Journal of Intelligent & Robotic Systems
https://doi.org/10.1007/s10846-019-01052-8

First-person Video Analysis for Evaluating Skill Level in the *Humanitude* Tender-Care Technique

Atsushi Nakazawa[1] · Yu Mitsuzumi[2] · Yuki Watanabe[1] · Ryo Kurazume[3] · Sakiko Yoshikawa[4] · Miwako Honda[5]

Received: 30 August 2018 / Accepted: 12 June 2019

Abstract

In this paper, we describe a wearable first-person video (FPV) analysis system for evaluating the skill levels of caregivers. This is a part of our project that aims to quantize and analyze the tender-care technique known as *Humanitude* by using wearable sensing and AI technology devices. Using our system, caregivers can evaluate and elevate their care levels by themselves. From the FPVs of care sessions taken by wearable cameras worn by caregivers, we obtained the 3D facial distance, pose and eye-contact states between caregivers and receivers by using facial landmark detection and deep neural network (DNN)-based eye contact detection. We applied statistical analysis to these features and developed algorithms that provide scores for tender-care skill. In experiments, we first evaluated the performance of our DNN-based eye contact detection by using eye contact datasets prepared from YouTube videos and FPVs that assume conversational scenes. We then performed skill evaluations by using Humanitude training scenes involving three novice caregivers, two Humanitude experts and seven middle-level students. The results showed that our eye contact detection outperformed existing methods and that our skill evaluations can estimate the care skill levels.

Keywords Dementia · Care · Deep neural network (DNN) · Skill evaluation · Wearable system · Computer vision · First person video

✉ Atsushi Nakazawa
nakazawa.atsushi@i.kyoto-u.ac.jp

Yu Mitsuzumi
yu.mitsuzumi.ae@hco.ntt.co.jp

Ryo Kurazume
kurazume@ait.kyushu-u.ac.jp

Sakiko Yoshikawa
yoshikawa.sakiko.4n@kyoto-u.ac.jp

Miwako Honda
honda-1@umin.ac.jp

[1] Graduate School of Informatics, Kyoto University, Kyoto, Japan

[2] NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation (NTT), Kanagawa, Japan

[3] Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan

[4] Kokoro Research Center, Kyoto University, Kyoto, Japan

[5] Geriatric Research Division, National Hospital Organization Tokyo Medical Center, Tokyo, Japan

1 Introduction

As the elderly population increases, the number of people suffering from dementia continues to grow. As a result, the care that needs to be administered to them is becoming increasingly important in social terms [8, 25, 42]. The population of people in Japan afflicted with dementia is expected to exceed seven million by 2025. A more serious problem is the shortage of caregivers. The number of caregivers that will be necessary in 2025 is estimated to be 2.53 million, but the actual number is estimated to be 2.15 million [29].

Dementia occurs when the brain is damaged by maladies such as Alzheimer's disease and Lewy body dementia and produces a set of symptoms that include memory loss and difficulties with thinking, problem-solving, and verbal communication. Dementia can be accompanied by psychosis, agitation and aggression; thus, caring for people with dementia is quite difficult [7].

Two approaches can be cited as ways to alleviate the difficulties this poses for caregivers. The first is providing patients with customized treatment, which can slow the progression of dementia and prevent side effects such as

Published online: 04 July 2019

infections. The second is reducing the burden on caregivers to prevent them from "burning out." Dementia can cause symptoms similar to those of mental illnesses, known as behavioral and psychological symptoms of dementia (BPSD), so caregivers' working conditions can be harsh. As a result, the number of caregiving staff members that leave and become burned out is increasing.

***Humanitude* tender-care style**: Due to this social background situation, the caregiving style *Humanitude* has been spotlighted by care professionals and family caregivers since it can reduce the occurrence of BPSD events and caregivers' burden [20]. Humanitude was developed by Y. Geneste and R. Marescotti 35 years ago [18] and has been introduced in more than 600 hospitals and nursing homes in Europe. Humanitude primarily uses a combination of four communication skills: gaze, verbal communication, touch, and helping care receivers to stand up. Several studies have reported that the cost-efficiency of introducing Humanitude is around 20 times that of care without it because of a 40% decrease in the use of psychotropic drugs and in the number of care staff members who leave [22]. In recent years, Humanitude has become popular in Japan; in the past three years, more than 2,600 people took Humanitude training over the course of more than 30 training sessions.

Computational tender-care science project: Since we believe improved care techniques using robotics and computer vision technologies are valuable to both caregivers and care receivers, we started a comprehensive project that aims to (a) quantize and visualize the Humanitude skills, (b) reveal the brain mechanism behind Humanitude-based communications and (c) develop a system that will help people to learn Humanitude skills (Fig. 1). This paper shows one of the topics in (a) skill quantization, but we briefly introduce all topics below.

For (a), we developed a system that automatically finds the skill elements by using wearable sensors that capture learner's and care receivers' behaviors. The system uses data mining and recognition algorithms developed to enhance computer vision and machine learning. We then obtained the *essence of Humanitude skills* through multi-modal analysis.

For (b), we tried to reveal why using Humanitude facilitates communication with people with dementia (PwD) and why it reduces BPSD through cognitive neuroscientific approaches. Namely, we conducted functional neuroimaging studies to find the differences between younger people and elderly people, and between healthy people and people suffering from Alzheimer's disease while giving emotional stimuli such as facial pictures that show eye contact or dynamic facial expressions.

For (c), on top of the (a) and (b) findings, we developed a tender-care education platform that presents the caregivers' skill level to learners by using our care-skill evaluation systems. With these systems, learners will be able to evaluate their current tender-care skill levels easily and at a low cost by themselves. It follows that this platform will be suitable for non-professional caregivers such as family caregivers, as well as for professional caregivers who want to periodically refresh their caregiving skills.

Wearable sensing technique for care behavior analysis: Since Humanitude consists of communication behaviors from close distances such as gaze and touch, we use wearable sensing devices to extract events in which such behaviors occur. We observed care techniques that do and do not use Humanitude and found behavioral differences and their outcomes through statistical computational analysis-based approaches, as illustrated in Fig. 2.

This paper describes the first step of this project – a system for extracting face-to-face communication behavioral skills by using a head-mounted camera worn by a caregiver. From camera images obtained using mutual facial distances, we obtained poses and eye contact states by using facial parts tracking and deep neural network (DNN)-based eye contact detection algorithms. We compared these behavioral elements from those we ascertained among care novices, middle-level Humanitude-care learners and Humanitude-care experts. Although many attempts have been made to obtain the aforementioned information by using third-person view videos, the videos were analyzed by video annotators. This required considerable cost and time and there was the risk of obtaining biased results caused by annotator subjectivity [21]. To address these problems, the contributions and limitations of this paper are as follows:

1. It describes our development of a prototype system that uses wearable cameras and image analysis for care skill evaluation.
2. It describes our development of DNN-based eye contact detection algorithms that outperform existing approaches by using eye contact datasets based on YouTube videos and first-person video (FPVs).
3. It describes how we obtained an FPV dataset while conducting Humanitude training sessions for novices, middle-level caregivers and expert caregivers and found the differences among them regarding face-to-face distance and pose (angle) as well as eye contact frequency.
4. It describes how we performed unsupervised principal component analysis (PCA) for the features obtained from FPVs and found significant correlation between caregiver levels and PCA scores.
5. The present research represents a preliminary analysis that uses a relatively small number of datasets. It will be thus be followed by our analysis of extensive studies using a much larger number of samples. It will be also important to use chronological behavioral data for the same individuals to observe how the skills were acquired or forgetten.

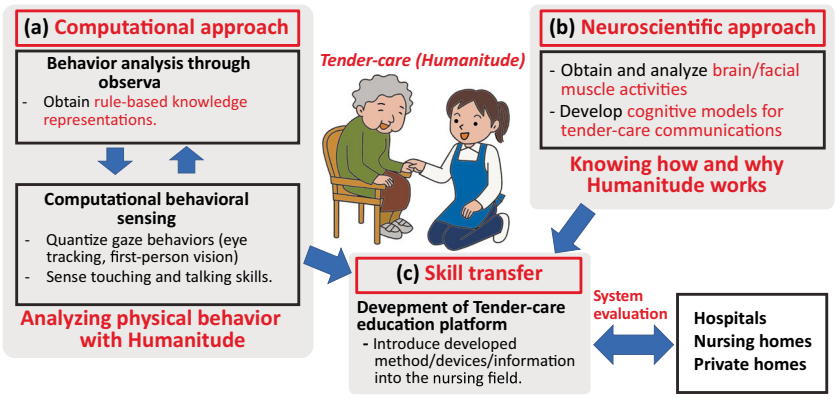


Fig. 1 Project overview. The Humanitude tender-care technique was analyzed from both computational and cognitive neuroscientific approaches

2 Related Work

In this section, we show related studies regarding caregiver's burden and effects of intervention, care skill evaluation, first-person video-based skill evaluations and eye contact detection from video images.

Caregiver's burden and effects of interventions: The burden of caregiver for dementia patients have been reported in a lot of literature [2, 10, 12, 15, 16, 23, 33, 41]. According to the recent meta-review paper[2], the larger caregivers burden is related to 1) female sex, 2) low education, 3) cohabitation with care recipient, 4) caregiving time and effort, 5) financial stress, 6) lack of choice and inability to continue regular employment. As the result, caregiver tends to having larger risk in mortality, weight loss, poor self-care and sleep deprivation.

Effects of interventions for reducing caregiver's burden have been reported as well [1, 13, 15, 19, 31, 32]. Interventions are categorized into several types: psychoeducational intervention, psychosocial intervention, cognitive behavior therapy, respite, caregiver support groups, anticholinergic and antipsychotic drugs, and skill training. As the results, practical interventions to reduce caregiver's burden are 1) encouraging caregivers to function as a member of the care team, 2) encouraging caregivers to improve self-care and maintain their health, 3) providing education and information, 4) coordinating for assistance with care, 5) encouraging caregiver to access respite care and 6) using the supports of technology [2]. Specifically, there are several reports that skills training such as coping skill training (CST) may reduce the caregiver strain, depression and fatigue in caregivers of the patient with cancers [9, 31].

Care skill quantization: There are several approaches that use care skill quantization. In computer science, Ishikawa et al. developed a method of care skill evaluation based on the knowledge of care experts [21]. They categorized care skills into three layers: intramodality, intermodality and multimodal-interaction. Intramodality consists of behavior primitives such as gaze, speech, touch, nodding and knocking on a door. Intermodality shows the relationships among intramodalities, such as comprehensiveness of care, waiting for elderly people's actions and consistency. Multimodal-interaction consists of actions that develop a relationship between actors, such as eye contact and

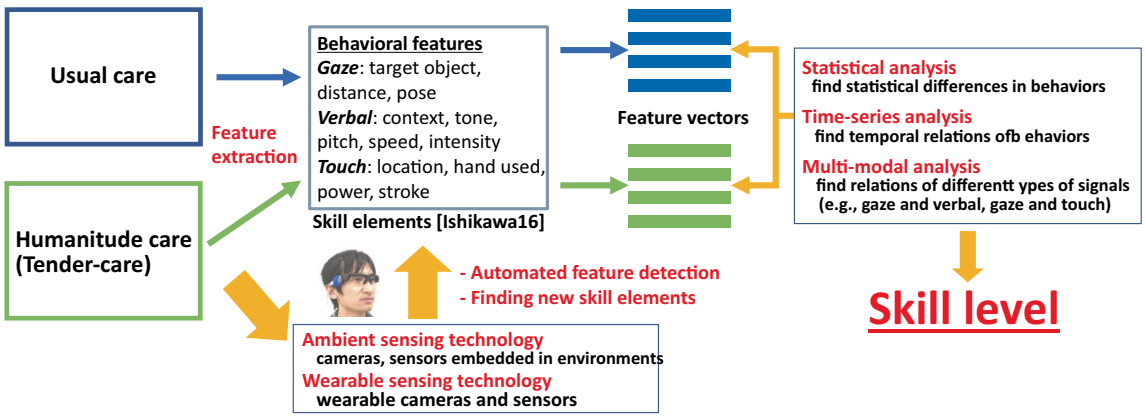


Fig. 2 Computational skill extraction obtained using behavioral observation and wearable sensors. We quantized care skill elements with and without Humanitude and found the differences between them

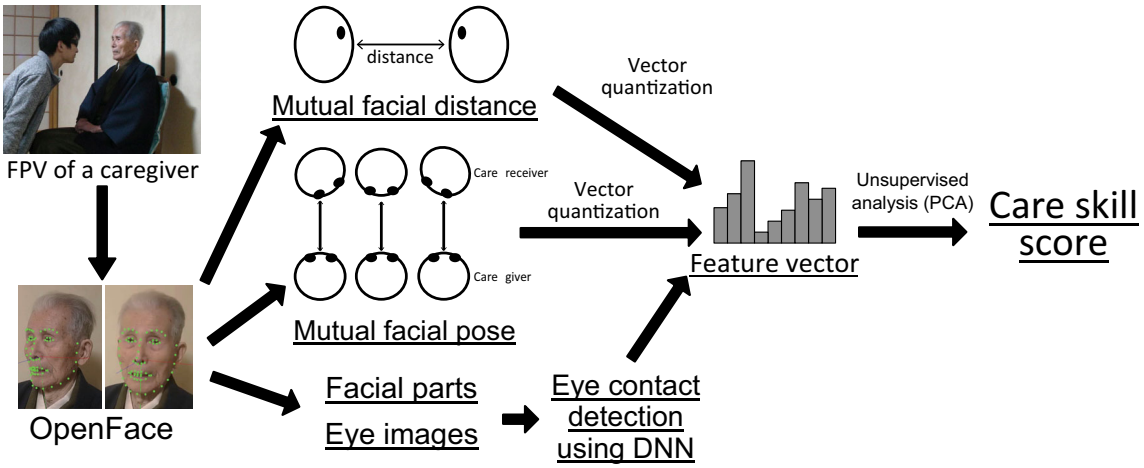


Fig. 3 Overview of the care skill estimation method using first-person video

verbal/nonverbal dialogue. They also developed a web interface that shows care learners care skills in visual form to confirm the effectiveness of the system.

First-person video analysis for skill evaluation: There have been a number of studies on action recognition and prediction using FPVs [17, 28, 37, 39, 40]. However, few studies have been conducted for *skill* evaluation. In recent years, Bertasius et al. showed a method to assess a basketball player's performance from FPVs. They designed and used temporal CNN and long short-term memory (LSTM) architecture to evaluate whether a particular play in basketball was good or not from a player's FPV [5]. In the medical field, Hei et al. proposed a method for evaluating skill in robotic surgical operations from video images. Their method tracks the keypoints of surgical robot instruments by using cloud sourcing or hourglass networks and evaluates the skill by support vector machine analysis [26].

Video-based eye contact detection: Detecting and making eye contact are important for understanding social communication and designing communication robots. Therefore, several studies in this area have been conducted. Smith et al. [38] proposed an algorithm to detect *gaze-locking* (looking at a camera) faces using eye appearances and PCA plus multiple discriminate analysis. Ye et al. developed a pioneering algorithm that detects mutual eye gaze using wearable glasses [43, 45]. In recent years, deep-learning-based approaches are being implemented for eye-contact detection. Mitsuzumi et al. developed the DNN-based eye contact deteciton algorithm (DeepEC)[30] that uses only cropped eye regions for eye-contact detection and performed better than existing methods. Eunji et al. develop the DNN-based PiCNN detector that accepts the facial region and output both facial postures and eye contact states [11]. Zhang et al. presented an eye-contact detection algorithm based on their deep neural network (DNN) based gaze estimations [48]. In robotics, Petric et al. developed an eye contact detection algorithm that uses facial images taken with a camera embedded in a robot's eyes [34] to develop robot-assisted ASD-diagnosis systems. These eye contact detection algorithms depend on facial landmark detection libraries or gaze estimation algorithms with which it is assumed that subject faces are not occluded.

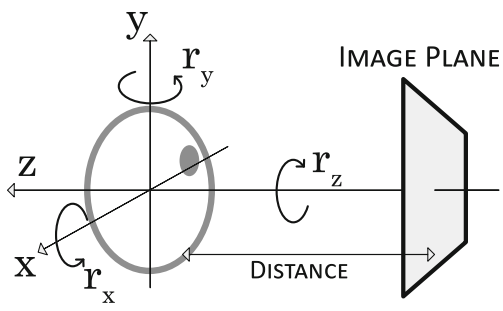
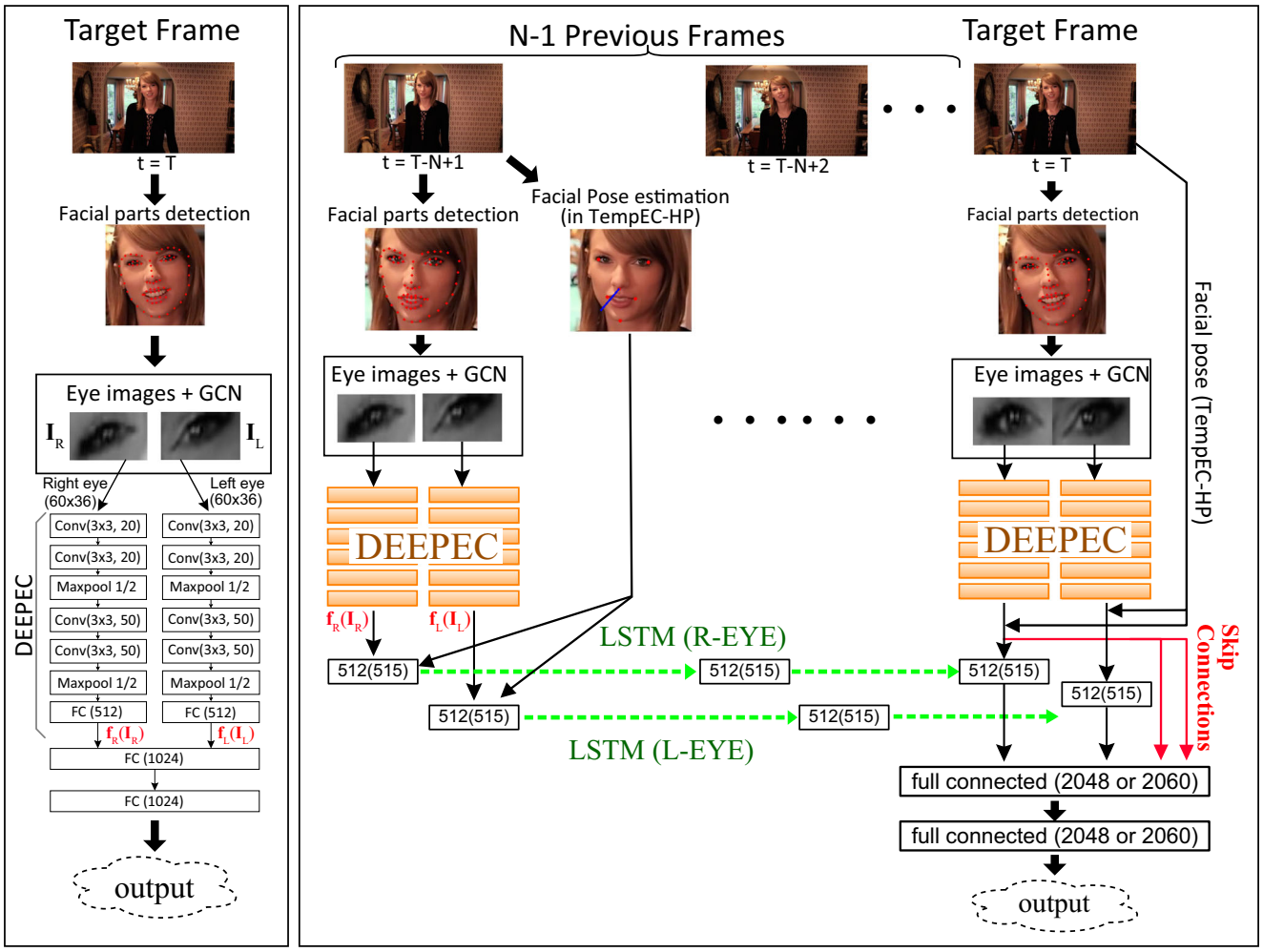


Fig. 4 Facial pose parameters obtained from OpenFace

Image-based gaze estimation algorithms have also been recently studied, although they differ in scope from our detection algorithms. The current trend in this area is deep learning-based approaches, namely, learning and predicting gaze directions according to datasets that describe the relation between facial images, facial landmarks, and gaze points. For example, Lu et al. developed a head pose-free gaze estimation method by synthesizing eye images from small samples. However, their method requires personal-dependent eye image samples taken under experimental setups [27]. Zhang et al. proposed a DNN algorithm that inputs eye images and 3D head poses obtained from facial landmark points [46]. They also developed a DNN-based algorithm using full facial images without occlusions [47]. Krafka et al. developed a DNN-based eye gaze



(a) Singleframe Eye Contact detection (DEEPEC)

(b) Proposed algorithm (TempEC, TempEC-HP)

Fig. 5 Eye contact detection algorithms. **a** The single-frame eye contact detection algorithm (DeepEC) first detects the eye regions of the target image frame and obtains a pair of right and left eye images. The eye contact state is obtained by only the CNN that inputs the eye images. **b** Proposed temporal eye contact detection algorithms that use multiple (i.e. N) image frames. First, they detect facial landmarks with OpenFace face detection, with which they then obtain eye regions in each of the N frames. The resulting N pairs of eye images are inputted to CNNs that have a structure similar to that of DeepEC. These CNNs are followed by an LSTM network, which learns the temporal state of the eyes. Finally, the target eye contact state is obtained by the following fully connected networks, which use not only the LSTM's outputs but also the CNN's outputs of the target frame ($t = T$) with skip connections

estimation algorithm that inputs full facial images as well as eye images [24].

In contrast, our detection algorithms only output binary (eye-contacted/averted) information. However, they do not require personal-dependent samplings and are robust to facial occlusions, which frequently occur in FPVs in caregiving and communication scenarios. This was achieved by designing a CNN that only uses images taken around the eyes.

3 Proposed Method

The flow of our skill evaluation is illustrated in Fig. 3. From a first-person camera worn by a caregiver we obtained mutual facial distances, mutual facial poses and eye contact states. Then, we estimated tender-skill scores through an unsupervised analysis. In the following subsections, we first describe the first-person camera hardware and then our algorithms we used for analyzing FPVs.

Table 1 First person eye-contact dataset (Youtube dataset)

Name	Duration (sec)	Fps	Image size (H,V)
Avec	350	29.97	1280×720
Azis	479	23.98	1280×720
Derek	490	23.98	1280×720
Elle	391	29.97	1280×720
Emma	555	23.98	1280×720
Wataru	55	29.97	540×360
James	380	23.98	1280×720
Kendall	447	23.98	1280×720
Liza	447	23.98	1280×720
Neil	687	23.98	1280×720
Selena	460	23.98	1280×720
Mai	337	29.12	1280×720
Taylor	581	23.98	1280×720

3.1 Hardware

We used two types of head-mounted first-person camera systems. One was a Pivothead Kudu camera [35], which is equipped with a front-view camera in the middle of a pair of glasses. The camera takes full HD (1920 × 1080 pixels) videos at 30 fps. The other was a Pupil Labs camera system [36], whose frontal camera also takes full HD (1920 × 1080 pixels) videos at 30 fps. The cameras' projection matrices were obtained by using the MATLAB camera calibrator.

3.2 Face Detection and 3D Pose Estimation

We then obtained facial positions, poses and eye locations from the input FPVs. We used the OpenFace library [4] and obtained 3D facial positions, poses and 68 facial landmark points. We computed the cameras' focal lengths from the camera projection matrices and used them to estimate 3D facial positions and rotations.

3.3 Histograms of Facial Distances and Poses

To quantize the face-to-face communication behaviors between caregivers and care receivers, we encoded the mutual facial distances and poses obtained from OpenFace as illustrated in Fig. 4. Namely, we computed the histograms $\mathbf{f}_{\text{dist}} = [f^1_{\text{dist}}, \ldots, f^{11}_{\text{dist}}]$, $\mathbf{f}_{r_x} = [f^1_{r_x}, \ldots, f^9_{r_x}]$, $\mathbf{f}_{r_y} = [f^1_{r_y}, \ldots, f^9_{r_y}]$ and $\mathbf{f}_{r_z} = [f^1_{r_z}, \ldots, f^9_{r_z}]$ that represent the mutual facial distances and poses from all frames in a care session. The bins were set to every 0.1 [m] from 0.0 to 1.0 [m] for the distance feature and 20 [deg] from -90 to +90

Fig. 6 Eye-contact dataset using publicly available videos from YouTube or conversational scnario (names with *)

Table 2 First person eye-contact dataset (Conversation dataset)

Name	Duration (sec)	Fps	Image size (H,V)
Imaizumi	220	29.97	1920×1080
Kitazumi	711	29.97	1920×1080
Ogawa	1865	29.97	1920×1080

[deg] for angular features. The distances larger than 1.0 [m] were voted to the last bin. Thus, for example, f^1_{dist} indicates the number of frames where the mutual facial distances were from 0.0 to 0.1 [m] and $f^4_{r_x}$ indicates the number of frames where the mutual facial rotation (r_x) was from -30 to -10 [deg].

3.4 Visualization

After obtaining histograms, we normalized the histogram and applied principal component analysis (PCA). While many data analysis or machine learning techniques have been proposed, we used PCA for its simplicity and reliability in exploratory data analysis (EDA). Since we tried to find tender-care technique skills in a bottom-up (data-driven) manner, this nature of PCA fitted our task better than other more complicated methods such as non-linear or supervised-learning based approaches.

We here denote $\mathbf{f}^{\text{s}}$ as a D × 1 column vector that represents a normalized histogram of either $\mathbf{f}_{\text{dist}}$, $\mathbf{f}_{r_x}$, $\mathbf{f}_{r_y}$ or $\mathbf{f}_{r_z}$ of a subject s ∈ {1, . . . , M}. Histograms of all subjects can be decomposed by using D × 1 column eigenvectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_{\text{M}}\}$ and eigenvalues $\{\lambda_{1,1}, \ldots, \lambda_{\text{D,M}}\}$, where D is the dimension of the histogram:

$$\left[\, \mathbf{f}^1 \ \cdots \ \mathbf{f}^{\text{M}} \,\right] = \left[\, \mathbf{e}_1 \ \cdots \ \mathbf{e}_{\text{D}} \,\right] \begin{bmatrix} \lambda_{1,1} & \cdots & \lambda_{1,\text{M}} \\ \vdots & \ddots & \vdots \\ \lambda_{\text{D},1} & \cdots & \lambda_{\text{D,M}} \end{bmatrix}.$$

We plot the eigenvalues of all subjects to visualize the distribution of their behaviors, as well as to analyze the elements of eigenvectors to find the relation between skill levels and behavioral features.

3.5 Eye Contact Features

Another feature is counting eye contact bids, which was introduced by Ye et al. [44]. They assume ”eye contact bids”, i.e., situations when a subject wearing an FPV camera is gazed at by other subjects. Since the definition of eye contact is making mutual eye gaze—two people look at each other at the same time – eye contact bids are not the same as actual eye contact. If we want to accurately detect eye contact, two people must wear FPV cameras or a caregiver must use an eye gaze tracker (EGT) device that detects observers' gaze information. However, from the practical point of view it is difficult to use two FPV cameras or an EGT device since 1) it is difficult for subjects with dementia to wear such devices and 2) even for caregivers it is difficult to use eye trackers in actual care scenes due to their noticeable appearance, calibration requirements and headmount drift. Therefore, rather than accurately detecting eye contact, we tried to measure and use eye contact bids for evaluating care skills. We used facial poses and eye images for detecting eye contact bids using DNNs.

Figure 5 illustrates the existing eye contact detection algorithm (Fig. 5a) and proposed TempEC and TempEC-HP (Fig. 5b). The TempEC uses only eye images whereas the TempEC-HP uses both eye images and 3D facial poses. These algorithms consist of the following components.

Eye region detection: From the landmark points detected by OpenFace, we obtained the right and left eye regions in the target frame, from which we obtained each eye image used as input for the CNN, after gray-scaling and

Fig. 7 Experimental scenes and example FPV frames of care learning scenes. **a** Data capturing, **b** first-person camera and recorder (Pupil Labs + smartphone) and (c) example frames (OpenFace annotations overlaid)

Table 3 The results of Experiment 1

Name	Total	Face	Eye	DeepEC [30]			DeepEC+CRF [30]			TempEC(Proposed)			TempEC-HP(Proposed)		
	Frames	Detected	Contacted	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Avec	10,498	8,940	4,671	0.8878	0.8262	0.8559	0.7881	0.8584	0.8217	0.8764	0.888	0.8822	0.8833	0.9241	0.9032
Aziz	11,508	6,711	5,058	0.9386	0.8695	0.9027	0.8838	0.9077	0.8956	0.935	0.8944	0.9142	0.9471	0.9636	0.9553
Derek	11,765	4,550	3,710	0.8865	0.8758	0.8811	0.8495	0.9292	0.8876	0.9244	0.8924	0.9081	0.8761	0.9478	0.9106
Elle	11,739	6,458	4,005	0.897	0.7371	0.8092	0.8522	0.7773	0.813	0.8479	0.9079	0.8769	0.8556	0.9486	0.8997
Emma	13,325	6,302	3,878	0.8196	0.8278	0.8237	0.752	0.8821	0.8119	0.8334	0.8701	0.8513	0.8441	0.7273	0.7814
Wataru	1,675	1,306	1,049	0.8785	0.9162	0.897	0.836	0.9615	0.8944	0.8492	0.9762	0.9083	0.8307	0.9964	0.906
James	9,125	3,270	2,110	0.8957	0.423	0.5746	0.8306	0.4803	0.6086	0.8839	0.9102	0.8968	0.9175	0.8197	0.8658
Kendall	10,735	4,331	3,094	0.8772	0.8423	0.8594	0.8102	0.894	0.85	0.8836	0.917	0.9	0.8654	0.9139	0.889
Liza	10,739	7,440	6,097	0.9313	0.8562	0.8922	0.8838	0.9156	0.8944	0.9539	0.894	0.923	0.9359	0.9144	0.925
Neil	16,487	8,904	5,677	0.8404	0.8894	0.8642	0.4223	0.9031	0.5755	0.8496	0.9729	0.9071	0.7952	0.9753	0.8761
Selena	11,043	6,052	3,634	0.8322	0.7179	0.7709	0.7767	0.7887	0.7826	0.8018	0.7517	0.776	0.7736	0.9282	0.8439
Mai	9,840	2,565	1,330	0.7349	0.7762	0.755	0.698	0.8231	0.7554	0.7022	0.9459	0.806	0.6315	0.988	0.7705
Taylor	13,945	9,444	5,489	0.8085	0.6674	0.7312	0.7473	0.7308	0.7389	0.8889	0.8028	0.8437	0.8141	0.9571	0.8798
Imaizumi*	6,594	2,343	1,664	0.565	0.8873	0.6904	0.5165	0.9543	0.6702	0.6188	0.9148	0.7382	0.6333	0.9312	0.7539
Kitazumi*	21,336	20,364	18,560	0.9024	0.6593	0.762	0.7993	0.7561	0.7771	0.9473	0.3612	0.523	0.9077	0.8818	0.8945
Ogawa*	37,917	29,220	28,032	0.9242	0.7207	0.8098	0.8622	0.7801	0.8191	0.9018	0.771	0.8313	0.8714	0.979	0.9221
Total	208,271	128,200	98,058	0.8512	0.7808	0.8319	0.7693	0.8339	0.7876	0.8561	0.8544	0.8706	0.8364	0.9248	0.8781

Table 4 t-test results (p-value) for Experiment 1

Methods	Precision	Recall	F_1-score
DeepEC - DeepEC+CRF	0.001**	0.000**	0.176
DeepEC - TempEC	0.296	0.035*	0.085
DeepEC - TempEC-HP	0.085	0.000**	0.001**
DeepEC+CRF - TempEC	0.001**	0.308	0.049*
DeepEC+CRF - TempEC-HP	0.006**	0.004**	0.001**
TempEC - TempEC-HP	0.013*	0.039*	0.118

The numbers with * and ** indicates respectively ≤ 0.05 and ≤ 0.01

normalizing with global contrast normalization (GCN). Using the landmarks as a basis, we obtained the coordinates of four corner points that determine the eye region. At this time, we applied a 10% margin to the height and width of the region so that facial landmark detection errors can be accepted.

Deep temporal eye contact detection: Given the images of both eye regions and the 3D facial pose, we implemented our two deep temporal eye contact detection algorithms, as shown in Fig. 5b. The algorithms use ten continuous video frames – the target frame and nine preceding frames – to make predictions.

As shown in Fig. 5b, each of these eye image pairs $\mathbf{I}_{\mathrm{R}}^{t}, \mathbf{I}_{\mathrm{L}}^{t}$ were input to the CNN. This CNN had the same structure as DeepEC with the exception of the last two fully connected layers; namely, it had two streams and six layers— two convolution layers followed by four max pooling layers. The CNN outputs a pair of 512-dimensional feature vectors for each eye image $\mathbf{f}_{\mathrm{R}}(\mathbf{I}_{\mathrm{R}}^{t})$ and $\mathbf{f}_{\mathrm{L}}(\mathbf{I}_{\mathrm{L}}^{t})$.

These feature vectors were input to two separate LSTM networks for the left and right eye images. In the TempEC algorithm, each LSTM accepts 10 vectors corresponding to a series of eye images and outputs one 512-dimensional feature vector. In the TempEC-HP algorithm, a series of 3D vectors that represent 3D head positions are additionally input to LSTM.

However, we found that a naïve LSTM could not perform satisfactorily. To solve this problem, we prepared fully connected layers that had 2048 (512×4) units at the last frame, which accepted the outputs of the left and right DeepEC's and the LSTM's cell state vectors. Because the DeepEC results for the current frames are directly used for eye contact detection, and since temporal inference is also merged to the fully connected layers, we were ultimately able to obtain better results than the naïve implementations of DeepEC and LSTM.

4 Dataset

We prepared two dataset for learning and evaluating the algorithms. The first one was an eye contact video dataset, which we prepared by using publicly available videos from YouTube and our original FPV videos that assume conversational scenarios. The second one was obtained in care learning scenes. We recorded the FPV videos equipped to the caregiver during Humanitude care teaching classes.

4.1 First-person Eye Contact Video Dataset

The first-person eye contact video dataset was used for evaluating the eye contact detection performance. The ground-truth eye contact states (1 or 0) were provided for every video frame in the dataset. We asked three people to annotate the eye-contact states and set the eye contact status as 1 if more than two annotators thought the eye contact was engaged at the frame.

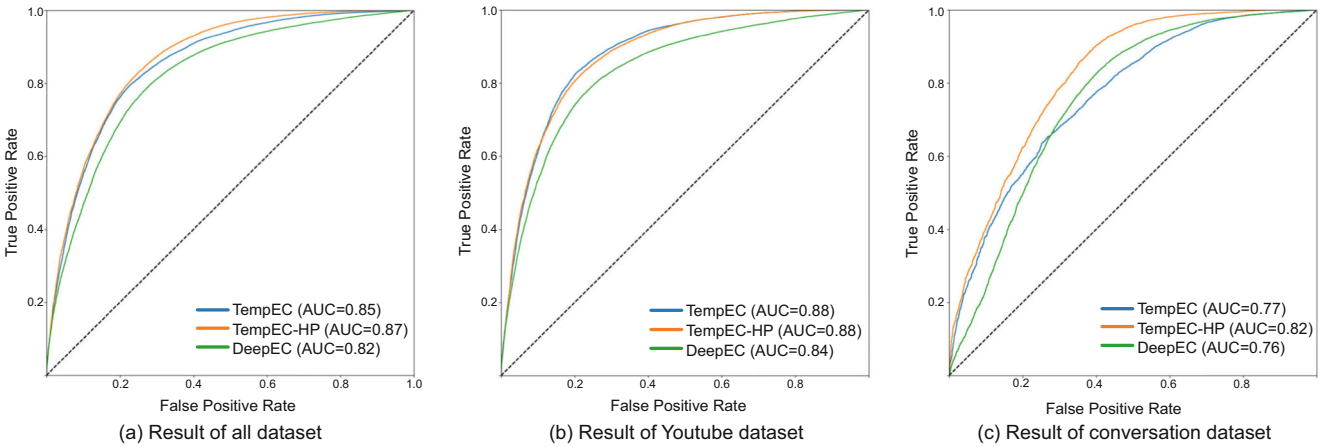


Fig. 8 ROC curves of (**a**) The test results with all dataset, **b** The test results with videos from Youtube, **c** The test results obtained during conversation

Table 5 The result of Experiment 2

Name	Total frames	Face detected	Eye contacted frames	Av. mutual facial distance [mm]	Av. mutual facial pose [deg]		
					r_x	r_y	r_z
Expert A	3442	2148 (62.4%)	1667 (48.4%)	314.3	6.69	7.37	−23.10
Expert B	4852	2070 (42.7%)	1321 (27.2%)	435.1	−9.31	−0.75	−19.98
Middle A	3214	1857 (57.8%)	1087 (33.8%)	406.8	−7.01	3.57	−30.61
Middle B	2659	1248 (46.9%)	1011 (38%)	404.4	2.80	0.46	−37.09
Middle C	2024	1171 (57.9%)	914 (45.2%)	470.6	−7.80	−2.30	−33.80
Middle D	2775	1108 (39.9%)	817 (29.4%)	489.3	−0.26	−4.07	−26.71
Middle E	4506	2547 (56.5%)	1989 (44.1%)	617.8	−5.30	0.91	−31.63
Middle F	3062	2176 (71.1%)	2114 (69%)	232.5	7.91	4.50	−22.04
Middle G	2485	683 (27.5%)	856 (34.4%)	686.5	−7.50	−9.80	−29.99
Novice A	6287	1648 (26.2%)	553 (8.8%)	485.0	−12.98	−7.37	−39.37
Novice B	6168	2675 (43.4%)	810 (13.1%)	422.2	−10.08	−2.08	−40.19
Novice C	4710	1002 (21.3%)	699 (14.8%)	587.2	−4.92	6.03	−25.00

a) First-person eye contact video Youtube (Youtube dataset)

We used 13 videos in which a person talked into a camera from Youtube. We took a consensus of the annotations and made ground-truth data. A list of the videos and their properties is shown in Table 1 and example frames are shown in Fig. 6.

b) First-person eye contact video dataset during conversation (Conversation dataset)

We additionally prepared first-person-view videos while two individuals were conversing. The scenarios were taken in a lab environment in which two participants were talking, where one of them wore a Pivothead Kudu first-person camera. A list of the videos and their properties is shown in Table 2 and Figure 6. We took three video clips from six participants and two test-video clips from two participants.

4.2 FPVs of Care-learning Scenes

To verify the applicability of the proposed algorithms, we prepared first-person videos of a) two Humanitude care experts (instructors), b) seven middle-level Humanitude caregivers and c) three novice Humanitude caregivers as shown in Table 5 and Fig. 7. In all videos, caregivers were equipped with the Pupil Labs first-person camera and performed the same task:

Step 1 Approach the simulated patient while making eye contact,
Step 2 Perform the care, and
Step 3 Leave the care receiver.

5 Experiments

In the first of two experiments we performed, we evaluated the performance of eye contact (bids) algorithms, comparing the two proposed approaches and an existing approach (DeepEC [30]). The second experiment was performed for an actual Humanitude care training scene. In it, we obtained data from a novice caregiver and a Humanitude care expert and compared the results through the use of an unsupervised learning algorithm.

5.1 Experiment 1: Evaluation of eye contact detection performance

As mentioned, we first conducted an experiment to compare the performances of the proposed algorithms and an existing algorithm by using the datasets. One video was chosen for testing and the others were used for learning. We iterated this step for 16 videos and obtained the average performance for them.

The learning of the networks with DeepEC, TempEC and TempEC-HP was conducted as follows. We first computed the bounding rectangles of eyes using the facial landmark points obtained by OpenFace. The obtained eye images were then rescaled such that the image was (60 × 36) pixels. We used static CNN hyper-parameters for all of the experiments. Specifically, the drop-out rate was 0.5 and Leaky ReLU activation function's α was set to 0.01. We used a Nadam optimizer [14] with the learning rate set to 0.001, the decay as 0.004 per epoch and β_1 and β_2 as 0.9 and 0.999, respectively. Learning was performed on a GPU-based workstation (Intel Core i7-7800X CPU 3.50GHz, 128GB RAM, NVIDIA GeForce1080Ti-11G). On average,

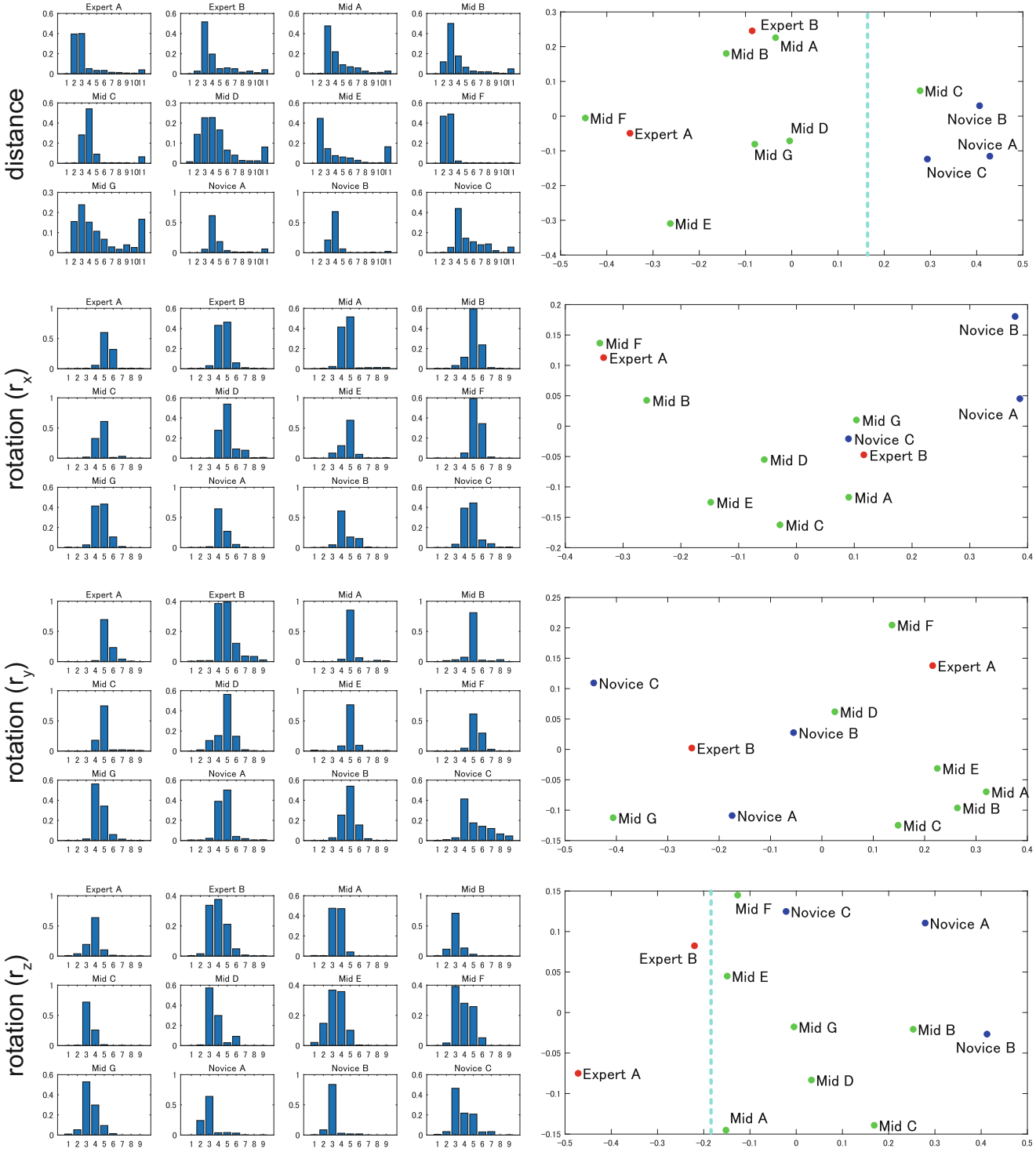


Fig. 9 (Left) Histograms of mutual facial distances and poses (r_x, r_y and r_z). (Right) PCA score results where the x and y axes show the first and second components. There are clear thresholds between novices and others at about $x = 0.16$ for the distance PCA score, and between experts and others at about $x = -0.18$ in r_z PCA score

it took 230 sec, 1900 sec and 1920 sec for learning 1 epoch of DeepEC, TempEC and TempEC+HP, respectively. The learning of CRF of DeepEC-CRF took about 57 sec. We took leave-one-out cross-validation strategy for splitting the learning data and test data.

The results are given in Table 3 and t-test results between four algorithms are given in Table 4. The results show that TempEC+HP algorithms generally outperform other approaches. Namely, the TempEC algorithm thoroughly outperforming DeepEC in precision = 0.8561, recall =

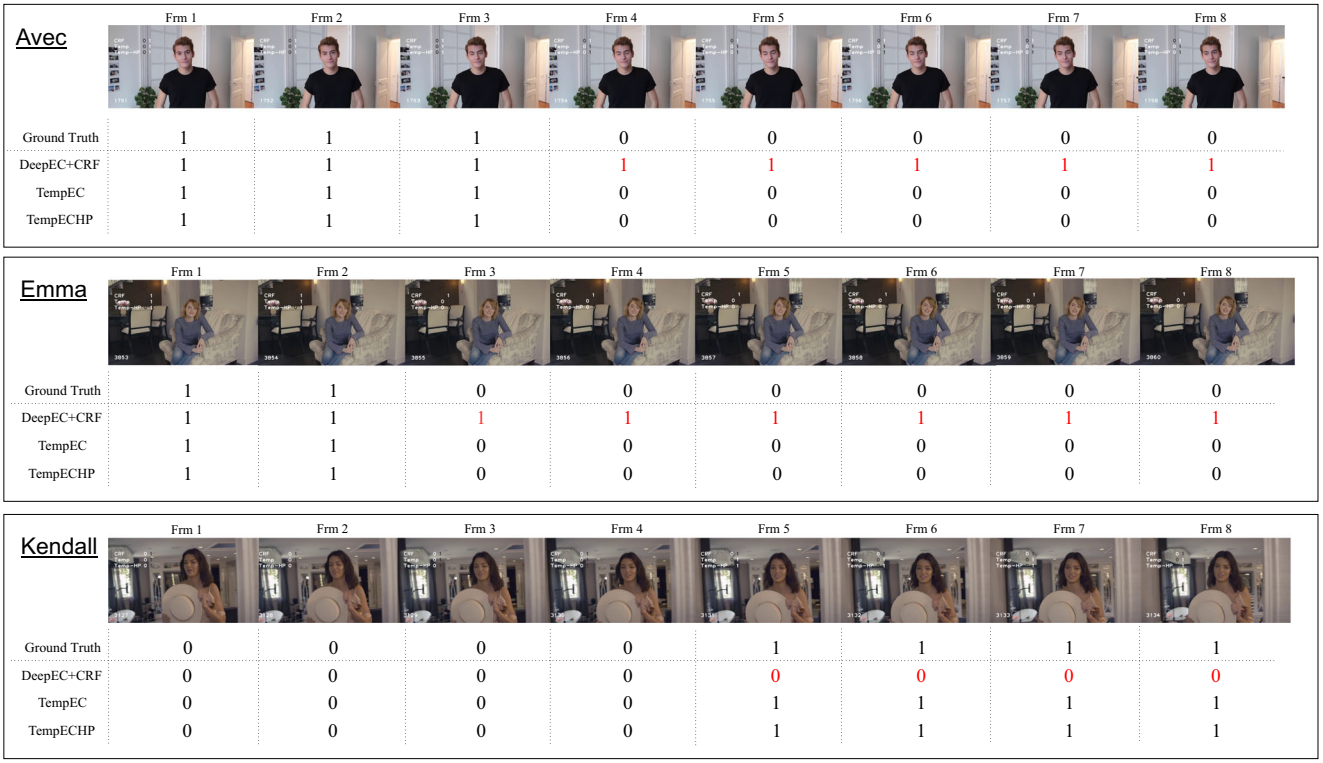


Fig. 10 Example differences among DeepEC+CRF, TempEC and TempEC-HP. Adding CRF to DeepEC tends to 'smoothen' the temporal inference while the other two algorithms correctly estimate the state change

0.8544 and F_1 score = 0.8706. TempEC-HP outperformed TempEC, DeepEC and DeepEC+CRF in recall and F_1-score, with a recall of 0.9248 and F_1 score of 0.8781 on average. Figure 8 shows the area under the curve (AUC) of our algorithm is larger than that of DeepEC. TempEC-HP had the best AUC (0.870) followed by TempEC (AUC = 0.85). With respect to accuracy, TempEC-HP achieved a 25% improvement in miss detection rate, with 0.1751 in comparison to DeepEC's 0.2330. Overall, the recall performances increased by introducing temporal inference (CRF or LSTM). In addition, the proposed methods using LSTM outperforms the method using CRF.

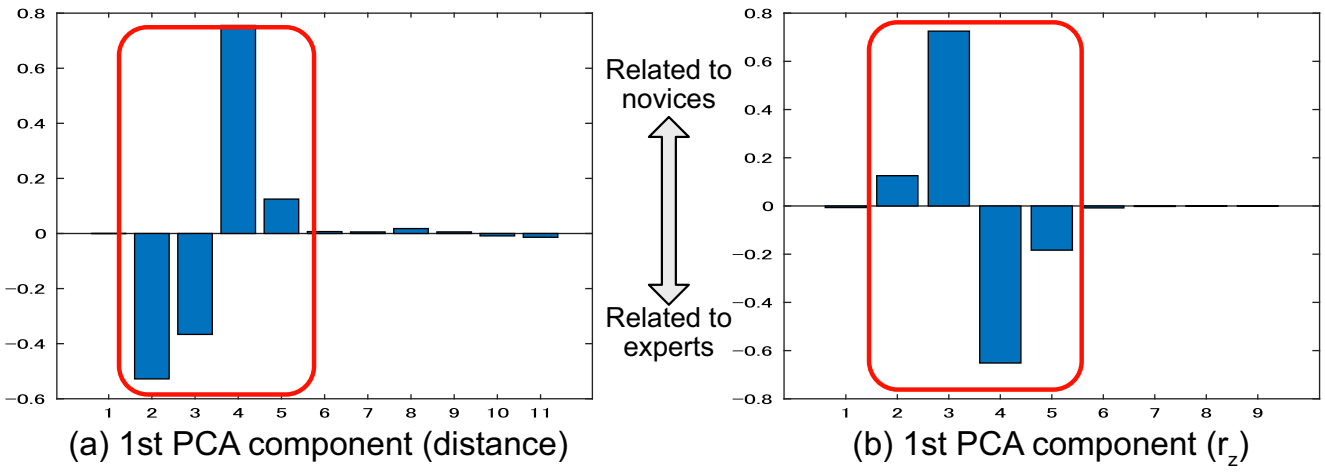


Fig. 11 1st components of PCA of the distance and r_z histograms. Negative elements are related to the experts' behaviors while positive elements are related to the novices' behaviors. In the distance histogram, bins 2 and 3 (0.2 - 0.4 [m]) are related to experts while bins 4 and 5 (0.4 - 0.6 [cm]) are related to novices. Regarding the r_z histogram, bins 4 and 5 (-30 - +10 [deg]) are related to experts while bins 2 and 3 (-70 - -30 [deg]) are related to novices

5.2 Experiment 2: Observation of Face-to-Face Communication Behavior During Care Learning Scenes

In the second experiment we compared the occurrence of Humanitude care skill between novice and expert caregivers using the FPVs of care learning scenes. We obtained the number of eye contact frames, mutual facial distances and poses from a care scene dataset and compared the results.

Analysis and results: The occurrences of eye contact frames, average mutual facial distance and poses (angles) are shown in Table 5 and normalized histograms of each feature are shown in Fig. 9 (left). We applied PCA to the histograms. The resulting PCA scores are shown in Fig. 9 (right), where the x-axis shows the scores of the first component and the y-axis indicates the scores of the second PCA component. From the eye contact rates and PCA analysis results, we were able to clearly distinguish the scores of novices and experts for eye contact rate, mutual facial distance and r_z PCA scores. There were significant differences in eye contact rate between the expert & middle-level and novice groups ($p = 0.0452$), and clear thresholds at about $x = 0.16$ for mutual facial distance and at about $x = -0.18$ for the r_z PCA scores. In the mutual facial distance category, the histograms showed that the expert caregivers and most of the middle-level ones approached the care receiver such that the distance was less than 30 [cm]. In the mutual facial pose category, there were clear dissimilarities in the z-rotation, which is the rotation of the care receiver's face in the FPV image plane (the plane perpendicular to the facial frontal direction) as shown in Fig. 4. Namely, the average and peak z-rotation values of the experts and the middle-level caregivers were located around 0 [deg] while those of novices were much larger.

6 Discussion and Conclusion

In this section, we will discuss the main points we have made for the proposed image-based eye contact detection algorithms and wearable care-skill evaluation system and draw conclusions from them.

Image-based eye contact detection: We developed eye-contact-detection algorithms that use temporal features as well as static image features. Our algorithms show better performance for various types of datasets. They combine CNNs and LSTM and successfully learned both static features and temporal dependence. In experiments, the proposed TempEC and TempEC-HP algorithms outperformed the DeepEC algorithm. In particular, TempEC-HP achieved a 25% improvement in the miss-detection rate over existing algorithms.

In a preliminary experiment, we found that a simple concatenation of CNNs and LSTM was not effective. We concluded that such a primitive combination was not suitable for learning both static and temporal features at the same time. Thus, in the final estimation step we introduced a skip connection that jumps over the LSTM networks and directly links the CNN outputs to the final fully connected layers. This structure improved our algorithms' performance, as the experiment 1 results showed. They showed that the skip connection enables the algorithms to successfully learn both static and temporal features at the same time.

Surprisingly, some tests showed that TempEC performed better than TempEC-HP. This was contrary to what we had expected because we believed the facial pose information in TempEC-HP would help in detecting eye contact for various face directions. However, these results do not indicate that facial pose information is useless. In our TempEC-HP algorithm, 3D facial pose estimation is based on facial landmarks, the detection of which is mostly accurate but has a certain degree of error. This error is not significant, which is why it is not a problem when used to obtain eye regions. However, in facial pose estimation, such a small error sometimes causes a large incorrect gap between two contiguous frames. The facial pose of two adjacent frames should be close because a human's face cannot move very much in a short time (namely, 0.03 sec because the videos were recorded at 30 fps). Due to this problem, facial pose estimation is occasionally not sufficiently reliable, which causes TempEC-HP to perform poorly. Hence, the performance of TempEC-HP can be improved by using a more accurate facial detection or facial pose estimation algorithm.

Another notable finding was that introducing temporal inference increased the algorithms' recall performance, which means the temporal information contributed to 'overlooked' effects of eye contact. Our experiments showed that adding a conditional random field (CRF) to the DeepEC algorithm could not improve its results. Several examples (Fig. 10) showed that CRF tends to "smoothen" temporal inference in DeepEC, which may help to avoid the 'jittering' effects of single frame estimations but does not solve the temporal inference problem fundamentally. Thus, we believe our current algorithms, which combine the internal states of single frame recognition and LSTM, are a better solution.

Our results show that our algorithms enable excellent eye contact detection performance to be achieved. They also show the potential of temporal learning of eye behavior, with which we can evaluate the care skills of caregivers.

Evaluation of wearable care-skill estimation system: Unsupervised analysis results of mutual facial distances and

facial poses enabled us to find significant differences between novices, middle-level and expert Humanitude caregivers. Specially, we found a clear threshold in eye contact frequency and PCA scores of facial distance and r_z-rotation histograms, which indicate that the important skills in Humanitude tender-care are related to a) frequent eye contact, b) a nearest mutual facial distance of less than 30 [cm] and c) mutual facial poses being in the same direction. This can also be seen from the 1st PCA components of the distance and r_z histograms (Fig. 11). In the distance histogram, bins 2 and 3 (0.2 - 0.4 [m]) are related to the experts while bins 4 and 5 (0.4 - 0.6 [m]) are related to the novices. For the r_z histogram, bins 4 and 5 (-30 - +10 [deg]) are related to the experts while bins 2 and 3 (-70 - -30 [deg]) are related to the novices. These skills are a part of Humanitude gaze skill: caregivers should communicate to the care receivers while keeping eye contact from a close distance and possessing the same facial angles of the care receiver's face. This is based on the idea of Humanitude care methodology that all behaviors are considered to imply non-verbal messages. To have the eye contact straight in front of the care receiver expresses the fairness, and the distance between caregivers and care receivers reflects their friendliness. The study results show that the experts expressed fairness and friendliness much more than the novices. This skill is a core skill with which to establish a good relationship that leads to high-quality care.

Open issues and future work: Though our analyses can quantize and visualize the Humanitude care communication skills, several open issues remain to increase the analysis quality, as we ascertained from the responses of Humanitude experts. The first point is the face detection stability of OpenFace. Specifically, OpenFace cannot detect care receivers' faces when x or y rotations are quite large (e.g., looks at the size of the faces). The second point is the temporal analysis. The estimation of facial poses and distances is currently performed frame-by-frame and TempEC considers only 1/3 seconds as the temporal duration. However, it has been reported that the duration of eye contact is about three seconds in typical communication scenes [6] and that a longer duration of mutual gaze is often effective in communicating with a dementia patient [3]. Thus, temporal inference using a longer duration can be expected to be effective in care-skill evaluation as well.

The tender-care concept involves multi-modal skills including gazing, speaking and touching. As the initial step in computational care communication analysis, we treated face-to-face communication skills. We are currently developing a method for detecting and analyzing voice signals and sensing touch behaviors through the use of wearable contact sensors or vision analysis. We believe that our findings for tender-care skills and systems for obtaining care skills will prove to be important and usable, not only for increasing the skills of caregivers but also for designing and evaluating care robots' behaviors.

Acknowledgements All the experiments were conducted in compliance with the protocol which was reviewed and approved by the ethical committee of Unit for Advanced Studies of the Human Mind, Kyoto University (Permit Number: 30-P-4). This work was supported by JST CREST Grant Number JPMJCR17A5 and JSPS KAKENHI 17H01779, Japan.

References

1. Acton, G.J., Kang, J.: Interventions to reduce the burden of caregiving for an adult with dementia: a meta-analysis. Res. Nurs. Health **24**(5), 349–360 (2001)
2. Adelman, R.D., Tmanova, L.L., Delgado, D., Dion, S., Lachs, M.S.: Caregiver burden: a clinical review. Jama **311**(10), 1052–1060 (2014)
3. Alzheimer's Society: Factsheet: Communicating. https://www.alzheimers.org.uk/site/scripts/documents_info.php?documentID=130, [Online; accessed 18-Nov-2016] (2016)
4. Baltrusaitis, T., Robinson, P., Morency, L.P.: (2016) OpenFace: An Open source facial behavior analysis toolkit. 2016 IEEE Winter Conference on Applications of Computer Vision WACV. https://doi.org/10.1109/WACV.2016.7477553 (2016)
5. Bertasius, G., Park, H.S., Stella, X.Y., Shi, J.: Am I a Baller? Basketball Performance Assessment from First-Person Videos. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2196–2204. IEEE (2017)
6. Binetti, N., Harrison, C., Coutrot, A., Johnston, A., Mareschal, I.: Pupil dilation as an index of preferred mutual gaze duration. Royal Society Open Science 3(7). https://doi.org/10.1098/rsos.160086, http://rsos.royalsocietypublishing.org/content/3/7/160086.full.pdf (2016)
7. Biquand, S., Zittel, B.: Care giving and nursing, work conditions and humanitude®. Work **41**(Supplement 1), 1828–1831 (2012)
8. Boseley, S.: Dementia research funding to more than double to $66m by 2015. The Guardian (2012)
9. Campbell, L.C., Keefe, F.J., Scipio, C., McKee, D.C., Edwards, C.L., Herman, S.H., Johnson, L.E., Colvin, O.M., McBride, C.M., Donatucci, C.: Facilitating research participation and improving quality of life for african american prostate cancer survivors and their intimate partners: a pilot study of telephone-based coping skills training. Cancer: Interdiscip. Int. J. Amer. Cancer Soc. **109**(S2), 414–424 (2007)
10. Casado, B., Sacco, P.: Correlates of caregiver burden among family caregivers of older korean americans. J. Gerontol. Ser. B: Psychol. Sci. Soc. Sci. **67**(3), 331–336 (2011)

11. Chong, E., Chanda, K., Ye, Z., Southerland, A., Ruiz, N., Jones, R.M., Rozga, A., Rehg, J.M.: Detecting gaze towards eyes in natural social interactions and its use in child assessment. Proc. ACM Interact. Mob. Wearab. Ubiquit. Technol. **1**(3), 43 (2017)
12. Clyburn, L.D., Stones, M.J., Hadjistavropoulos, T., Tuokko, H., et al.: Predicting caregiver burden and depression in alzheimer's disease. J. Gerontol. Ser. B **55**(1), S2–S13 (2000)
13. Coon, D.W., Thompson, L., Steffen, A., Sorocco, K., Gallagher-Thompson, D.: Anger and depression management: psychoeducational skill training interventions for women caregivers of a relative with dementia. Gerontologist **43**(5), 678–689 (2003)
14. Dozat, T.: Incorporating nesterov momentum into adam. http://cs229.stanford.edu/proj2015/054_report.pdf, [Online; accessed 25-Aug-2018] (2016)
15. Dunkin, J.J., Anderson-Hanley, C.: Dementia caregiver burden: a review of the literature and guidelines for assessment and intervention. Neurology **51**(1 Suppl 1), S53–S60 (1998)
16. Etters, L., Goodall, D., Harrison, B.E.: Caregiver burden among dementia patient caregivers: a review of the literature. J. Am. Acad. Nurse Pract. **20**(8), 423–428 (2008)
17. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social Interactions: a First-Person Perspective. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1226–1233. IEEE (2012)
18. Gineste, Y., Pellissier, J.: Humanitude: comprendre la vieillesse, prendre soin des hommes vieux. A. Colin (2007)
19. Given, B., Sherwood, P.R., Given, C.W.: What knowledge and skills do caregivers need? J. Soc. Work. Educ. **44**(sup3), 115–123 (2008)
20. Honda, M., Ito, M., Ishikawa, S., Takebayashi, Y., Tierney, L.: Reduction of behavioral psychological symptoms of dementia by multimodal comprehensive care for vulnerable geriatric patients in an acute care hospital: a case series. Case reports in medicine 2016 (2016)
21. Ishikawa, S., Ito, M., Honda, M., Takebayashi, Y.: The skill representation of a multimodal communication care method for people with dementia. JJAP Conf. Proc. **011616**, 4 (2016)
22. Ito, M., Honda, M.: An examination of the influence of humanitude caregiving on the behavior of older adults with dementia in japan. In: Proceedings of the 8th International Association of Gerontology and Geriatrics European Region Congress (2015)
23. Kim, H., Chang, M., Rose, K., Kim, S.: Predictors of caregiver burden in caregivers of individuals with dementia. J. Adv. Nurs. **68**(4), 846–855 (2012)
24. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye Tracking for Everyone. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
25. Larson, E.B., Yaffe, K., Langa, K.M.: New insights into the dementia epidemic. Engl. J. Med. **369**(24), 2275–2277 (2013)
26. Law, H., Ghani, K., Deng, J.: Surgeon technical skill assessment using computer vision based analysis. In: Machine Learning for Healthcare Conference, pp. 88–99 (2017)
27. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Gaze estimation from eye appearance: a head pose-free method via eye image synthesis. IEEE Trans. Image Process. **24**(11), 3680–3693 (2015)
28. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1894–1903 (2016)
29. Ministry of Health, Labour and Welfare, Japan: Supply and demand estimation for nursing care personnel for 2025. https://www.mhlw.go.jp/stf/houdou/0000088998.html, [Online; accessed 11-Aug-2018] (2015)
30. Mitsuzumi, Y., Nakazawa, A., Nishida, T.: Deep Eye Contact Detector: Robust Eye Contact Bid Detection Using Convolutional Neural Network. In: Proceedings of the British Machine Vision Conference (BMVC) (2017)
31. Northouse, L.L., Katapodi, M.C., Song, L., Zhang, L., Mood, D.W.: Interventions with family caregivers of cancer patients: meta-analysis of randomized trials. CA: Cancer J. Clin. **60**(5), 317–339 (2010)
32. Ostwald, S.K., Hepburn, K.W., Caron, W., Burns, T., Mantell, R.: Reducing caregiver burden: a randomized psychoeducational intervention for caregivers of persons with dementia. Gerontologist **39**(3), 299–309 (1999)
33. Papastavrou, E., Kalokerinou, A., Papacostas, S.S., Tsangari, H., Sourtzi, P.: Caring for a relative with dementia: family caregiver burden. J. Adv. Nurs. **58**(5), 446–457 (2007)
34. Petric, F., Miklić, D., kovačić, Z.: Probabilistic eye contact detection for the robot-assisted asd diagnostic protocol. In: Lončarić S., Cupec, R. (eds.) Proceedings of the Croatian Compter Vision Workshop, Year 4, Center of Excellence for Computer Vision, pp. 3–8. University of Zagreb, Osijek (2016)
35. Povithead: Pivothead KUDU. http://www.pivothead.com/, [Online; accessed 29-Aug-2016] (2016)
36. Pupil Labs: Pupil labs camera system. https://pupil-labs.com/pupil/, [Online; accessed 25-Aug-2018] (2018)
37. Singh, S., Arora, C., Jawahar, C.: First person action recognition using deep learned descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2620–2628 (2016)
38. Smith BA, Yin Q, Feiner SK, Nayar SK: Gaze locking: passive eye contact detection for human-object interaction. In: Proceedings of the 26th annual ACM symposium on User interface software and technology, pp. 271–280. ACM (2013)
39. Soo Park, H., Shi, J., et al.: Force from motion: decoding physical sensation in a first person video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3834–3842 (2016)
40. Su, S., Hong, J.P., Shi, J., Park, H.S.: Predicting Behaviors of Basketball Players from First Person Videos. In: CVPR, vol. 2, pp. 3 (2017)
41. Win, K.K., Chong, M.S., Ali, N., Chan, M., Lim, W.S.: Burden among family caregivers of dementia in the oldest-old: an exploratory study. Front. Med. **4**, 205 (2017)
42. World Health Organization, et al.: Dementia: Fact sheet N 362 (2012)
43. Ye, Z., Li, Y., Fathi, A., Han, Y., Rozga, A., Abowd, G.D., Rehg, J.M.: Detecting eye contact using wearable eye-tracking glasses. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 699–704. ACM (2012a)
44. Ye, Z., Li, Y., Fathi, A., Han, Y., Rozga, A., Abowd, G.D., Rehg, J.M.: Detecting eye contact using wearable eye-tracking glasses. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp'12, pp 699–704. ACM, New York, https://doi.org/10.1145/2370216.2370368 (2012b)
45. Ye, Z., Li, Y., Liu, Y., Bridges, C., Rozga, A., Rehg, J.M.: Detecting Bids for Eye Contact Using a Wearable Camera. In: 2015 11Th IEEE International Conference and Workshops On Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–8. IEEE (2015)
46. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4511–4520 (2015)
47. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze

estimation. arXiv:1611.08860, https://perceptual.mpi-inf.mpg.de/wp-content/blogs.dir/12/files/2016/11/zhang16_arxiv.pdf (2016)
48. Zhang, X., Sugano, Y., Bulling, A.: Everyday eye contact detection using unsupervised gaze target discovery. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, pp. 193–203. ACM (2017)

Atsushi Nakazawa is an associate professor in the Department of Informatics at the Kyoto University. He received his doctorate from the Osaka University in 2001 in Systems Engineering. Afterward, he worked in Institute of Industrial Science, University of Tokyo and then in Cybermedia Center, Osaka University. From 2013, he joined the Kyoto University. From Oct. 2017, he becomes a program investigator (PI) of the JST CREST project "Computational and cognitive neuroscientific approaches for understanding the tender care". His research interests are in human behavior/mental analysis using computer vision, eye tracking, eye imaging and motion capture systems. Dr. Nakazawa got the best paper award in International Conf. on Virtual Systems & Multimedia (VSMM2004) and Japan Robotics Society (RSJ).

Yu Mitsuzumi is an employee at NTT Communication Laboratories. He received Master of Infromatics from Kyoto University in 2019. His research interests include computer vision, machine learning and deep neural nets.

Yuki Watanabe is a master course student in the Department of Informatics at the Kyoto University. He received B.E from Faculty of Engineering, Kyoto University in 2018, then he joined the Department of Informatics, Kyoto University. His research interests include computer vision, deep neural nets and computer animation.

Ryo Kurazume is a Professor at the Graduate School of Information Science and Electrical Engineering, Kyushu University. He received his M.Eng. and PhD in Mechanical Engineering from Tokyo Institute of Technology in 1989 and 1998. He was a director of the Robotics Society of Japan (RSJ) from 2009 to 2011 and 2014 to 2015, and a director of the Society of Instrument and Control Engineers (SICE) from 2013 to 2015. Currently he is a chairman of the Japan Society of Mechanical Engineers (JSME) Robotics and Mechatronics Division. He received JSME Robotics and Mechatronics Academic Achievement Award in 2012, RSJ Fellow in 2016, SICE System Integration Division Academic Achievement Award in 2017, and JSME Fellow in 2018. His current research interests include legged robot control, computer vision, multiple mobile robots, service robots, medical imaging, and biometrics.

Sakiko Yoshikawa is a professor of Psychology, Kokoro Research Center, Kyoto University. She received her Ph.D. from Kyoto University in Education. After working at the Psychology Department of Otemon-Gakuin University, she joined the Kyoto University in 1997. From 2007 to 2018 she worked as a first director of Kokoro Research Center that promotes interdisciplinary research on the human mind. Her recent research interests include basic mechanisms of face-to-face communication and recognition of social signals from faces. She received a Distinguished Paper Award from the Japan Psychological Association and a Best Paper Award of JPA annual convention in 2017.

Miwako Honda is director of Geriatric Research Center, National Hospital Organization Tokyo Medical Center. After graduating the Department of Medicine, Tsukuba University at 1993, she became the clinical fellow of general medicine, Kameda General Hospital, then worked as the Clinical fellow of infectious disease, International Medical Center of Japan during 1997-1998, followed by the Resident of internal medicine, Thomas Jefferson University, Philadelphia, and clinical fellow of geriatrics, at Cornel University. From 2002 to 2011, she joined International Medical Center of Japan. From 2011, she worked in the current position. She wrote more than five books regarding geriatric care and Humanitude.