Expert Systems With Applications 56 (2016) 100–115

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Gesture phase segmentation using support vector machines

Renata Cristina Barros Madeo, Sarajane Marques Peres*, Clodoaldo Aparecido de Moraes Lima

School of Arts, Sciences and Humanities, University of São Paulo, R. Arlindo Béttio, 1000, São Paulo, SP 03828-000, Brazil

ARTICLE INFO

Article history:
Received 26 October 2014
Revised 11 February 2016
Accepted 14 February 2016
Available online 26 February 2016

Keywords:
Gesture analysis
Gesture phase segmentation
Machine learning
Support vector machine

ABSTRACT

An interaction between humans or between a human and a machine will be more effective if it is supported by gestures. In different levels of complexity, the communication system used in human interaction includes the use of gesture. In natural conversation, for instance, speakers use gestures for both to enhance the expressiveness of their speech and to support their own linguistic reasoning. The audience absorbs the content being transmitted also based on the speakers' gesticulation. Thus, an analysis of gestures should add value to the purpose of the interaction. One of the concerns in the analysis of gestures is the problem arising from the segmentation of phases of a gesture (rest position, preparation, stroke, hold and retraction), which, from the standpoint of Gesture Theory, may reveal information on prosody and semantics of what is being said in a discourse. Finding an automation solution to this problem involves enabling the development of theoretical and application areas that are based on the analysis of human behavior and on the interpretation and generation of natural language. In this study, the problem of gesture phase segmentation is modeled as a problem of classification, and then support vector machine is employed to design a model able to learn the patterns of gesture that are inherent to each phase. This work presents two main highlights. The first is to address the limitations of the segmentation approach through the study of its performance in different scenarios that represent the complexity of analyzing patterns of human behavior. In this study, we reached an *F*-score around 0.9 for rest position and around 0.8 for stroke and preparation as segmentation results in the best cases. Moreover, it was possible to investigate how classification models are influenced by human behavior. The second highlight refers to the conduction of an analysis by considering the standpoint of specialists concerned with gesture phase segmentation in the area of Linguistics and Psycholinguistics, through which we obtained impressive results. Thus, in regard to the suitability of our approach, it is a feasible means of supporting the development of the Gesture Theory as well as the Computational Linguistics and Human Machine Interaction fields.

1. Introduction

In the last few years, there has been an increase of studies on gesture analysis in Computer Science (Chen, Jafari, & Kehtarbavaz, 2015; Han, Shao, Xu, & Shotton, 2013; Madeo, Wagner, & Peres, 2013b; Mitra & Acharya, 2007). Currently, both the academic world and industry are interested in gesture analysis, mainly with the aim of developing new methods of Human–Computer (or Machine) Interaction. These methods have been developed considering different aspects, such as: different ways of interacting (Turk, 2014); execution of gestures with body parts that are little explored in gesture analysis (Tran, Doshi, & Trivedi, 2012); multiple users interacting at the same time with one or more devices (Baraldi, Bimbo, & Landucci, 2008); and interaction with avatars (Kopp, Sowa, & Wachsmuth, 2004) or robots (Salem, S. Kopp, Wachsmuth, Rohlfing, & Joublin, 2012).

The increasing interest in gesture analysis is partly due to the development of low-cost sensors, for which there are Software Development Kits that provide the means to create applications with different levels of complexity (Guna, Jakus, Poganik, Tomai, & Sodnik, 2014; Lun & Zhao, 2015; Tashev, 2013). These sensors are able to acquire information that can describe human movements at different levels of detail. For example, they have been used to build large-scale benchmark datasets which allow more extensive research to be carried out in the fields of gesture segmentation and gesture recognition (Escalera et al., 2013). Although these sensors have become increasingly popular, video cameras (RGB and RGB-D) and webcams are still the main sources of data for human movements. The methods regarding the use of different types

* Corresponding author. Tel.: +551130918897.
E-mail addresses: renata.si@usp.br (R.C.B. Madeo), sarajane@usp.br, smperesbr@hotmail.com (S.M. Peres), c.lima@usp.br (C.A.d.M. Lima).

http://dx.doi.org/10.1016/j.eswa.2016.02.021
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

of cameras to capture and analyze human motion are discussed by Moeslund and Granum (2001). One example of a recent research about the use of this type of data is presented by Chaquet, Carmona, and Fernández-Caballero (2013) and a recent survey on this theme is presented by Pisharady and Saerbeck (2015). Furthermore, there are several studies that combine video cameras and (low and high cost) sensors in order to improve results in distinct types of problems concerning recognition of human actions. Chen et al. (2015) present a review on studies that use a combination of sensors. Also, the evolution of resistive flex sensors, which are more sensitive and accurate, has enabled the development of applications that need more detailed date. A review on this type of sensor and its applicability in gesture analysis is presented by Saggio, Riillo, Sbernini, and Quitadamo (2016).

Gesture analysis is important for different application fields, since it is concerned with the interactions between humans, and between humans and the environment or machines. These applications usually require the development of methods for: detecting movements made by users; identifying the segments of the movement that concern the application; and recognizing the movement within a range of expected movements. In order to build these applications, different strategies can be used for solving each aforementioned task (Huang, Hu, & Chang, 2011; Kyriazis et al., 2015; Ozturk, Aksac, Ozyer, & Akhajj, 2015; Suarez & Murphy, 2012).

Currently, one of most common applications of human–machine interaction using gestures is the interaction through gestures performed by means of touch screen devices and, on a smaller scale, in three-dimensional space (Baraldi et al., 2008; Kim, Song, & Kim, 2007; Spano, Cisternino, & Paterno, 2012). This type of application may be classified as recognition of a “finite vocabulary of gestures with predefined meaning”. Initiatives that aim at recognizing this type of gestures are specially studied to contribute with the evolution of software able to interpret sign language (Ong & Ranganath, 2005) considering hand configuration (Dong, Leu, & Yin, 2015; Lamberti & Camastra, 2012; Tsai & Lee, 2011), hand configuration and movement (Lee, h. Yeh, & Hslao, 2016; Madeo, Peres, Bíscaro, Dias, & Boscarioli, 2010), grammatical facial expressions (Freitas, Peres, Lima, & Barbosa, 2014), or a combination of these and other elements of a sign language (Koller, Foster, & Ney, 2015); or still able to support the analysis of human behavior in specific tasks, such as playing an instrument (Seger, Wanderley, & Koerich, 2014) or learning how to conduct an orchestra (Brown & Sasson, 2012).

Following a different line of research, there are researchers that analyze natural gestures, which according to Kim et al. (2007) are “meaningless and uncertain” with “cultural and local diversity”. These efforts are usually related to the analysis of human behavior through the study of movements of different parts of the human body. The studies conducted by Jacob and Wachs (2014), Drosou, Ioannidis, Moustakas, and Tzovaras (2012), Bremner, Pipe, Frase, Subramanian, and Melhuish (2009), and Salem et al. (2012) are examples of how sophisticated problems can be tackled in this field. Jacob and Wachs (2014) are concerned with decision making on whether a gesture is or is not intentional; Drosou et al. (2012) discuss the area of biometric authentication; and Bremner et al. (2009) and Salem et al. (2012) are concerned in developing the ability of natural gesticulation for a robot and analyzing how important this gesticulation is during an interaction with humans. It is a different paradigm of gesture analysis in which the recognition task must consider gesture patterns which are frequently disconnected from their shape or trajectory, but more related to their periodicity, velocity, acceleration, strength and tension.

Also concerning natural gestures, there has been a growing number of studies which support research on Gesture Studies—which can be defined as an interdisciplinary area that seeks to study the use of the hands and other parts of the body for communicative purposes—by combining knowledge from linguistics, psychology, neuroscience, the social sciences and other disciplines (Kendon, 1996). In fact, Gesture Studies are mostly based on the study of different types of gesture and its phases.

Basic types of gesture were established by McNeill (2005) and include iconic, metaphoric, beat, cohesive, and deictic gestures; and by Gibbon (2009), including affective gestures and emblems. The automated identification of these types of gesture was studied by Kettebekov (2004), and Kettebekov, Yeasin, and Sharma (2005) automating the identification of beat and iconic gestures in the analysis of monologues and automating the identification of deictic gestures in the analysis of narratives. The authors’ motivation is the confirmation that coverbal gesticulation presents features that help the identification of prosodic structures which are useful for the automation of natural communication interpretation and natural gesture recognition. More recently, the analysis of different types of gesture has been acquiring more complex motivations. For instance, Hsieh, Hidayati, Chen, Hu, and Hua (2014) present an approach for video data mining aiming at finding the iconic movements that better correspond to specific semantic concepts, considering that such movements vary under different behavioral contexts; Sathyanarayana et al. (2014) present a study referring to automated identification of deictic gestures used by teachers and students during teaching and learning process, and the analysis of the importance of such gestures in the learning context.

Gesture phases are a hierarchy of movements that composes or describes gesticulation. The hierarchy in its basic formulation, proposed by Kendon (1980) and briefly described by McNeill (2005) and Kita, van Gijn, and van der Hulst (1998), divides gestures into segments, or phases, called *preparation, stroke, hold*, and *retraction*, which are the main object of study in this paper. Several authors have been studying the automation of gesture phase analysis considering different strategies, motivations and application contexts. Studies concerned with gesture phase segmentation (Bryll, Quek, & Esposito, 2001; Gebre, Wittenburg, & Lenkiewicz, 2012; Madeo, Lima, & Peres, 2013a; Martell & Kroll, 2007; Okada, Bono, Takanashi, Sumi, & Nitta, 2013; Ramakrishnan, 2011; Wagner, Madeo, Peres, & Lima, 2013; Wagner, Peres, Lima, Freitas, & Madeo, 2014; Wilson, Bobick, & Cassell, 1996; Yin & Davis, 2014) are discussed further in this paper (Section 6), including a comparison between these studies and our approach. There are also efforts that apply concepts of gesture phases in gesticulation synthesis, such as studies that provide natural gesticulation capabilities for robots (Bremner et al., 2009; Salem, Rohlfing, Kopp, & Jaoublin, 2011; Salem et al., 2012).

In Gesture Studies, it is a common practice to record videos of people communicating, through oral or sign language, and then analyze the gestures that they make in these recordings. In conducting this analysis, it is necessary to transcribe some features of the video, such as speech, gesture and its descriptive features, posture and facial expression, performing a task also known as video annotation. One stage of gesture transcription is the gesture phase segmentation, which is usually performed manually by specialists and is a laborious task (Quek et al., 2002). There are some computational tools that support this task (Brugman & Russel, 2004; Kipp, 2012; Maricchiolo, Gnisci, & Bonaiuto, 2012), but only as tools to facilitate information edition, storage and visualization, with no automated support for decision making. Thus, the development of an automated process to support the gesture transcription would be useful (Kita et al., 1998).

The aim of this paper is to outline an approach to support research in the field of Gesture Studies tackling the problem of gesture phases analysis in the context of natural gesticulation. The study presented in this paper illustrates the complexity of the automated analysis of human behavior, since gesticulation is influenced by several variables such as stress, tiredness, mood,