the overall configuration of the system.



Figure 1: Demographic input phase



Figure 2: Mock interview phase

Left: Experimental set up; Right: View of the display on the right screen



Figure 3: Feedback phase

Left: Experimental set up; Right: View of the display on the right screen, during standby state (top) and feedback state (bottom)

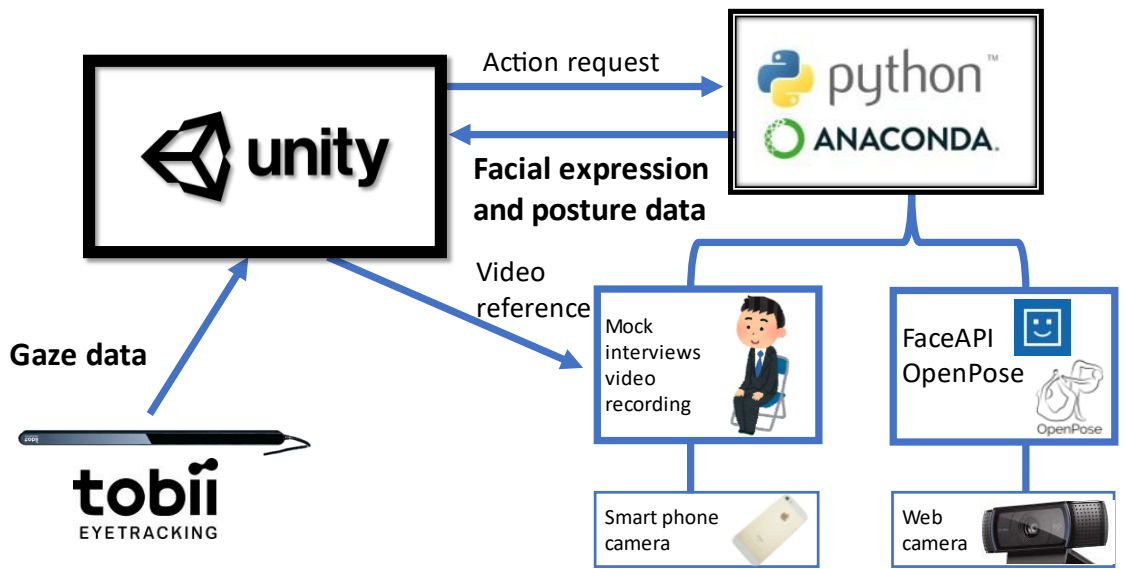


Figure 4: Overall system configuration

2.2 Interviewer

Our job interview training system used a video recording of a real person to conduct an interview. The videos of the interviewer were captured with the cooperation of the Employment Department of our university. There were 10 videos in total, and they comprised two types of questions and five interview patterns. For the questions, interviewees were asked to talk for one minute about either “self-promotion” or “their focus during university studies.” The interview patterns included “normal version,” “affirmative version,” “strict version,” “note-taking version,” and “look closely at the resume version.” From the 10 available videos, a random video would be played for each interview. Figure 5 shows an example of an interviewer's video.

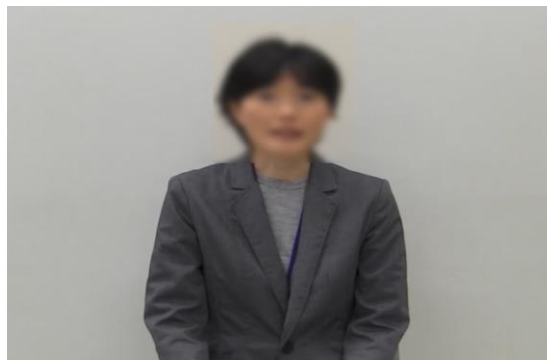


Figure 5: Interviewer (Employment Department of our university)

2.3 Feedback CG Agent and Feedback Method



In this system, a CG agent is used to provide feedback. The CG agent we use (Figure 6) is the Unity Asset "NATASHA" [26,27,28].



Figure 6: CG agent "NATASHA"

While the video taken during the mock interview is played back, the system can pause the video at any time according to the feedback algorithm and the agent provides advice. For example, " Was your posture not slumped at this time? Let's pay attention." This method allows the user to clearly distinguish between the parts of the mock interview where they performed well and where they can still improve. Table 1 shows examples of the feedback that the CG agent can give.

Table 1: Feedback examples

		Feedback content	
Detection information	【Gaze】 User's gaze take off the interviewer's face for more than five seconds.	"At this time, your gaze will be off the interviewer for a period of time. Let's pay attention." "Are you having trouble concentrating?"	
	【Facial expression】 Level of straight face	"Your expression may be stiff at this time."	
	【Posture】 User's legs are wide open.	"Aren't you opening your legs too much at this time? Let's be careful."	

2.4 Detectable non-verbal behaviors

The nonverbal behaviors were acquired at 1-second intervals for gaze and 3-second intervals for facial expression and posture. From the acquired behaviors, the following information was obtained:

- gaze rate,
- user's gaze moving off the interviewer's face for more than five seconds,
- number of times the gaze point moved to the upper right or upper left,
- level of smile or straight face,
- six facial expressions (anger, contempt, disgust, fear, sadness, surprise),
- posture (forward and backward leaning),
- legs open,
- legs opening gradually,
- shake of the neck, and
- protrusion of the elbow.

2.5 Method of detecting

2.5.1 Gaze Detection Method

We used a collision-detection method in order to detect inappropriate gazes. In order to determine when the interviewee's gaze moved off of the interviewer's face, we preset an area-of-interest (AOI) on the interviewer's entire face. The system determined inappropriate gaze when the interviewee's gazing point moved out of the AOI for five seconds. The gaze rate was calculated by dividing the number of frames in which the interviewee was looking at the interviewer's face by the total number of frames and displaying it as a percentage. The number of times the gazing point moved to the upper right or upper left was determined by setting up another AOI in the upper right and upper left areas of the screen (next to the interviewer). The system determined inappropriate gaze if the gazing point entered these areas more than 10 times, it was counted. Eventually, these metrics were used to determine the feedback given by the CG agent during the feedback phase.

2.5.2 Facial Expression Detection Method

For inappropriate facial expression detections, we used "smile" and "emotion" from FaceAPI. The smile and straight face scores were set at 0 and 1, respectively. A smile was determined as a result when a score of 0.5 or more was detected, and a straight face was determined as the result when 0 was detected three times in a row. Figure 7 shows an example of the detection results obtained from the FaceAPI.

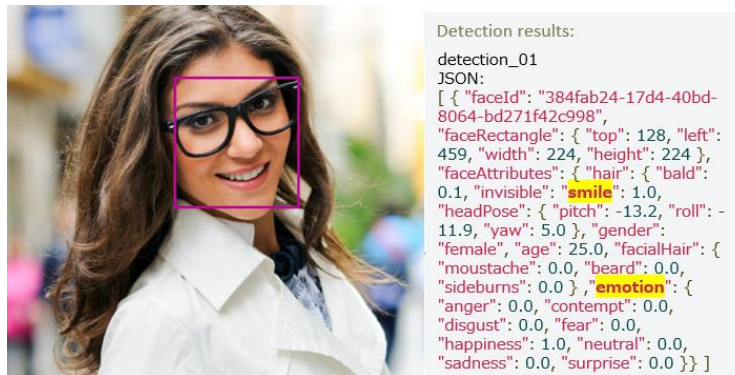


Figure 7: Example of the detection results from FaceAPI [23]

2.5.3 Posture Detection Method

To detect posture, we observed whether the interviewee

- leaned forward or backward,
- opened their legs,
- opened their legs gradually,
- shook their neck, or
- protruded their elbows.

Inappropriate postures were detected by comparing a correct posture model with the posture of the interviewee. A correct posture model (Figure 8) was created for each gender using OpenPose under the guidance of the Employment Department of our university. For example, whether an interviewee was leaning forward and backward was judged when there was a difference of more than 20 degrees between the model and the interviewee, legs in an open position was judged when the legs of the interviewee was wider than the width of model's legs, and protrusion of the elbow was judged when there was more than 15 degrees difference between the model and the interviewee. Figure 9 shows examples of the male posture model.

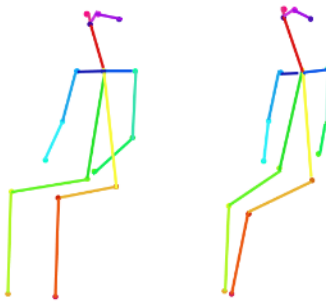


Figure 8: Correct posture model (left: male, right: female)

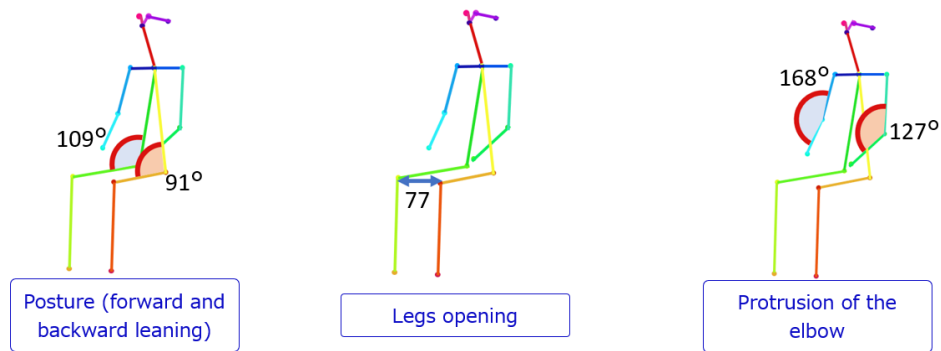


Figure 9: Detailed examples of a male model

2.6 Feedback Algorithm

The detected information on inappropriate nonverbal behaviors was stored in the gaze, facial expression, and posture arrays each second and then compiled into a single array using weighted prioritization. The weight was set in the order of gaze, facial expression, and posture, based on the order of importance during the interview. This also helped to avoid duplication of multiple detections in the same number of seconds and biased results that pointed to the same type of feedback. The CG agent would refer to these arrays when giving feedback. Figure 10 shows the flow of feedback from the acquired data, as described above.

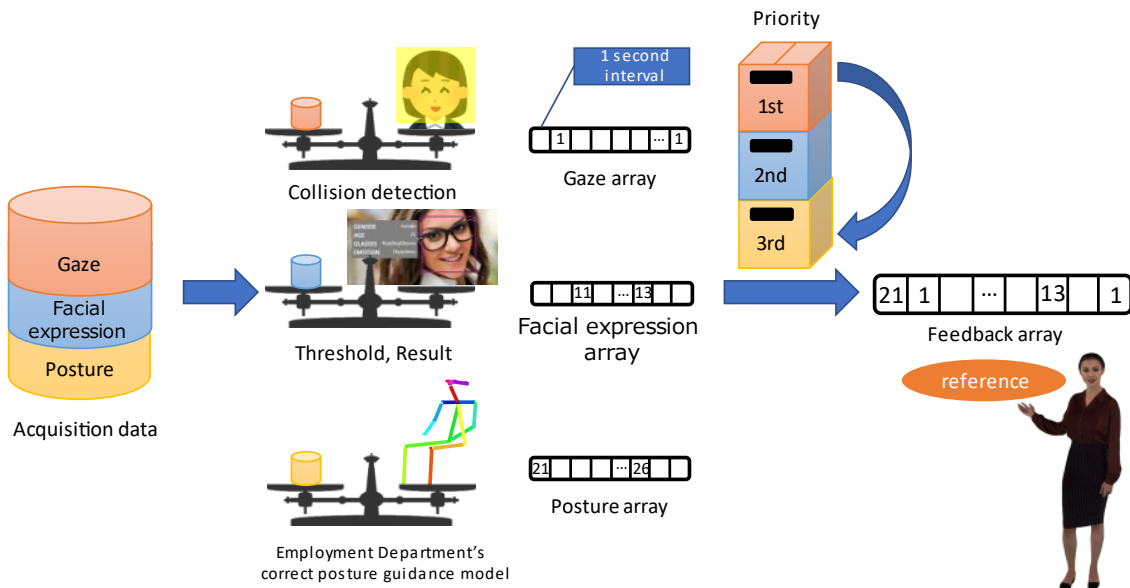


Figure 10: Feedback Algorithm

3 INITIAL EVALUATION EXPERIMENT

3.1 Experiment Overview

We conducted an initial evaluation of the system: we asked five university students (three males and two females) aged between 21 and 22 to use the system and interviewed them after the experiment to set their feedback. The experiment using human participants was approved by the Life Science Committee of our university.

Because the experimental environment was designed to simulate an actual interview situation, the subject was positioned considering the distance between the display and the user to ensure it was not too close and that the EyeTracker would work. In addition, we placed a webcam at an angle of 45° to the subject to estimate the posture using OpenPose, because it would not be possible to detect the forward and backward leaning posture and the angle of the feet using a frontal view. Figure 11 shows the experimental environment.

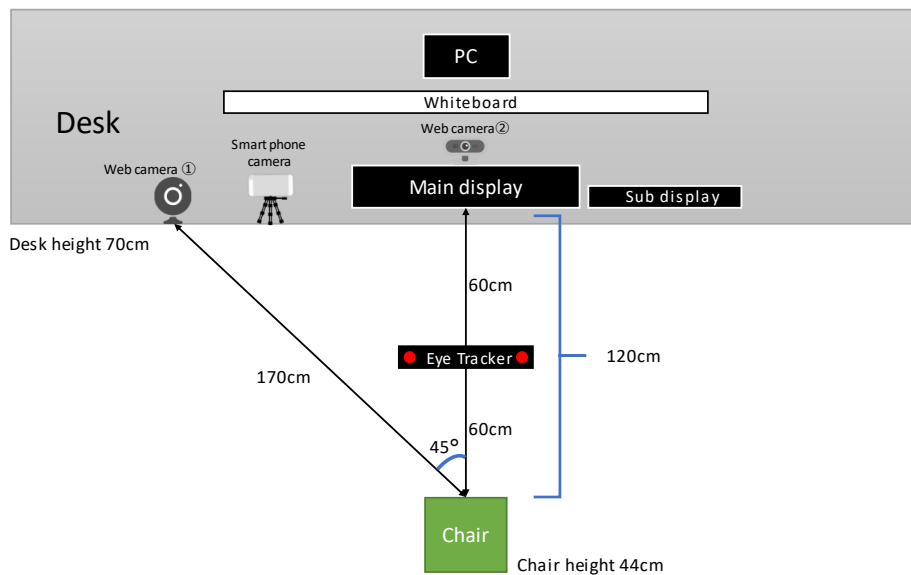


Figure 11: Experimental Environment

3.2 Impressions of the Presence of the CG Agent

We asked our participants on their impressions of the CG agent. During the interviews with the participants, we received positive comments such as “the appearance of an agent/secretary specializing in interviews,” “the image of a person who can point things out firmly and properly,” and “the feeling of being able to ask questions honestly without feeling uncomfortable.” On the other hand, we received negative comments such as “practice with actual humans will give you a more detailed view” and “it is practice with humans that is beneficial.” We asked the participants who gave negative comments if they would prefer receiving similar feedback from a human or a CG agent. They responded: “Either is fine as long as the feedback is the same,” and “I feel nervous

and embarrassed with a human, so it is easier to practice with a CG agent.” This shows that practice using a CG agent is more casual. It also shows that face-to-face interviews with humans could be considered as more serious practice. Therefore, as a future prospect, we plan to conduct a contracting experiment under the conditions of receiving feedback from humans and CG agents. We can use this to determine which of the two are more effective in listening and providing feedback.

3.3 Impressions of the Feedback by the CG Agent

We also asked our participants for their impressions of the feedback given by the CG agent. We received positive comments such as “Since the video stops and the CG agent gives pointers and advice each time, I can clearly see what was good or bad in each scene.” We also received negative comments such as “In the case of the CG agents, they only play back what they were originally prepared to play back, so you don't feel like they're really giving you personalized advice,” “CG agents are easy to use, but feedback from human beings are more acceptable and convincing.”

This may be because the interaction with CG agents is one-sided and uniform and deemed inferior to those with humans. As a countermeasure, it is necessary to prepare multiple voice patterns for the same point and increase the opportunities for interaction with the CG agent. For example, it may be possible to create a phase during which participants could talk about their impressions of the mock interview or give a self-evaluation, and the CG agent could respond to the content with praise, or encouragement.

3.4 Impressions on Recognition Accuracy

The participants provided us with feedback on the recognition accuracy of our system. The negative comments included: “I think there was an error in pointing out (the posture)” and “I thought it would be pointed out, but it was not.” In addition, judging from the comments and logs, the overall accuracy was approximately 50%.

The first reason for this is that the detection accuracy of OpenPose may be low. In this system, a web camera was placed at an angle of 45° to obtain information about the posture like leaning forward and backward. Because of the nature of OpenPose, it was not possible to detect the degree of sitting in a chair or whether the sitting posture was perpendicular using only an oblique angle, which reduced the accuracy. As a countermeasure, we believe that adding a web camera to give a frontal or side view or using a Kinect as a substitute may improve the accuracy.

Further, it was assumed that there could have been a detection error regarding the posture. Since we compared the interviewee's posture using only one reference model per gender and detected deviations through threshold judgment, we considered that the errors in the detection of postural deviations may have been caused by the fact that the reference model did not take individual differences into account. As a countermeasure to this problem, we believe that the accuracy can be improved by preparing several reference models and using one that is as close as possible to the user's height and body size for comparison and detection.

Finally, it was assumed that the empirical determination of the thresholds for gaze, facial expression, and posture may be the cause of low recognition accuracy. In the future, we need to refine the threshold value and investigate how the quality of the feedback can be improved by changing the threshold.

3.5 Impression of Feedback Content

Regarding the feedback content, we received negative comments such as “The feedback (type and number of points) is few.” In fact, each person received feedback approximately two to six times. This may be attributable to the detection accuracy, but could also be due to the weighted priority algorithm used to create the feedback array. The system prioritizes gaze, facial expression, and posture, in that order, based on its importance during interviews and tendency to change over time, and adds a weighting (which reconstructs the ranking when the ranking has not updated for 10 seconds) to the format of lowering the ranking of the detected types. Because there may be cases where the detection occurs, but it is not reflected in the feedback sequence, the algorithm needs to be improved. In addition, it was indicated that additional detection points may be necessary. One commenter said, “It would be nice if you could point out hand and finger movements, body swaying, loudness of voice, conjunctions, content of answers to questions, and speaking style,” indicating the need to point out verbal and nonverbal information.

3.6 Impressions on Usability of the System

We received positive comments on the usability of the system; e.g., “I don’t often have the opportunity to watch my own interview videos, so it’s good that they give me advice while watching the videos objectively during the feedback,” “If I could, I would like to practice again,” “The interviewers were very realistic. There was a sense of tension as if it was a real interview, and I felt like I was being watched,” and “I think I can do it alone (could be used by reluctant people who can’t actively go to the employment office to practice).”

On the other hand, we received negative comments such as “I don’t really want to watch my own interview videos.” Possible solutions to this comment include changing the user’s voice or appearance, like VTubers and avatars.

4 CONCLUSION

This paper introduces a job interview training system that recognizes the nonverbal behaviors of an interviewee, namely, gaze, facial expression, and posture. The proposed system uses a Tobii eye tracker for gaze recognition and camera images for facial expression and posture recognition. The system compares the recognition results of the interviewee’s nonverbal behaviors with exemplary models and points out the behaviors that need improvement while playing back the interview recording. Our development goal for the system was to construct an inexpensive and easy-to-use system using commercially available HWs, open-source code, and a CG agent that could provide feedback to the interviewee.

Initial evaluation experiments were conducted, and we identified some points that need improvement: the accuracy was measured at approximately 50% and is not sufficient, the interaction with the CG agent was one-sided and deemed inferior to that with humans, and the number of times feedback was given was approximately two to six times per person, which is too little in terms of volume and content.

As a future prospect, we need to improve the system regarding the above, and conduct an experiment with more participants. In addition, we will attempt to acquire nonverbal and verbal information other than gaze, facial expressions, and posture. Functions will be added to allow the interviewee to practice the corrections that were pointed out, and to review their own interview video after using the system. Improvements will be made to the algorithm for setting the priority of feedback content. We will also add a detection method that combines

two or more pieces of nonverbal information. We believe that these improvements are necessary. Furthermore, it is necessary to compare and verify the listening ability and feedback of CG agents and humans.

ACKNOWLEDGMENTS

REFERENCES

- [1] Yuko Matsuda, Minoru Nagasaku, Kunijiro Arai : Influence of job-hunting anxiety on job-hunting: From the viewpoint of coping, *The Japanese Journal of Psychology*, 2010, Vol. 80, No. 6, pp. 512-519
- [2] Mehrabian A, *Silent messages: Implicit communication of emotions and attitudes*, Wadworth Publishing.Co.,California, 1981
- [3] Washburn P.V., Hakel M.D. , Visual cues and verbal content as influences on impressions after simulated employment interviews.*Journal of AppliedPsychology*, pp.58,137-140, 1973
- [4] R. D. Arvey , J. E. Campion, "The employment interview: A summary and review of recent research," *Personnel Psychology*, vol. 35, no. 2, pp. 281–322, 1982.
- [5] Shogo Okada, Yoshihiro Matsugi, Yukiko Nakano, Yuki Hayashi, Hung-Hsuan Huang, Yutaka Takase, Katsumi Nitta, Estimating Communication Skills based on Multimodal Information in Group Discussions, *Journal of the Japanese Society for Artificial Intelligence*,Vol.31,No.6, A130-E,2016
- [6] MIDAS Information Technology Co., Ltd., <https://www.inair.co.jp/>
- [7] ZENKIGEN Co., Ltd., <https://harutaka.jp/>
- [8] Naim, I., Tanveer, M. I., Gildea, D., Hoque, M. E.: Automated prediction and analysis of job interview performance: The role of what you say and how you say it, *IEEE FG*, 2015
- [9] Rao S. B. P., Rasipuram, S., Das, R., Jayagopi, D. B.: Automatic assessment of communication skill in non-conventional interview settings: A comparative study, *ICMI*, pp. 221–229, 2017
- [10] Nanaho Goda, Keitaro Ishihara, Tomoko Kojiri, *Job Interview Support System Based on Analysis of Nonverbal Behavior*, *IEICE Technical Report*,Vol.116,No.517,ET2016-98,25-30,2017
- [11] T. Barur, T. Ionut, G. Patrick, P. Kaska, A. Elisabeth.: A Job Interview Simulation: Social Cue-Based Interaction with A Virtual Character, *IEEE International Conference on Social Computing (SocialCom2013)*, pp.220–227, 2013
- [12] J. Matthew, B. Laura, F. Micael, J. Neil, A. Michael, J.Emily, W. Katherine, O. Dale, Morris, D, B.: Virtual Reality Job Interview Training for Veterans with Posttraumatic Stress Disorder, *Journal of Vocational Rehabilitation* 42, pp. 271–279, 2015
- [13] H. Tanaka, S. Sakti, Graham. N, T. Toda, H. Negoro, H. Iwasawa, S. Nakamura: Automated Social Skills Trainer, *IUI '15 Proceedings of the 20th International Conference on Intelligent User Interfaces*, pp. 17–27, 2015
- [14] Anderson, K., Andre, E., Baur, T., Bernardini, S., Chollet, M., Chryssaifidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., Jones, H., Ochs, M., Pelachaud, C., Porayska-Pomsta, K., Rizzo, P., Sabouret, N.: The TARDIS framework: Intelligent virtual agents for social coaching in job interviews, *ACE*, pp. 476–491, 2013
- [15] Damian, I., Baur, T., Lugin, B., Gebhard, P., Mehlmann, G., Andre, E.: Games are better than books: In-situ ´ comparison of an interactive job interview game with conventional training, *AIED*, pp. 84–94, 2015
- [16] Hoque, M. E., Courgeon, M., Martin, J.-C., Mutlu, B., Picard, R. W.: MACH: My automated conversation coach, *UBICOMP*, pp. 697–706 ,2013
- [17] Langer, M., Konig, C. J., Gebhard, P., Andr ´ e, E.: ´ Dear computer, teach me manners: Testing virtual employment interview training, *International Journal of Selection and Assessment*, Vol. 24, No. 4, pp. 312–323, 2016
- [18] Baur, T., Damian, I., Gebhard, P., Porayska-Pomsta, K., Andre, E.: A job interview simulation: Social cue-based interaction ´ with a virtual character, *SocialCom*, pp. 220–227, 2013
- [19] Callejas, Z., Ravenet, B., Ochs, M., Pelachaud, C.: A computational model of social attitudes for a virtual recruiter, *AAMAS*, pp. 93–100 ,2014
- [20] Gebhard, P., Baur, T., Damian, I., Mehlmann, G., Wagner, J., Andre, E.: Exploring interaction strategies for virtual ´ characters to induce stress in simulated job interviews, *AAMAS*, pp. 661–668, 2014
- [21] A. N. Joinson. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2), pp.177–192, 2001
- [22] T. Bickmore, A. Gruber, R. Picard. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient Education and Counseling*, 59(1) pp.21–30, 2005.
- [23] MicrosoftAzure, <https://azure.microsoft.com/ja-jp/services/cognitive-services/face/>

- [24] tf-pose-estimation, <https://github.com/apulis/tf-pose-estimation>
- [25] Tobii Technology K.K.,<https://www.tobiipro.com/ja/>
- [26] Metastage: "NATASHA" - Seated Listening,<https://assetstore.unity.com/packages/3d/characters/humanoids/humans/natasha-seated-listening-153787>
- [27] Metastage: "NATASHA" - Presentation,<https://assetstore.unity.com/packages/3d/characters/humanoids/humans/natasha-presentation-153909>
- [28] Metastage: "NATASHA" - Serious Talking,<https://assetstore.unity.com/packages/3d/characters/humanoids/humans/natasha-serious-talking-153785>