Human Identification through Kinect’s Depth, RGB, and Sound Sensor

P Shunmugam¹, K Kamarudin¹,², A A A Mosed¹,² and S A A Shukor¹

¹School of Mechatronics Engineering, Universiti Malaysia Perlis (UniMAP), Arau, Malaysia,
²Centre of Excellence for Advanced Sensor Technology (CEASTech), Universiti Malaysia Perlis, Arau, Malaysia

Abstract. Human identification is a very important subject in computer field and has been researched widely. This paper proposes human identification system based on Kinect’s individual sensor as well as combination of all its available sensors. In the first part of the project, each sensor on the Kinect (i.e IR depth sensor, RGB camera and microphones) was used for skeleton recognition, face recognition and speech recognition respectively. Then, these individual recognition methods are combined as step-by-step process into a multi-sensor recognition system. Few experiments were carried out to test the reliability of the developed human identification systems. The results show that multi-sensor based human identification system is highly efficient compared to the single-sensor system. This is because, multi-sensor recognition system involves many recognition stages as each recognition stage needs particular biometric information of the user.

1. Introduction

Human identification has good application prospect and potential economic value. Human identification is a systematic process, involves biometrics technology using human biological characteristics. Human identification is beneficial in many fields mainly on authentication and security system [1].

One of the most significant inventions in the year 2010 is the Microsoft Kinect because of the high-resolution depth, four arrays of microphones, and visual (RGB) features that it provides for a relatively much lower cost when compared to other 3D cameras such as stereo cameras and Time-Of-Flight cameras[2]. Kinect is used for applications like playing a virtual violin, to applications in health care and physical therapy, retail, education, and training [3]. Figure 1 shows the various biometric technologies [1][4][5][6] and components of the Kinect sensor. The followings are the specification of the Kinect; an RGB camera with a 640 x 480-pixel, an infrared (IR) emitter and an IR depth sensor to capture depth images an array of four microphones with 16-kHz, 24-bit mono pulse code modulation (PCM) and a 3-axis accelerometer to determine the orientation of the Kinect [2].
The availability of various sensors available on the Kinect can be utilized for multi-typed sensor data based human identification purpose where more data can be obtained and collected, which resulting in more accurate and efficient identification. Although, Kinect consists of many sensors, but the price of Kinect much lower compared to other multi sensor devices [7]. By using Kinect, low cost, highly efficient and accurate human identification system can be developed. In this research, speech recognition system using microphone, skeleton recognition system using the IR depth sensor and face recognition system using the RGB camera was developed for the single-typed sensor data based human identification system. Speech recognition recognizes speech command, skeleton recognition recognizes user’s 19 skeleton mean information while face recognition recognizes user’s face, eye, nose and lips template. For multi-typed sensor data based human identification system, speech, skeleton and face recognitions are combined into one system where the process will run step-by-step. The reliability of the developed human identification methods was tested and analyzed.

2. Related Work

In previous study, the Mel-frequency cepstrum coefficients, the logarithmic power and their related values are calculated from the personal voice [8]. Another study in [9] was on facial depth data of a speaking subject, captured by the Kinect device, as an additional speech informative modality to incorporate to a traditional audiovisual automatic speech recognizer. They presented feature extraction algorithm for both visual and accompanying depth modalities, based on a discrete cosine transform of the mouth region-of-interest data, further transformed by a two-stage linear discriminant analysis projection to incorporate speech dynamics and improve classification. An autonomous voice and motion-controlled video camera system was developed by Jody (Pritchett) Shu [10]. The Kinect data is processed in real time by the software system on the laptop to track the lecturer’s movement and automatically change focus to the board, projector screen, or other visual aids. In addition to posture and voice-command recognition, the system filters out background sounds and acts as a directional microphone to pick up the lecturer’s speech in an optimal fashion for voice recognition [10].
3. Methodologies

3.1. Single-type sensor data based human identification

3.1.1. Speech Recognition System (Microphone)

This system uses automatic speech recognition method which is speech to text technique. The flow chart of the system is shown in Figure 2. The Kinect’s microphone array collected the 24-bit audio data and pre-processed to remove the background noise and echoes of the signal using automatic echo cancellation (AEC) algorithm [11]. When there is a multiple set of microphones, the time that sound arrives from an audio source to each microphone is slightly different. Audio data captured from Kinect microphones is fed to the preliminary processor and at that section Beam forming and Sound localization techniques determine the direction of the sound source and the set of microphones is used as a directional microphone.

![Speech Recognition System Flow Chart](image)

Speech module section consists of two parts which are Microsoft Speech Recognition Engine [12] and Microsoft speech recognition grammar. Microsoft speech recognition engine matches vocal inputs with words and phrases, as defined by grammar rules. Two grammar rules exist in the speech recognition grammar. First is the simple rule which recognize small words and commands. Another rule can recognize and organize semantic content with various user accents. In this system, we had implemented the simple rule where the user needs to add the word into the command list and then the word will be initialized. This system built by grammar builder and speech recognizer from the “System.Speech(4.0.0.0)” of NET Constructor with LabVIEW software [13].

3.1.2. Skeleton Recognition System (IR Depth sensor)

Depth sensor is used to recognize human skeleton in this research using Kinesthesia toolkits. It utilizes the “Distance and Displacement between Joints VI” which obtain 20 joint coordinates of the skeleton [14] starting from head until feet. When the Microsoft Kinect starts operating, the subject to be recognized must stand at a minimum distance of 1.5 meters from the Kinect. 20 joints will be detected, and the coordinates are obtained. All the joints are connected, where 19 pairs of adjacent joints make up the nones for the whole skeleton. The distance between joint pair (i.e bone) calculated using Euclidean distance algorithm (refer Equation 1) [15]:

\[
d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}
\]

where \(d\) is the distance between 2 joints, \((x', y', z')\) are the coordinates of the first joint and \((x, y, z)\) are the coordinates of the second joint. By applying this algorithm, again it is possible to obtain 19 absolute distances between the 20 adjacent joints. The data then is written in one spreadsheet file continuously. After Euclidean distances have been calculated, the mean for each of 19 absolute distances is computed using Equation 2:
\[ \bar{x} = \frac{1}{n} \left( \sum_{i=0}^{n} x_i \right) = \frac{x_1 + x_2 + \ldots + x_n}{n} \] 

(2)

Where \( n \) is the total number of readings. The data saved as the subject’s name and will be registered as user. During the recognition, the 19 mean distance data stored in the file (.csv) will be compared \( \pm 1 \) range with real time data of the 19 Euclidean distance data of subject to be recognized. The 19 outputs are either logic 1 or logic 0. Then all of these are summed using “Compound Arithmetic” function and choosing the “Add” option. If the score is equal to or more than 15 thresholds, then the subject is recognized as the registered user.

3.1.3. Face Recognition System (RGB camera)

One of the most remarkable abilities of human vision is that of face recognition [16]. RGB camera was used to recognize face in this section. This recognition was developed using IMAQ Vision Development Toolkit [17]. The system uses built-in colour and pattern matching capabilities to perform face recognition on a Kinect camera stream. Given right templates and adequate settings it does perform well with respect to its simplicity. The whole approach can be split into two main parts; Face Detection and point following. Face detection was accomplished by matching eye, nose and lips templates against the camera stream image. Figure 3 shows the sample of the eye, nose and lips template in .png file format.

![Figure 3. Sample of pattern matching templates: (a) face (b) eye (c) nose (d) lips.](image)

3.2. Multi-sensor based Human Identification

This system developed for human identification based on multiple types sensor data particularly depth, RGB, sound sensor. It is the combination of all the above speech recognition, skeleton recognition and face recognition approaches which are executed step by step. Firstly, the user must set up the data of his speech command, skeleton information and face template in the system. To be recognized by the system, the subject must speak the speech command. If the speech command is recognized, then the subject will undergo second step which is the skeleton recognition. When the skeleton is recognized, then the subject will go through face recognition as the last step. The user will be identified and authorize the system only if all 3 stages of recognitions are passed. If any of the stage failed to be recognize, he or she will not be recognized, and the system will end the process.

4. Experimental Result

Few experiments were carried out on eight different subjects to test the reliability of the method for the single and multi-typed sensor data based human identification system. The results of the experiments are shown and discussed in this chapter.

4.1. Single-typed sensor data based human identification system

4.1.1. Speech Recognition System (Microphone)

The efficiency of the speech recognition system was tested by using “measure” and “scan” commands. Results are shown in Table 1. Each test consists of two rounds where in each round, subject is allowed to say the command for a of maximum of 10 attempts to be recognized. If more than 10 attempts, the subject will not be recognized, and the system ends the recognition operations.
Table 1. “Measure” command test on speech recognition system

| Subjects | Command “Measure” | Command “Scan” |
|----------|-------------------|----------------|
|          | 1<sup>st</sup> round | 2<sup>nd</sup> round | 1<sup>st</sup> round | 2<sup>nd</sup> round |
| Osamah   | Fail              | Fail            | 1              | 1              |
| Fiqa     | Fail              | 3               | 8              | 7              |
| Izyan    | 6                 | 2               | 1              | 1              |
| Ling     | 4                 | 3               | 4              | 2              |
| Khei     | Fail              | 1               | 3              | 2              |
| Raj      | 1                 | 2               | 2              | 1              |
| Teven    | 2                 | 1               | 1              | 1              |
| Puva     | 3                 | 2               | 1              | 1              |

From Table 1, we can observe that Osmanah failed to pronounce “measure” command as he has Arabic mother tongue and he faced difficulties in pronouncing the command correctly. But Raj, Teven and Puva was able to pronounce the “Measure” command correctly and easily in both rounds. Also, we can observe Osamah, Izyan, Teven and Puva were able to easily pronounced the “Scan” command and was recognised by the system at the very first attempt. But still few subjects like Fiqa unable to be recognised early by the system as she has very soft voice and very low sound. This shows that the subject must say the command bold and confident. The weakness of this system is, any subject with the right command will be recognized as the user by the system. To increase the efficiency of this system, the command must be unpredictable word, difficult to be pronounced and must kept secretly by the user.

4.1.2. Skeleton Recognition System (IR Depth sensor)
Few experiments were conducted on eight subjects to test the reliability of this skeleton recognition method. Firstly, the matching score (threshold) is fixed to 15. Each subject tested two random users’ skeleton information. The results are shown in Table 2.

Table 2. Testing random user data on random subject.

| Subjects | User Match score | Recognized? | User Match score | Recognized? |
|----------|-----------------|-------------|-----------------|-------------|
| Osamah   | Puva 7          | NO          | Khei 10         | NO          |
| Fiqa     | Puva 10         | NO          | Osamah 16       | YES         |
| Izyan    | Puva 8          | NO          | Fiqa 17         | YES         |
| Ling     | Puva 9          | NO          | Izyan 14        | NO          |
| Khei     | Ling 15         | YES         | Fiqa 5          | NO          |
| Raj      | Khei 11         | NO          | Fiqa 15         | YES         |
| Teven    | Raj 11          | NO          | Osamah 12       | NO          |
| Puva     | Ling 8          | NO          | Fiqa 12         | NO          |

From Table 2, we can observe misrecognition happened when subject (Khei) tried to access the user (Ling) skeleton data. Also, we can conclude that few misrecognitions occurred for Fiqa, Izyan and Raj. This happened because subject and user have the similar same height and body build. The weakness of this system is, few subjects were misrecognized by the system where the subject and the user have similar body built.

4.1.3. Face Recognition System (RGB camera)
Efficiency of the system was determined by testing random user’s face template data on random subject. Each subject tested the system twice. The results are shown in Table 3.
From Table 3, we can observe misrecognition occurred while testing subject (Ling) with user (Izyan) face template data. Also, we can observe that Osamah was misrecognized by Khei data and Puva was misrecognized by Osamah data. This happened because both the user and subject have fair skin, and have similar face features such as eyes, nose and lips. From the experiment result we can conclude that 700 is the best threshold as user can be recognized easily. But still misrecognition occurred when tested with random user data on random subjects. This happened because the subject and the user have the similar face features such as same type of nose, eyes and lips.

4.2. Multi-sensor based human identification system

For multi sensor based human identification system, two experiments were carried out to test the reliability of the system. First experiment used eight subjects to test user (Vishnu) system. User chooses “measure” as the voice command password as it’s difficult to pronounce by others and he can pronounce it correctly and easily. User’s skeleton information and face template also set up in the system. The results are shown in Table 4.

| Subject | User | Stage 1: Speech Command: Measure | Stage 2: Skeleton Match Score:15 | Stage 3: Face Threshold 700 | Authorised? |
|---------|------|---------------------------------|---------------------------------|-----------------------------|------------|
| Osamah  | Vishnu | Failed                          | X                               | X                           | X          |
| Fiqa    | Vishnu | Failed                          | X                               | X                           | X          |
| Izyan   | Vishnu | Yes                             | Failed                          | X                           | X          |
| Ling    | Vishnu | Yes                             | Failed                          | X                           | X          |
| Khei    | Vishnu | Failed                          | X                               | X                           | X          |
| Raj     | Vishnu | Failed                          | X                               | X                           | X          |
| Teven   | Vishnu | Yes                             | Failed                          | X                           | X          |
| Puva    | Vishnu | Yes                             | Yes                             | Failed                      | X          |
| Vishnu  | Vishnu | Yes                             | Yes                             | Yes                         | Yes        |

From Table 4, we can observe that only 4 subjects (Izyan, Ling, Teven and Puva) able to pass first stage (speech recognition). Out of four subjects only one subject (Puva) able to pass second stage (skeleton recognition). But this subject (Puva) too failed to pass the last stage (face recognition) of the system. At the end, this multi-typed sensor based human identification system can only be authorized by the user (Vishnu). Second experiment was conducted using cross user method where five persons act as user as well as subjects tested the system 25 times. The results are shown in Table 5. While Figure 4 shows an example of the process where the user (Puva) has been successfully recognized by the multi-sensor data based human identification system.
Table 5. Five user/subject cross test the system 25 times.

| Subjects | Osamah | Fiqa   | Izyan   | Ling   | Puva   |
|----------|--------|--------|---------|--------|--------|
| Osamah   | Recognised | Not    | Not     | Not    | Not    |
| Fiqa     | Not    | Recognised | Not   | Not    | Not    |
| Izyan    | Not    | Not    | Recognised | Not   | Not    |
| Ling     | Not    | Not    | Not    | Recognised | Not    |
| Puva     | Not    | Not    | Not    | Not    | Recognised |

Figure 4. Puva was recognized by the multi-sensor data based human identification system.

From Table 5, we can observe that the system did not misrecognize any subjects. All these experiments proved that multi-sensor human identification system is more efficient compared to the single-sensor human identification system as only the correct user was recognized and authorized by the system.

5. CONCLUSION
Multi-typed sensor human identification algorithm was developed, analyzed and implemented successfully by a detailed evaluation for real-life scenarios such as highly efficient and low-cost user authentication system. From all the experiments carried out, we can conclude that single-typed sensor data based human identification system are less efficient compared to the multi-typed sensor data based human identification system.

One of the limitations are, voice command password must be kept secretly. Secondly, the person to be identified and tracked must not move his/her body for 10s. Thirdly, for the face recognition the person should maintain the appearance, for example if he or she used glasses in the template image, then he or she must wear glasses during the recognition process. Finally, since Microsoft Kinect uses infrared, camera and microphone, this system only can be used indoors where there is no direct sunlight, and under controlled brightness.

Kinect originally can detect 6 people and track 2 at the same time, but with Kinesthesia toolkit in LabVIEW, only one person is recognized and tracked which reduces the practicality of the algorithm developed. As future work, the Kinesthesia toolkit can be improved to recognize all the six detectable people and track two, which will enhance the practicality of the system.

6. ACKNOWLEDGMENT
The authors would like to acknowledge the support from the Fundamental Research Grant Scheme (FRGS) under a grant number of FRGS/1/2018/TK04/UNIMAP/02/15 from the Ministry of Education Malaysia.
REFERENCES

[1] Bhattacharyya D, Ranjan R, Alisherov F and Choi M 2009 Biometric authentication: A review *Int. J. u-and e-Service, Sci. Technol.* 2 13–28

[2] Recognition H 2013 Human Recognition, Identification and Tracking using Microsoft Kinect Interfaced with DaNI Robot Author: Diyar Khalis Bilal 2012–3

[3] Tashev I 2013 Kinect development kit: A toolkit for gesture-and speech-based human-machine interaction [Best of the Web] *IEEE Signal Process. Mag.* 30 129–31

[4] Kumar M S N and Babu R V 2013 Human gait recognition using depth camera 1–6

[5] Anon 2013 Face Recognition And Registration For Home Surveillance System Using Iterative Closest Point And Haar Cascade Lgorithm By Kranthi Kumar Kandi , B . tech Presented to the faculty of The University o f Houston Clear Lake In Partial Fulfillment of the Requir

[6] Engineering C 2013 Person Authentication Using Face And Voice 71–9

[7] Rahman M W, Zohra F T and Gavrilova M L 2018 Rank level fusion for kinect gait and face biometrie identification 2017 *IEEE Symp. Ser. Comput. Intell. SSCI 2017.*1–7

[8] Kita E, Zuo Y, Saito F and Feng X 2017 Personal Identification with Face and Voice Features Extracted through Kinect Sensor *IEEE Int. Conf. Data Min. Work. ICDMW* 545–51

[9] Galatas G, Potamianos G and Makedon F 2012 Audio-Visual Speech Recognition Incorporating Facial Depth Information Captured By The Kinect Institute of Informatics and Telecommunications, NCSR “ Demokritos “, Athens , Greece Heracleia Lab , Dept . of Computer Science and Engin ., University of Tex 2714–7

[10] State C 2013 Autonomous Voice and Motion Controlled Video Camera System for Instructional Technology by

[11] Rathnayake K A S V, Diddeniya S I A P, Wanniarachchi W K I L, Nanayakkara W H K P and Gunasinghe H N 2017 Voice operated home automation system based on Kinect sensor 2016 *IEEE Int. Conf. Inf. Autom. Sustain. Interoper. Sustain. Smart Syst. Next Gener. ICIAfS 2016.*1–5

[12] Rami M, Svitlana M, Lyashenko V and Belova N 2017 Speech Recognition Systems : A Comparative Review *IOSR J. Comput. Eng.* 19 71–9

[13] Ertugrul N 2000 with LabVIEW™ Simulation

[14] Anon 2017 Copyright by Sanobar Kadiwal 2017

[15] Lip C J, Yeon A S A, Kumarudin L M, Kumarudin K, Visvanathan R, Zaidi A F A, Mamduh S M, Zakaria A and Nooriman W M 2018 Human 3D Reconstruction and Identification Using Kinect Sensor 2018 *Int. Conf. Comput. Approach Smart Syst. Des. Appl. ICASSDA 2018.*1–7

[16] IEEE 1998 Face Recognition: Features versus Templates *IEEE Trans. Pattern Anal. Mach. Intell.* 15 1–11

[17] Adiloğlu S 2016 We are IntechOpen , the world ’ s leading publisher of Open Access books Built by scientists , for scientists TOP 1 % *Heavy Met. Remov. with Phytoremediation* i 13