Computers and Electrical Engineering 000 (2016) 1–11

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots☆

Ing-Jr Ding*, Jia-Yi Shi

Department of Electrical Engineering, National Formosa University, Taiwan

ARTICLE INFO

Article history:
Received 9 May 2015
Revised 2 December 2015
Accepted 3 December 2015
Available online xxx

Keywords:
Kinect microphone array
Speaker recognition
Speech recognition
Humanoid robot
Kinect-SSA
Kinect fuzzy–DTW

ABSTRACT

This study developed a Kinect microphone array-based method for the voice-based control of humanoid robot exhibitions through speech and speaker recognition. A support vector machine (SVM), a Gaussian mixture model (GMM), and dynamic time warping (DTW) were used for speaker verification, speaker identification, and speech recognition, respectively; they were effectively combined for realizing advanced voice-based control of humanoid robot exhibitions. Speech recognition capability was enhanced by using the Kinect microphone array and by combining the DTW-based recognition decisions associated with all the microphones through a fuzzy control scheme. A humanoid robot with the proposed voice-based control can be controlled through voice commands by authenticated users. The robot first verifies the authenticity of the personal operator, following which it identifies the operator and validates the command. Subsequently, it executes the command if both the user and command are valid. Experimental results demonstrated the effectiveness and accuracy of the proposed method.

1. Introduction

Currently, the use of voice control–based human–computer interaction for device operations, including the exhibition control of humanoid robots, is widespread. Speech pattern recognition is a critical task in speech processing and is vital in the voice-based control of devices, including the control of robot exhibitions. Speech and speaker recognition are two primary tasks in speech pattern recognition [1–5]. Mature automatic speech recognition techniques are widely used in numerous commercial products, such as intelligent mobile devices and smart human–machine interactive games. With the maturing of speech recognition techniques [3–5], another class of speech pattern recognition—speaker recognition—has gradually emerged; it is more functional than speech recognition [1,2]. Speaker recognition is employed in surveillance and security systems, such as for access control and in community surveillance systems [1]. Although both speech and speaker recognition can be accurately achieved in specific systems, effective integration of speech and speaker recognition for controlling the action of a robot has rarely been reported. The main objective of this study was to develop a Kinect microphone array-based speech pattern recognition scheme involving integration of speech and speaker recognition for the exhibition control of humanoid robots.

The Kinect device developed by Microsoft is a sensor that can be used as a signal receiver for collecting sensed video and audio data [6]. The Kinect sensor is known for its excellent performance in simplifying acquired image data. The Kinect-

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. S. D. Prior.

* Corresponding author. Tel.: +886 56315630.
E-mail address: ingjr@nfu.edu.tw (I.-J. Ding).

http://dx.doi.org/10.1016/j.compeleceng.2015.12.010
0045-7906/

Please cite this article as: I.-J. Ding, J.-Y. Shi, Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2015.12.010

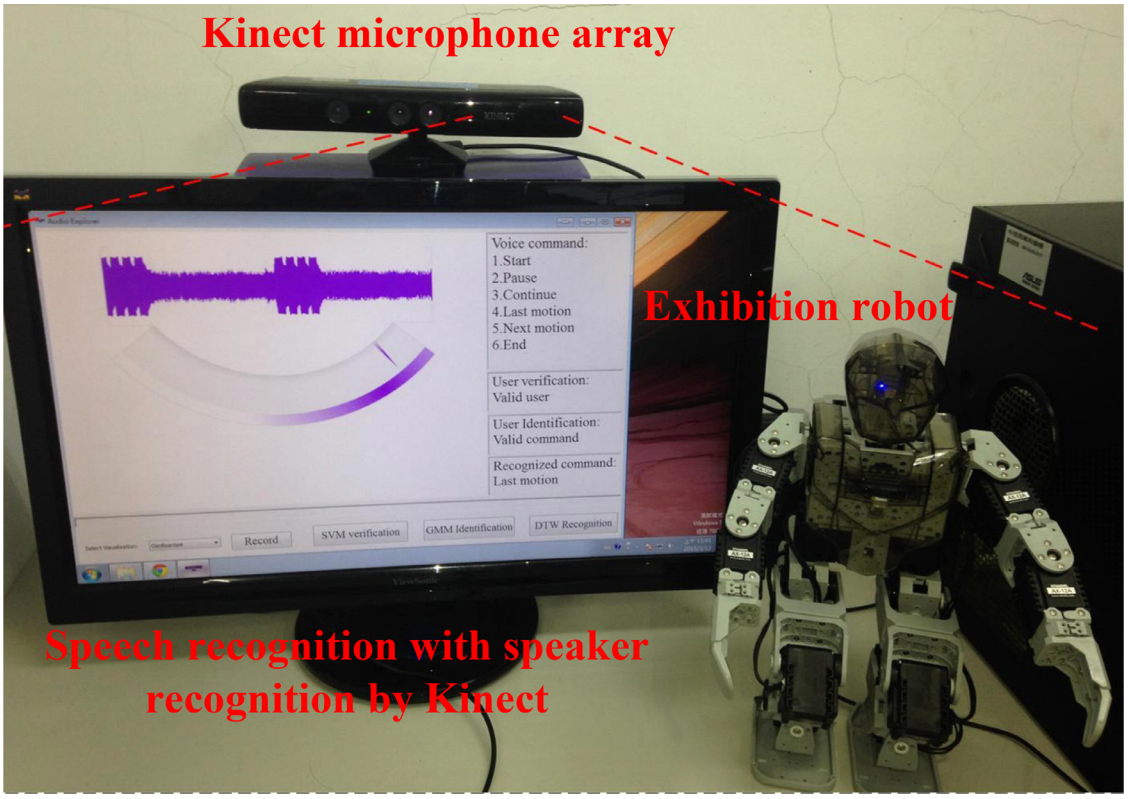

Fig. 1. Developed Kinect microphone array-based speech and speaker recognition for exhibition control of the humanoid robot.

derived human skeleton, with each human joint position represented by a corresponding three-dimensional data point, is used for easy human gesture recognition [7–11]. In human gesture recognition, both RGB and depth-sensing Kinect cameras are employed to obtain three-dimensional data of human joint positions. The vast majority of Kinect applications are in the field of video-based gesture recognition. Studies have employed the Kinect microphone array for audio-based tasks. For example, the Kinect sensor was used for audio-based human behavior recognition in [12].

The current study developed a Kinect microphone array–based voice control scheme for regulating the actions of a robot. Studies have employed Kinect sensors for controlling robots [8, 13–15]. In [13], a Kinect-based control system with a proportional-derivative control algorithm was proposed for a manipulator robot; a combined scheme was used for controlling both the robot arms and Kinect. In [14], Kinect was used to recognize body gestures and provide an interactive interface between the body gesture module and the humanoid robot Nao (manufactured by Aldebaran Robotics). Furthermore, in [15], Kinect, gesture recognition systems, and mobile devices were integrated for enabling interactive discussions with the users. In the author's previous study [8], Kinect-based gesture recognition based on an adaptive hidden Markov model was presented for the use in a humanoid robot that imitates human actions. The main objective of all studies that have used Kinect for robot control has been to achieve gesture recognition-based control; few studies have used speech pattern recognition-based methods for controlling robots. The Kinect microphone array-based voice control scheme for robot control presented herein is a speech pattern recognition-based method wherein speech and speaker recognition are effectively combined for realizing fine robot control; moreover, speaker recognition is divided into speaker identification and speaker verification for accurate robot operator authentication.

Fig. 1 depicts a practical application scenario for the operational exhibition control of a humanoid robot through an operator's voice commands captured by the microphone array of a Kinect sensor. The humanoid robot responds to valid voice commands. The robot performs three tasks: it authenticates the operator, identifies the operator, and executes the voice commands following successful operator authentication and identification; these tasks are performed using three recognition schemes: support vector machine (SVM) [16], Gaussian mixture model (GMM) [17], and Kinect fuzzy–dynamic time warping (DTW) schemes, respectively. For increasing speech recognition accuracy, the conventional DTW method [18] is improved using the Kinect microphone array and a fuzzy logic control (FLC) scheme [5, 19]; the improved method is termed the Kinect fuzzy–DTW method; the method is detailed later in this paper. Note that the author's previous work in [2] is completely different to the current study herein. The main research of [2] is to develop a method by integrating GMM, SVM and DTW for the application of speaker recognition. In this study, a Kinect microphone array-based scheme to properly

Please cite this article as: I.-J. Ding, J.-Y. Shi, Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2015.12.010