Pattern Recognition Letters 36 (2014) 196–203

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Context-based hand gesture recognition for the operating room

Mithun George Jacob, Juan Pablo Wachs *

School of Industrial Engineering, Purdue University, West Lafayette, IN 47906, USA

ARTICLE INFO

Article history:
Available online 6 June 2013

Communicated by Luis Gomez Deniz

Keywords:
Continuous gesture recognition
Operating room
Human computer interaction

ABSTRACT

A sterile, intuitive context-integrated system for navigating MRIs through freehand gestures during a neurobiopsy procedure is presented. Contextual cues are used to determine the intent of the user to improve continuous gesture recognition, and the discovery and exploration of MRIs. One of the challenges in gesture interaction in the operating room is to discriminate between intentional and non-intentional gestures. This problem is also referred as spotting. In this paper, a novel method for training gesture spotting networks is presented. The continuous gesture recognition system was shown to successfully detect gestures 92.26% of the time with a reliability of 89.97%. Experimental results show that significant improvements in task completion time were obtained through the effect of context integration.

1. Introduction

Due to advances in computer-assisted surgery, human–computer interaction (HCI) in the operating room (OR) is gradually becoming commonplace. Several surgical procedures such as tumor resections mandate the use of computers (WHO, 2009) intra-operatively and during pre-operative planning. Since HCI devices are possible sources of contamination due to the difficulty in sterilization, clinical protocols have been devised to delegate control of the terminal to a sterile human assistant (Albu, 2006; Schultz et al., 2003; Oliveira et al., 2012). However, this mode of communication has been shown to be cumbersome (Maintz and Viergever, 1998) and prone to errors (Albu, 2006) and therefore increase the overall duration of the procedure. As a secondary effect, such indirect interaction could increase the surgeon's cognitive load (Firth-Cozens, 2004; Halverson et al., 2011; Lingard et al., 2004) and highlights the need for a sterile method of HCI in the operating room.

Computer systems used to navigate MRI images before and during the surgery (PACs) (Boochever, 2004; Lemke and Berliner, 2011; Mulvaney, 2002) conventionally requires the use of keyboard, mice or touchscreens for MRI browsing. This paper proposes a sterile method for the surgeon to naturally, and efficiently manipulate MRI images through touchless, freehand gestures (Gallo et al., 2011; Ebert et al., 2011; Johnson et al., 2011; Mentis et al., 2012; Micire et al., 2009).

Image manipulation through gestural devices has been shown to be natural and intuitive (Ebert et al., 2011; Hauptmann, 1989) and does not compromise the sterility of the surgeon. An example of a touchless mouse (Gratzel et al., 2004) utilizes stereo vision to localize the hand in 3D which allows the user to control the interface with hand gestures. A multimodal solution (Keskin et al., 2007) for obtaining patient input using gestures has also shown to be effective. Systems based on voice recognition have also been utilized in the OR such as AESOP, a voice controlled robotic arm which handle a camera during surgery (Mettler et al., 1998). The main drawback with voice recognition systems are the long reaction times, erratic responses and user dependency (Nishikawa et al., 2003). The uncontrolled and noisy environment characteristic to the OR has led to the development of gesture-based (Jacob, 2011; Jacob et al., 2012) interfaces for the operating room.

The need for sterile image manipulation has led to the development of touchless HCI based on the use of facial expressions (Nishikawa et al., 2003), hand and body gestures (Gratzel et al., 2004; Keskin et al., 2007; Zeng et al., 1997; Grange et al., 2004) and gaze (Nishikawa et al., 2003; Yanagihara and Hama, 2000). It should be noted that none of this research have incorporated surgical contextual cues to disambiguate recognition of false gestures and improve gesture recognition performance.

An alternate modality is proposed to replace HCI devices such as the keyboard, mouse, and touch-screens traditionally used to navigate and manipulate a sequential set of MRI images. The proposed system extends the work previously developed by the authors (Wachs et al., 2008; Jacob et al., 2012) with the use of dynamic two-handed gestures and contextual knowledge. Additionally, the authors provide a novel, analytical method to optimize

* Corresponding author. Tel.: +1 765 496 7380.
E-mail addresses: mithunjacob@purdue.edu (M.G. Jacob), jpwachs@purdue.edu (J.P. Wachs).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.05.024

the gesture recognition system using *a priori* data and have introduced a new contextual cue for improved navigational performance.

2. System overview

2.1. MRI image browser

An MRI browser was developed with OpenGL (Shreiner et al., 2005) and OpenCV (Bradski, 2000) for the navigation and manipulation of MRI images inspired by the OsiriX system (Rosset et al., 2004). Several MRI sequences and slices within sequences are displayed in the browser for selection, navigation and manipulation. The browser supports a set of 10 commands typically used during surgery to manipulate and revise MRI images. The commands and corresponding gestures were obtained by consulting nine veterinary surgeons that are used to working with MRI browsing software. Additionally, the image browser can also be operated through the keyboard/mouse.

The lexicon consists of ten gestures (see Fig. 1(a)) which encompasses image manipulation tasks such as zooming (zoom in, and zoom out), rotation (clockwise, and counter-clockwise) and brightness change (brightness up, and brightness down) as well as image navigation tasks such as browsing (up, down, left, and right).

2.2. Intent and gesture classification

Anthropometric information of the user was obtained through a Microsoft Kinect (Kinect – Xbox.com, 2012) using the OpenNI SDK (OpenNI, 2012). This includes the 3D coordinates of the left and right shoulders, head, and both hands (see Fig. 1(b)). The positions of the shoulders and head were used to classify the pose of the user (the user is assumed to make intentional gestures when facing the system). The position of the both hands is used to obtain the trajectory of the hands over time (only hand positions during an "intentional" pose are recorded). These cues are henceforth referred to as *visual* contextual cues.

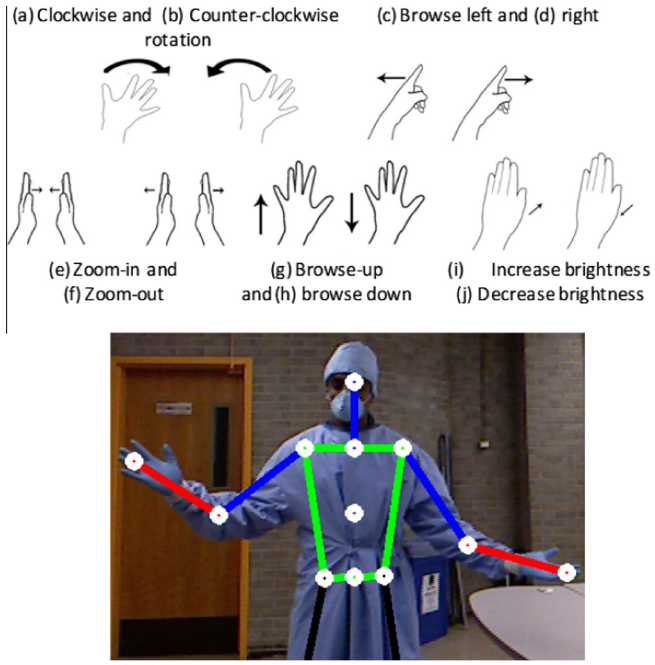

Fig. 1. (a) Gesture lexicon (b) skeleton model and tracked marker-less points.

Non-visual contextual cues include the history of commands performed during the user interaction as well as the time between successive commands. Another important non-visual cue is inferred from the biopsy procedure. The tip of the biopsy needle in the brain is tracked as it penetrates the patient's brain tissue (see Fig. 2). Since the MRI images of interest are usually related to the position of the needle tip inside the brain, this location is mapped to a sequence and a slice from the available set of MRI images. The MRI image browser uses this information to display the mapped slice saving the user from executing several navigational commands to reach the slice.

In our previous work (Jacob et al., 2012), we established that the inclusion of context significantly reduced the false positive rate of gesture recognition. The following sections briefly describe the integration of contextual cues within the recognition framework.

3. Intent classification through visual context

It has been shown that gaze is a critical contextual cue used to determine the intent of a user to interact with an entity (Emery, 2000). Other important anthropometric cues (Langton, 2000) such as torso orientation have also been used to aid in intent recognition.

3.1. Torso orientation (T_θ)

The orientation of the torso is computed using the skeletal joint coordinates in 3D. The coordinates of the left and right shoulder was used to compute the azimuth orientation T_θ w.r.t the X-axis. A pose is counted as intentional if the user faces the Kinect sensor (i.e. T_θ is approximately 180°), this resembles the face to face interpersonal interaction.

3.2. Head orientation (H)

The skeletal coordinates of the user's body is used to reduce the search space for the head. The Viola–Jones frontal face detector (Viola and Jones, 2004) is used to compute the location of the head in the reduced search space. Since the aforementioned detector is a frontal face classifier, the mean H of the binary response of the detector (detected/not detected) is computed over a window of $W = 10$ frames. A high value of H corresponds to a high degree of confidence that the head is oriented towards the sensor.

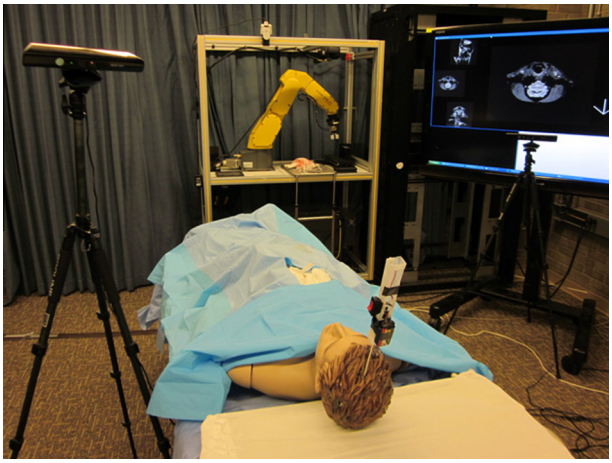

Fig. 2. Experimental setup with the mock biopsy needle inside the model's head.