# Automated Assessment of Nonverbal Behavior of the Patient during Conversation with the Healthcare Worker Using a Remote Camera

Takashi Imabuchi, Oky Dicky Ardiansyah Prima, Hisayoshi Ito
Faculty of Software and Information Science
Iwate Prefectural University
Takizawa, Iwate, Japan
e-mail: g236o001@s.iwate-pu.ac.jp, {prima, hito}@iwate-pu.ac.jp

***Abstract*—The importance of nonverbal behavior between the healthcare worker and the patient has led many studies to focus on quantitatively evaluating this behavior. Those existing studies utilize the Kinect 3D sensor to visualize the nonverbal behaviors for a range of healthcare applications. In this study, we propose a framework to automatically assess the nonverbal behavior of the patient during conversation with the healthcare worker. Instead of using the Kinect, we detect the skeleton information of the targeted body from a single camera using an advanced computer vision approach. The proposed framework collects data consisting of facial expression, eye movement, and head nodding and analyzes this data to assess the quality of the nonverbal communication.**

***Keywords-Nonverbal communication; Eye tracking; Facial expression; Pose estimation.***

## I. INTRODUCTION

In the health sector, good communication between healthcare workers and patients is important to improve the quality of care and to promote patient-centered healthcare. These communications are effective to improve the patient satisfaction [1]. Assessing the quality of communication in a healthcare setting is a difficult task in today's hospitals. The assessment includes both verbal and nonverbal communications [2]. While the verbal communication directly conveys the patient's needs, the nonverbal communication represents communicative acts which may be even more important than the matter under verbal discussion. The nonverbal communication consists of a variety of non-words information such as gestures, physical features and paralanguages. The nonverbal interaction in the healthcare sector may represent as much as 65 percent of the hidden thoughts and emotions of the patient [3]. To establish a good relationship of trust with a patient, the healthcare worker is required to pay close attention to this interaction. This skill would be essential for all healthcare workers.

There are basically five types of nonverbal behaviors (body language) related to movement of the body [4]: emblems, regulators, illustrators, affective display and adaptors. During a communication with patients, skilled healthcare workers (health professionals) use four elements of the body language: body posture, eye contact, facial expression, and gesture. Hence, the quality of communication can be observed from the appearances of the visual components of their faces in accordance with the movements of other parts of their bodies.

Information and Communications Technologies (ICTs) play an important role to automatically assess the nonverbal communication. Depth camera sensing enables the observation of human pose in three dimensions. The Kinect 3D sensor shows adequate performance for a range of healthcare imaging applications [5]. It provides not only the changes of the body posture, but also the pose (skeleton) information of the targeted body, which can be used to estimate gestures. Facial image processing enables the detection of head movements and the subtle changes of facial expression. Using Internet-of-Things (IoT) [6], these tasks can be processed with a low investment cost.

The major challenge of implementing camera sensing into the healthcare sector is determining where to place the camera so that it can capture both the healthcare worker and the patient at the same time, with adequate image resolution. Since the camera's existence must not disturb the process of care, it should be placed around corners or on the ceiling in the hospital room. However, placing the camera far away from the target will decrease its ability to reveal the detail structures of the target's body. An attempt using a combination of multiple cameras, depth sensing, and fiducial markers has been conducted to measure hand movements of the healthcare worker with respect to the patient's bed [7]. While this approach enables to estimate the potential hand movements at the bedside, the overall system is complicated for a practical use in a hospital.

Recent computer vision applications enable the detection of 2D human poses from a single image [8]. Furthermore, the 3D human pose can be estimated by using human pose libraries taken from motion capture devices as a reference [9]. Unlike the Kinect which needs a proper distance setting to the targets, these approaches are more flexible. The skeleton information can be derived for the targeted bodies located more than 5m away from the camera. Based on this information, the detection of the head and the visual components of the face can be analyzed easier without requiring face detections as in the conventional image processing [10].

In this study, we propose a framework to automatically assess the nonverbal behavior of the patient during conversation with the healthcare worker from a single camera. The proposed framework starts analysis at the time it detects a skeleton other than that of the patient. Using the part of the skeleton that consists of the bones of the head (axial skeleton), the face area of the patient is determined and up-sampled to

 ISBN: 978-1-61208-618-7 

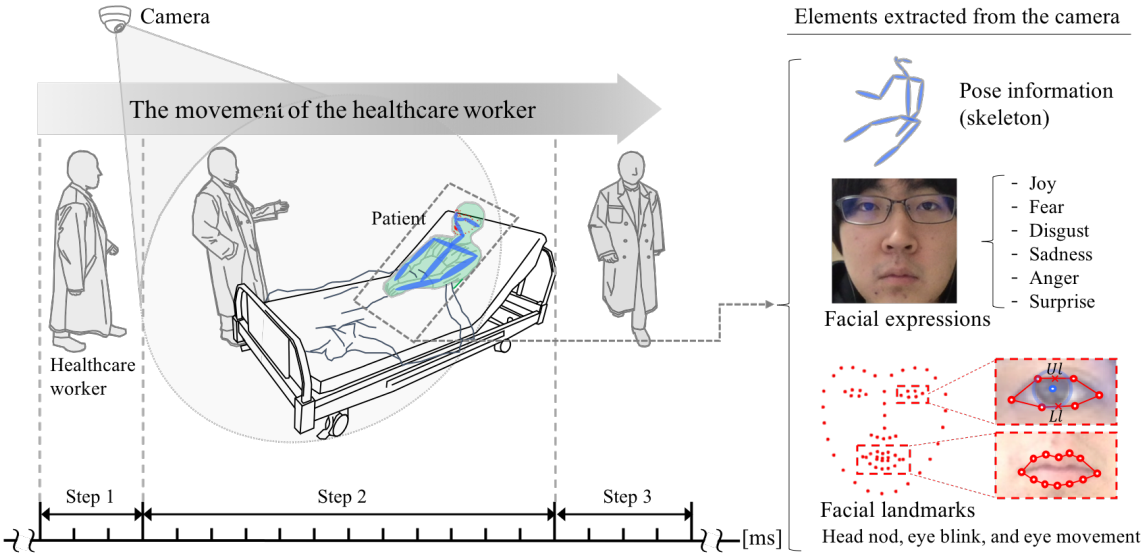



Figure 1. The proposed framework to assess nonverbal behaviors from a single camera.

observe the facial information in detail. Information consisting of facial expression, eye movement, and head nodding is statistically analyzed against the pre-calculated learning data to show how well the nonverbal communication is being constructed. For further application, we also discuss other possibilities on the uses of skeleton information to improve the quality of healthcare service.

This paper is organized as follows. Section II describes related works in healthcare sector. Section III introduces our approach to assess nonverbal behavior using a remote camera. Section IV describes our experimental setting and its results. Finally, Section V presents our conclusions and future works.

## II. Related Work

Traditional methods to measure nonverbal behavior rely on manual coding system. This measurement involves the duration, the response latency, and the inter-response time during a behavior [11]. Since there are a lot of ambiguities on judging a particular behavior, many observers tend to perform in an inconsistent manner, thus degrading the quality of the measurements. Some studies have developed automated detection of the nonverbal behavior based on velocities changes derived from gyroscopes [12], video-based motion analyses, and a multi-modal sensor consisting of red, green and blue (RGB), depth, and audio data. For a more complex and dynamic behavior, high-level sensing is required to assess [13].

In the health sector, many studies have been conducted to assess the nonverbal behavior between the healthcare worker and the patient. These assessments covered conversational agents, securities, healthcare settings and medical acts in the hospital room. The Kinect 3D sensor has been used to extract and analyze head pose and hand gestures of the healthcare worker. The typical hand gestures can be classified using machine learning algorithm [14]. The depth sensing in the Kinect is also shown to be adequate to monitor the respiratory rate of the patient [15].

Although the Kinect 3D sensor provides pose information to measure the high-level behavior, it has drawbacks for a practical use in a hospital. The measurement of nonverbal behavior in the hospital setting requires a sensor to capture a wider area where the healthcare worker and the patient are located. Moreover, operating the Kinect and its processing computer will incur a high investment cost.

The advances brought about by machine learning and Artificial Intelligence (AI) have contributed to solve the problem on localizing anatomical key points to find body parts inside an image. The "OpenPose" library enables realtime multi-person 2D pose estimation from a single camera [8]. This library applies a bottom-up approach by encoding the location and rotation of limbs over the image domain to allow a greedy parse to connect the detected body parts. Thus, the library can detect human pose not only from the front but also from the back. Martinez et al. [16] used deep neural networks to map between 2D and 3D poses. The standard protocol of Human3.6M was used to normalize the pose data [17]. Human pose estimation from a single camera has become a potential alternative to the Kinect.

## III. Automated Assessment of Nonverbal Behavior

### A. Nonverbal Behavior Measurement

Here, we assume a situation in a hospital where the nonverbal communication between the healthcare worker and the patient mostly occurs. Figure 1 shows three steps for our proposed method to assess the nonverbal behavior. A single camera is located facing the patient's bed. For each step, we use the "OpenPose" library to detect human pose found in the camera image. A pre-defined region of interest (ROI) is applied to the image to assign the location of the bed.

 ISBN: 978-1-61208-618-7 

*Step 1: Detecting the presence of the healthcare worker*

When the healthcare worker is heading towards the patient's bed, his posture will be detected by the time he stands in front of the bed. "OpenPose" detects the existence of human pose inside ROI. It enables to distinguish the basic body postures: standing, sitting (reclining), and laying. The patient is recognized to have laying or reclining postures while the health worker to have a standing posture. Once the healthcare worker is detected when he enters the ROI, he will be tracked until he leaves the area.

*Step 2: Extracting nonverbal behavior during conversation*

Nonverbal behaviors of the patient during speaking and listening are extracted separately. The periods of speaking and listening are defined by utterances, occurrences of changes in shapes of the patient's mouth detected from the camera image [18]. To calculate the facial shape, 68 facial landmarks are estimated using "Dlib" library [10]. The extracted nonverbal behaviors include head nod, eye blink, eye movement, and facial expressions. While the calculations of head nod, eye blink, and eye movement make use of facial landmarks, the facial expression is estimated using "Affectiva" library, a state-of-the-art emotion recognition using Deep Learning [6].

*Step 3: Detecting the end of conversation*

The observation will finish when the healthcare worker leaves the room as detected by the absence of his posture in the camera image.

The details of methods for detecting utterances, head nod, eye blink, eye movement, and facial expressions are as follows.

*1) Utterance:* Utterance is measured by detecting the relative location changes of the boundary of the mouth represented by 12 landmarks. These changes are observed as the standard deviation of those landmarks within a duration. Utterance is quantized by

$$Utterance = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(L_{nj_{(x,y)}} - \overline{L_n}\right)^2}, \qquad (1)$$

where, $n$ is the number of image frames, $m$ is the number of landmarks, and $L_{nj}$ is $j$-th landmark point ($x$, $y$) at $n$-th frame.

*2) Head nod:* Head nod is calculated by detecting the changes of head-pose in pitch angle. This calculation is done by fitting six 3D anthropometric points to the associated facial landmarks. The changes of pitch angles are calculated by

$$Head\ nod = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\varphi_n - \bar{\varphi})^2}, \qquad (2)$$

where, $n$ is the number of image frames and $\varphi_n$ is the *pitch* angle at $n$-th frame.

*3) Eye blink:* The eye blink is considered as the degree of eye openness which can be measured from the relative ordinate changes of the midpoints of upper and lower eyelids represented by previously calculated landmarks by

$$Eye\ openness = \sqrt{(Ul_y - Ll_y)^2}, \qquad (3)$$

where, $UL_y$ and $Ul_y$ are ordinates of midpoints of upper and lower eyelids, respectively.

*4) Eye movement:* We measure the eye movement by tracking the iris inside the region of eye in the image. The region of eye is extracted from the area of image surrounded by eye landmarks. The tracking method is based on eye tracking from visible-spectrum camera [19]. This method calculates the gaze direction with head pose compensation. Let $(x, y)$ the coordinate of the iris center, horizontal and vertical eye movements $(E_{dx}, E_{dy})$ in n-number of frames are calculated by

$$E_{dx} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ex_i - \bar{e})^2} \quad , \quad E_{dy} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ex_i - \bar{e})^2}. \quad (4)$$

*5) Facial expression:* The "Affectiva" library estimates six fundamental expressions: joy, fear, disgust, sadness, anger, and surprise from the patient face in the image. The occurrence of each expression is represented as a probabilistic value. Among these expressions, we use only the accumulated occurrence of "joy" for our purpose, as described later on this paper.

### *B. Assessment of nonverbal behaviors*

Our strategy to assess nonverbal behaviors is to create two scenarios of communication scenes between the healthcare worker and the patient. The first scenario is "high trust" where a good quality relationship between the healthcare worker and the patient has been established. Contrary, the second is "low trust" scenario. While the "high trust" will lead to greater emotional stability that facilitates acceptance and openness of expression, "low trust" will result in less accurate communication. We believe that these differences can be revealed from the nonverbal behaviors extracted in this study.

Figure 2 shows our preliminary results of nonverbal behaviors of the two scenarios. Both manual and automatic coding of utterances are provided to show our utterance detection is effective to define the cognitive tasks: listening and speaking. Nonverbal behaviors for each scenario are assessed as follows.

*1) Repetition of head nod, eye blink, and eye movement:* Here, to distinguish communicative from noncommunicative types head nod, eye blink, and eye movement, we count the repetition for those behaviors in a duration of time and in each cognitive task: listening or speaking. The number of repetitions is computed by counting the number of values

 ISBN: 978-1-61208-618-7 

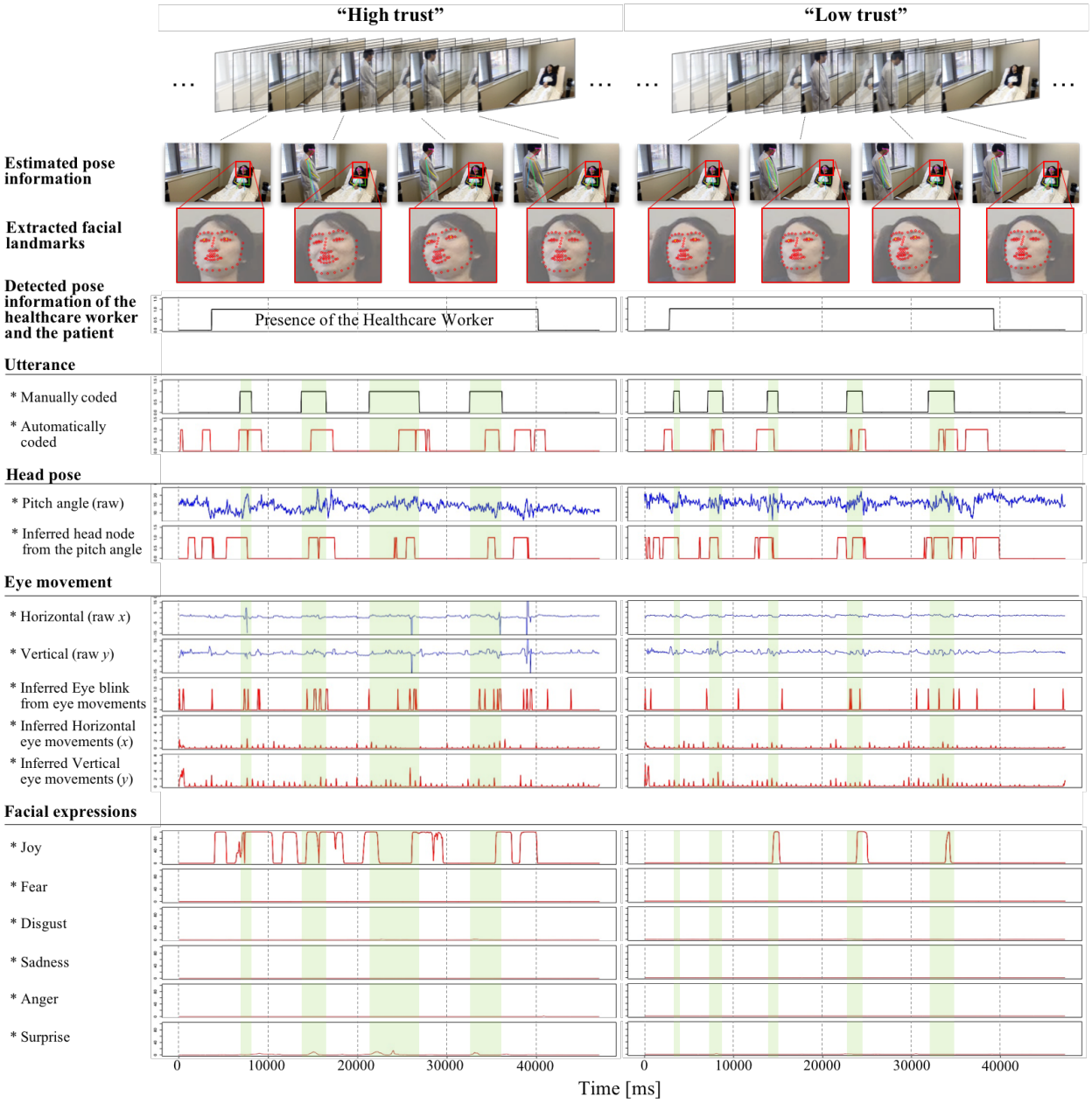



Figure 2. Preliminary results of nonverbal behaviors of the two scenarios.

above a given threshold in (1), (2), (3), and (4). Different patterns for head nod, eye blink, and eye movement during listening and speaking can be observed, as shown in the filled and unfilled area in Figure 2.

*2) Accumulated occurrences of facial expressions:* Since facial expression will have a range of amplitude, occurrence and duration, we calculate the probability accumulation for each facial expression in each cognitive task. Figure 2 shows different patterns of six fundamental expressions during speaking and listening. Since the fluctuations of "joy" show high responses, for further analyzes, only "joy" will be calculated to represent the facial expression changes in this study.

## IV. Experiments and Results

Experiments were conducted by an experimenter and three participants as patients to communicate in "high and low trust" scenarios. Both scenarios were conducted for a duration of 40 seconds. The experimenter talked with the participant, in the way to cause the participants to alter their behavior to match each scenario. All scenes were recorded using a 60fps single camera with 1280×720 pixels resolution.

 ISBN: 978-1-61208-618-7 

TABLE I. FREQUENCIES OF HEAD NOD, EYE BLINK, AND EYE MOVEMENT IN TWO SCENARIOS

| Nonverbal Behaviors | Speaking [Hz] | | | | | | Listening [Hz] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Subject #1 | | Subject #2 | | Subject #3 | | Subject #1 | | Subject #2 | | Subject #3 | |
| | LT | HT | LT | HT | LT | HT | LT | HT | LT | HT | LT | HT |
| **Head nod** | 0.4 | 0.2 | 0.0 | 0.1 | 0.6 | 0.5 | 0.2 | 0.4 | 0.1 | 0.3 | 0.3 | 0.1 |
| **Eye blink** | 0.5 | 0.6 | 0.2 | 0.5 | 0.6 | 1.1 | 0.7 | 0.4 | 0.4 | 0.3 | 0.2 | 0.1 |
| **Eye movement:** | | | | | | | | | | | | |
| **-Horizontal (*x*)** | 1.1 | 1.0 | 0.6 | 1.0 | 1.0 | 1.0 | 0.9 | 0.4 | 0.3 | 0.7 | 0.9 | 0.8 |
| **-Vertical (*y*)** | 1.3 | 1.1 | 1.4 | 1.3 | 1.8 | 1.3 | 0.9 | 1.2 | 0.5 | 0.9 | 1.2 | 0.8 |

LT: "Low trust", HT: "High trust"

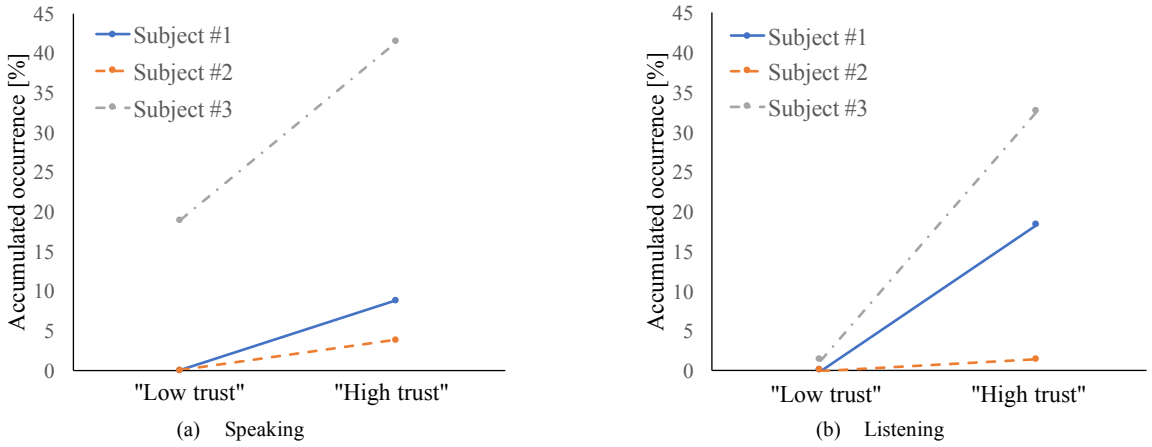



Figure 3. Accumulated occurrences of facial expression "joy" in two scenarios.

Table I shows frequencies of head nod, eye blink, and eye movement of the participants detected in two scenarios. Nodding frequency does not show significant patterns. Although patterns of nodding frequency are expected to change as the communication proceeds from "low trust" to the "high trust", we consider that the communication duration is too short to create emphatic responses of the participants. For further observation, it is necessary to break down the communication of each scenario into stages (initial stage, exploration stage, struggling stage, and closing stage). The frequency of eye blink shows a significant increase from "low trust" to the "high trust" during speaking. Conversely, it shows a significant decrease during listening. This result is consistent with the previous research where during tasks requiring higher cognitive load, subjects' blink rate tends to decrease [20]. The frequency of eye movement shows a significant increase in the vertical direction regardless of the scenarios. This behavior is highly affected by the experimental setting, where the participants' heads were in lower position than the experimenter. Therefore, when the participants tried to maintain eye contact with the experimenter, the vertical gaze movements were remarkable than the horizontal.

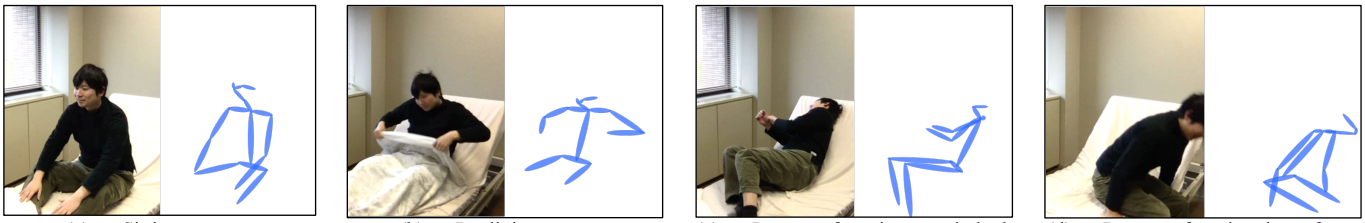
Figure 3 shows the accumulated occurrences of facial expression "joy" in two scenarios. The expression of "joy" increases from the "low trust" to the "high trust" scenarios in both speaking and listening behaviors. This result represents the feeling of being nervous or uncomfortable of the participant during communication in the "low trust" scenario.

## V. CONCLUSION AND FUTURE WORK

We have demonstrated our framework to assess the nonverbal behavior of the patient from a single camera. With its ability to detect human pose information and facial landmarks, the presence of the healthcare worker can be detected to start analyzing the nonverbal behavior of the patient automatically. The proposed framework will also eliminate the need for a number of sensor devices such as gyroscopes and eye trackers, to analyze the nonverbal behavior.

The proposed framework can be extended to monitor the activity of the patients. For example, by using the detected pose information, the patient body movements such as sitting posture, posture of turning over in bed, and posture of getting down from the bed can be monitored and characterized for each patient. This information will lead to an improvement of

 ISBN: 978-1-61208-618-7 



(a) Sitting posture (b) Reclining posture (c) Posture of turning over in bed (d) Posture of getting down from the bed

Figure 4. Using pose information to detect patient's body movements.

the healthcare service. Figure 4 shows four different postures of a user detected by our framework.

We will continue to conduct more experiments with various settings to confirm the robustness of our framework and to increase its functionality in order to improve the communication between the healthcare worker and the patient in particular and the healthcare service in general. By building a private cloud environment, data derived by our framework can be stored as big data to analyze possible efforts to increase patient satisfaction. Authorized persons can access the results by using cloud applications on hand-held devices. As a result, any issues related to the communication between the healthcare worker and the patient can be found at the early stage.

## REFERENCES


[1] P. G. Northouse and L. L. Northouse, Health Communication: Strategies for Health Professionals, second ed., Northouse, 1992.

[2] L. S. Pettegrew, Some boundaries and assumptions in helthcare communication, Explorations in provider and patient interaction, Louisvillem KY:Humana, 1982.

[3] A. J. Davis, Listening and Responding : Cornerstone of Helping, Elsevier Health Sciences, 1984.

[4] P. Ekaman and W. V. Friesen, "The repertoire or nonverbal behavior: categories, rigins, usage, and coding," Semiotica, vol. 1, pp. 49–98, 1969.

[5] S. T. L. Pohlmann, E. F. Harkness, C. J. Taylor, and S. M. Astley, Evaluation of Kinect 3D Sensor for Healthcare Imaging, Journal of Medical and Biological Engineering, vol. 36, pp. 857–870, 2016.

[6] S. Thibaud, M. Daniel, and K. Rana, "Facial Action Unit Detection Using Active Learning and an Efficient Non-linear Kernel Approximation," 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 2015.

[7] J. Chen, J. F. Cremer, K. Zarei, A. M. Segre, and P. M. Polgreen, "Using Computer Vision and Depth Sensing to Measure Healthcare Worker-Patient Contacts and Personal Protective Equipment Adherence Within Hospital Rooms," *Open Forum Infect Dis*, vol. 3, no. 1, pp. 200–206, Feb. 2016.

[8] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person2D pose estimation using part affinity fields," Computer Vision andPattern Recognition, 2017.

[9] H. C. Chen and D. Ramanan, "3D Human Pose Estimation = 2D Pose Estimation + Matching," arXiv:1612.06524v2 [cs.CV], pp. 1–9, 2017.

[10] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), pp. 1867–1874, 2014.

[11] J. O. Cooper, T. E. Heron, and W. L. Heward, Applied Behavior Analysis 2nd ed., Pearson, (2007).

[12] M. Inoue, T. Irino, N. Furuyama, R. Hanada, T. Ichinomiya, and H. Masaki, "Manual and accelerometer analysis of head nodding patterns in goal-oriented dialogues," Interaction Techniques and Environments - 14th International Conference, HCI International 2011, Orlando, FL, pp. 9–14, 2011.

[13] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 11, pp. 1473–1488, 2008.

[14] A. S. Won, J. N. Bailenson, S. C. Stathatos, and W. Dai, "Automatically Detected Nonverbal Behavior Predicts Creativity in Collaborating Dyads," Journal Nonverbal Behavior, vol. 38, no. 3, pp. 389–408, 2014.

[15] N. Burba, M. Bolas, and D. M. Krum, "Unobtrusive measurement of subtle nonverbal behaviors with the Microsoft Kinect," Proceeding VR '12 Proceedings of the 2012 IEEE Virtual Reality, pp. 1–4, 2012.

[16] J. Martinez, R. Hossain, J. Romero, and J.J. Little, "A simple yet effective baseline for 3d human pose estimation," arXiv:1705.03098 [cs.CV], pp. 1–10, 2017.

[17] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1325 1339, 2014.

[18] Y. Jaana, O. D. A. Prima, T. Imabuchi, H. Ito, and K. Hosogoe, "The development of automated behavior analysis Software," Proc. SPIE 9443, Sixth International Conference on Graphic and Image Processing (ICGIP), 2014.

[19] T. Imabuchi, O. D. A. Prima, and H. Ito, "Visible Spectrum Eye Tracking for Safety Driving Assistance," in Trends in Applied Knowledge-Based Systems and Data Science, vol. 9799, no. 4, pp. 428–434, 2016.

[20] L. Vincze and I. Poggi, Communicative Functions of Eye Closing Behaviours, Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues Lecture Notes in Computer Science, vol. 6800, 2011.


 ISBN: 978-1-61208-618-7 