Image and Vision Computing 52 (2016) 56–72

Contents lists available at ScienceDirect

Image and Vision Computing

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

Review article

Recent trends in gesture recognition: how depth data has improved classical approaches☆

T. D'Orazio*, R. Marani, V. Renò, G. Cicirelli

ISSIA - C.N.R.Italy

ARTICLE INFO

Article history:
Received 25 May 2015
Received in revised form 5 May 2016
Accepted 6 May 2016
Available online 26 May 2016

Keywords:
Gesture recognition
RGB-D data
Features extraction
Classification approaches
On-line experiments

ABSTRACT

This paper analyzes with a new perspective the recent state of-the-art on gesture recognition approaches that exploit both RGB and depth data (RGB-D images). The most relevant papers have been analyzed to point out which features and classifiers best work with depth data, if these fundamentals are specifically designed to process RGB-D images and, above all, how depth information can improve gesture recognition beyond the limit of standard approaches based on solely color images. Papers have been deeply reviewed finding the relation between gesture complexity and features/methodologies suitability. Different types of gestures are discussed, focusing attention on the kind of datasets (public or private) used to compare results, in order to understand weather they provide a good representation of actual challenging problems, such as: gesture segmentation, idle gesture recognition, and length gesture invariance. Finally the paper discusses on the current open problems and highlights the future directions of research in the field of processing of RGB-D data for gesture recognition.

Contents

☆ This paper has been recommended for acceptance by Stanley Sclaroff.

* Corresponding author at: Institute of Intelligent Systems for Automation — National Research Council of Italy (CNR-ISSIA), via Amendola 122 D/O, 70126 Bari (BA), Italy.
E-mail address: dorazio@ba.issia.cnr.it (T. D'Orazio).

http://dx.doi.org/10.1016/j.imavis.2016.05.007
0262-8856/

1. Introduction

A gesture is defined as a form of non-verbal communication in which visible bodily actions communicate particular messages, either in place of, or in conjunction with speech. A gesture can include movements of hands, face, or other parts of the body. Gestures are the oldest means of human communication. Nowadays gestures are still important as people use them also in an unconscious way in everyday life, but they can be essential in many situations which involve communications in hazardous contexts. From the scientific point of view, gestures are used and then analyzed in several domains such as sign language recognition, vision-based augmented reality, smart surveillance, virtual environments, and human–computer interaction.

Different definitions of the term *gesture* have been provided in literature and sometimes this term has been interchangeably used as a synonym of the term *action*. In this paper, the definition provided in [1] has been used: a gesture is a physical movement or posture of hands, arms, face or body, made with the intent of conveying meaningful information. We point out the distinction between gestures, which are intentional movements of the body, and actions which are unconscious elementary movements of the body and can be used to understand human daily activities such as running, walking, skating, jumping, or, in a home environment, go to bed, get up, eat a meal, drink water, sit down, stand up, take off the jacket and put on the jacket and many others. According with this definition of gesture, in this review we consider and classify the papers that propose algorithms for gesture/action recognition where the gesture or action terms meet our definition of intentional movement or body posture for communication.

Gestures can be static, when the user assumes a certain pose or configuration, or dynamic with a pre-stroke, stroke and post-stroke phase, as pointed out in [2]. Some gestures also have both static and dynamic elements, as in sign language applications. The automatic recognition requires in the first case the characterization of the spatial disposition of the body parts performing the gesture, whereas in the second case it requires the observation of the sequence of movements generated by the human body.

Many good reviews on action recognition approaches summarized the researches carried out for the recognition of human movements such as walking, jumping, running, and so on [3,4]. Gesture recognition surveys have also been published [1,2,5], giving particular emphasis on hand gestures and facial expressions by the analysis of images acquired by conventional RGB cameras. Although intensity images contain rich information, they are very sensitive to lighting conditions, different point of views, camera resolutions, and cluttered backgrounds. As a consequence, tasks such as people segmentation, motion detection, or interest point detection can be affected by these factors and perform well only in very specific and limited situations. The recent introduction of low cost depth sensors, such as the widespread Microsoft Kinect sensor [6], allowed the development of new gesture recognition approaches. Depth images provide a 3D model of the scene which can be easily used to simplify many tasks such as people segmentation and tracking, body part recognition, motion estimation and so on. Recent reviews on human activity recognition and motion analysis from 3D data have been published in [7,8,9]. Human activities are characterized by sequences of atomic actions, by person–object interactions and by person–person interaction or group activities. A 3D gesture recognition survey, published in [10] provides recent trends on the general issues of sensing, recognition, and experimentation.

In this paper, we will review the literature which uses depth information for gesture recognition approaches from a different perspective. We will focus our attention on the main problems related to the application of gesture recognition approaches in real contexts: the identification of the beginning/ending parts of a gesture; the invariance to gesture length; the normalization with respect to different speeds during gesture executions. We will give particular attention to the most recent literature on gesture recognition which sees a number of publications on approaches based on depth data extracted by RGB-D sensors. The aim of this review is to highlight the main advantages of using depth data as additional information to traditional RGB data and to point out both technological and methodological limits which prevent real application of these approaches to commercial interfaces.

The rest of this review is organized as follows. Section 2 reviews the types of gestures and public datasets that have been used in literature. Then, the challenging problems related to the development of an automatic gesture recognition system will be considered. In particular, Section 3 describes the RGB-D features that better and distinctively characterize a specific movement or posture, setting them apart from similar items. In Section 4, Gesture Recognition is seen as a classification problem in which examples of gestures are used into supervised learning schemes (such as SVM or NN) to model the gestures and to address the recognition problem as a class association problem. In Section 5, the temporal Segmentation of dynamic gestures is approached as the task of determining, in a video sequence, the starting and ending frames of each gesture execution. Finally, after a discussion about the general topic of gesture recognition, Section 6 provides insights into open problems and future research directions. Section 7 reports final conclusions and remarks.

2. Types of gestures and datasets

According to the definition of gestures as intentional movements or posture with the intent to communicate a semantic message, different parts of the body can be involved in the communication. Intentional gestures can be executed by movements of hands, arms, head, torso, and full body. Static gestures are characterized only by postures or shapes of the involved body parts. For example, hand gestures can be characterized by the positions or orientations of the fingers (see Fig. 1). Similarly body gestures can be characterized by the relative positions of hands and legs with respect to the torso. Dynamic gestures are instead characterized by a movement which includes a starting and ending pose of the involved body parts (see the first and last frames of the gestures in Fig. 2).

The recent literature on gesture recognition proves the large interest in the use of public datasets as this allows the scientific community to compare different approaches. Table 1 summarizes the most used datasets. Some action recognition datasets are also cited in the Table as they contain, among the others, some actions which can be considered gestures such as hand waving, hand clapping, boxing, and so on. For this reason, in this review, some action recognition approaches will be also discussed as they can be applied to gesture recognition systems.