Computer Vision and Image Understanding 131 (2015) 72–87

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Gesture recognition corpora and tools: A scripted ground truthing method

Simon Ruffieux [a,b,*], Denis Lalanne [b], Elena Mugellini [a], Omar Abou Khaled [a]

[a] University of Applied Sciences and Arts of Western Switzerland, 1700 Fribourg, Switzerland
[b] University of Fribourg, 1700 Fribourg, Switzerland

ARTICLE INFO

Article history:
Received 4 December 2013
Accepted 8 July 2014
Available online 20 August 2014

Keywords:
Human–computer interaction
Framework
Gesture recognition
Dataset
Corpora
Ground truth

ABSTRACT

This article presents a framework supporting rapid prototyping of multimodal applications, the creation and management of datasets and the quantitative evaluation of classification algorithms for the specific context of gesture recognition. A review of the available corpora for gesture recognition highlights their main features and characteristics. The central part of the article describes a novel method that facilitates the cumbersome task of corpora creation. The developed method supports automatic ground truthing of the data during the acquisition of subjects by enabling automatic labeling and temporal segmentation of gestures through scripted scenarios. The temporal errors generated by the proposed method are quantified and their impact on the performances of recognition algorithm are evaluated and discussed. The proposed solution offers an efficient approach to reduce the time required to ground truth corpora for natural gestures in the context of close human–computer interaction.

1. Introduction

These last years, the field of human gesture and activity recognition has been evolving rapidly due to the research and development in novel sensors for human action, activity and gesture recognition. These new sensors can be split in three types: vision (color, depth or heat), position (inertial motion units, global positioning system, or motion capture) and physiological (temperature, heart rate or electromyography). The advances in technology allowed engineers to produce smaller, more efficient and cheaper sensors and the possibility to embed them in wearable devices such as necklaces, watches, and controllers. These new sensors offer interesting exploration paths for research but also complexify the quantitative comparisons of methods, algorithms and sensors.

We identified three linked issues hindering research in the domain of natural gesture recognition. The recognition of gesture performed in the air by a human is often only considered as a subdomain of action and activity recognition and may confuse researchers, the lack of standards and common structure amongst corpora restraint valid quantitative comparisons of methods and the increasing complexity and cost of creating multi-purposes corpora may become a problem for researchers.

The first issue concerns the confusion between research domains. Three main paths of exploration can be distinguished: human action and activity recognition, human surveillance and human gesture recognition. These three areas of research share many common aspects and are often confused. Action and activity recognition focuses on recognizing high-level actions or activities performed by humans such as walking, hiking, cycling, eating, lying in a couch, and working or preparing a meal. The result of the recognition is mainly applied to monitor or generate statistics about users, contextualize interaction or automatize the environment [1]. Human surveillance focuses on recognizing activity, actions or situations that may indicate an undesired behavior such as shoplifting, dangerous situations, potential threats to persons or simply to facilitate everyday life such as ambient-assisted living [2]. Gesture recognition slightly differs from the two latter, which may be seen as a single field of research with different purposes. Gesture recognition focuses on recognizing gestures performed by a human in order to control or interact with devices; the notion of intention is important. The recognized gestures can be waving a hand, pointing at a device, opening a hand, touching both hands, clapping, sign languages or any gesture performed in the environment. Those air-gestures are often considered as a more natural way to interact with our surrounding environment than physical controllers or buttons [3]. The confusion amongst these three fields complexifies research in gesture recognition; these areas share many common terms and aspects, sometimes with different

[*] Corresponding author at: University of Applied Sciences and Arts of Western Switzerland, 1700 Fribourg, Switzerland.

E-mail address: simon.ruffieux@hes-so.ch (S. Ruffieux).

http://dx.doi.org/10.1016/j.cviu.2014.07.004
1077-3142/© 2014 Elsevier Inc. All rights reserved.

meaning and are easily mixed in the literature. Furthermore, gesture recognition also has specificities that are not considered; specifically datasets and evaluation metrics are often shared amongst research domains despite their many differences. We believe that dedicated guidelines should be developed specifically for the domain of gesture recognition.

The second issue concerns the lack of standards and common structure in gesture recognition datasets. This situation probably originates from the previously mentioned weak differentiation from action and activity recognition, which encourages the reuse of generic tools, guidelines and frameworks and from the fact that datasets are often initially produced only for usages internal to a laboratory. The availability of a dedicated framework supporting the complete chain of operations for developing applications and corpora for gesture recognition should hopefully lead the researchers to share common standards and structures. Some existing frameworks have already been developed to handle multimodal inputs in the generic context of activity, action and gesture recognition. However, these frameworks mainly focus on rapid prototyping functionalities to facilitate the creation of small applications, proof-of-concepts or demonstrators with only basic knowledge of programming. They have notably been used for artistic purposes [4–6]. The need of corpora and related tools for developing and evaluating algorithms is crucial for fields relying on machine learning algorithms. However, none of the reviewed framework has been developed to support facilitated corpora creation in the context of gesture recognition.

The third issue concerns the availability of corpora. In order to train, optimize and evaluate supervised algorithms, researchers have two options: use existing publicly available corpora or create their own. Corpora consist in raw or processed sensor data and their corresponding ground truth. The ground truth, also called labeling or annotation, refers to precise description (textual or equivalent) describing what is in the data or what should be recognized from the data at a specific instant. Creating a true all purposes ground truth would then imply a complete description of explicit and implicit information contained in every frame of the data, which is practically not feasible. Generally, only three types of information are considered for the ground truths in gesture recognition corpora: the name of the gesture, its temporal segmentation and the spatial segmentation of specific body-parts. Therefore the use of existing dataset is not always possible due to missing ground truth or potential specificities of datasets and algorithms. The creation of a corpus is a time-consuming and costly task. The sole acquisition of sensor data for a corpus is already a time-consuming task; it requires recording multiple subjects in different controlled conditions potentially with multiple synchronized sensors. Once the acquisitions completed, the ground truthing of the data is usually performed by an expert human annotator spending numerous hours analyzing the data, frame by frame, and labeling names of gestures, temporal start and end of gestures, spatial position of body-parts, etc. Several programs are already available to facilitate this ground truthing process but only few valid automatic or semi-automatic solutions are applicable to gesture datasets. These limitations often hinder the number and the quality of the datasets produced. This situation notably limits the quantity of elements labeled during the ground truthing process, as each additional label implies additional manual work.

In this article, we present a framework supporting the rapid prototyping of applications, the creation and management of datasets and the quantitative evaluation of algorithms for gesture recognition. The central part of the paper proposes a novel method that has been developed to automatize the acquisition of ground truth when creating new corpora in the context of gesture recognition for human–computer interaction. Concretely, our contributions are:

- A review of the available frameworks, tools and corpora for gesture recognition.
- A framework supporting rapid prototyping, the creation and management of datasets and the evaluation of algorithms in the context of multimodal gesture recognition.
- A method facilitating the ground truthing of data when creating new datasets.

The article proceeds with a discussion of related work in Section 2. Then our "Framework for the Evaluation and Optimization of Gesture Acquisition and Recognition Methods" (FEOGARM) is presented, illustrated with two practical examples of applications and discussed in Section 3. In the central part of the article, Section 4, our novel ground truthing method is presented, followed by an evaluation of its accuracy and potential impact on the recognition rate for machine learning algorithms and discussed. In Section 5, we provide a general conclusion for the article, a conclusion for each section and we point toward potential future works.

2. Frameworks, corpora and ground truths in gesture recognition

This section focuses on a literature review for the three topics addressed in this article. We first clarify what is referred to when using the term gesture recognition and we present a brief review of the recent advances in the field. In the first subsection, we review the frameworks used in literature for the rapid prototyping of gesture recognition applications and we detail three of the most popular ones. Then we review the available corpora for gesture recognition and highlight their main characteristics in the second subsection. Finally, we describe the tools and methods used for ground truthing and illustrate them with practical examples from the literature.

The term gesture recognition is often incorrectly referred as being similar to human action and activity recognition in literature. In this article, gesture recognition precisely refers to a subset of the human activity and action recognition field [7] and can be defined as the process by which specific gestures, intentionally performed by a user, are recognized and interpreted by a machine. Natural gestures refer to expressive and meaningful body movements involving physical motion of the fingers, hands, arms, head, face or body with the intent of conveying information [8]. It can be summarized as an ergonomic body-command performed with the intent of interacting with an automatic system. Gesture recognition has initially reached most of its popularity based on devices measuring acceleration or inertia such as the Wii controller [9] and then some success with video processing techniques based on cameras [10], and time-of-flight cameras [11].

The apparition of commercially available cheap and efficient RGB-D cameras has given a new impulse on vision-based techniques for gesture recognition [12,13]. The use of multimodal inputs has been studied for a long time in the context of human–computer interaction, such as the famous "Put-That-There" example from 1980 [14]; it consists in a human–computer interaction system based on multiple communication channels such as speech, gesture, and writing [15]. Multimodal or multi-sensors systems for human computer interaction have largely evolved and have driven to new research paths thanks to the recent advances in sensors size, price and availability. Multimodal research also drives the research toward contextual and opportunistic systems for the selection and use of available sets of sensors to interact with the environment [16] or toward methods to improve recognition by fusing multiple types of sensors or modalities [15]. This article focuses on the domain of multimodal gesture recognition