Computer Vision and Image Understanding xxx (2014) xxx–xxx

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 xxxx

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/

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

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

performed in mid-air in the context of close-human computer interaction.

2.1. Frameworks for multimodal applications

Several frameworks have been developed to handle and facilitate the development and evaluation of multimodal applications. In [17], most frameworks applicable for pervasive research have been reviewed and classified into design and evaluation frameworks. These frameworks, also called rapid prototyping toolkits, facilitate the development of the entire design process: from early concepts based on low fidelity prototypes to the deployment of high-fidelity prototypes for practical tests and evaluations [17]. An issue with many of these earlier frameworks is that they have been developed mainly for end-users and have limited expressive power [18]. Most frameworks share common properties and technical features such as distributed processing, synchronization, modularity and rapid prototyping. It is worth noting that most frameworks used in research for gesture recognition are extensions or reuse of frameworks initially developed for activity or emotion recognition or for pervasive computing.

Best-known frameworks include Context Recognition Network Toolbox (CRNT) [4], EyesWeb XMI [5] and Social Signal Interpretation (SSI) [6]. CRNT has been developed to facilitate the rapid construction of multimodal context recognition systems and their rapid deployment in targeted environment [4]. The framework provides a set of tools and modules for the connection of devices (drivers), the processing of data and some pre-implemented algorithms and machine learning systems. It has notably been used to acquire the dataset of the Opportunity Challenge where multiple sensors distributed in the environment and on a user were synchronously recorded in an everyday life scenario [16]. EyesWeb XMI has been developed as a toolbox and visual interface for processing full-body human movements and interaction in the context of emotions and non-verbal expression recognition [19]. A collection of software modules has been developed and released publicly: the "EyesWeb Expressive Gesture Processing Library". This collection implements models and modules to handle the processing of data for motion, space and trajectory analysis. The main research directions for the framework include analysis and classifications of expressive actions in musical signals and human movements, real-time generation of visual and audio content based on results of data analysis and user interactions. The framework has been widely used in dance and artistic fields due to the simplicity of its visual "drag&drop" user interface [5]. SSI has been developed to complement existing tools by providing support for the development of online recognition systems in the context of multiple sensors. It has been specifically tuned for machine learning pipelines and the acquisition of data [6]. The main research directions for SSI are the analysis of physiological signals in real-time; notably expressivity in user speech and affective recognition from video. The framework has later been extended to cover a larger area of research and to facilitate the manual annotation of recorded data through a specific interface [20]. Our proposed framework bears much resemblance with SSI but it has been specifically designed for gesture recognition and for supporting the creation and management of datasets in the context of multiple sensors and/or modalities.

2.2. Corpora for gesture recognition

Most of the research performed in the field of gesture recognition relies on corpora to develop, to train, to optimize and to evaluate algorithms and systems. In the context of gesture recognition, a corpus consists in a machine-readable collection of labeled files containing data such as video, text or audio which characterize the gestures. The files are labeled with meta-data describing the gestures temporally and/or spatially and potentially also describing the context of the data acquisition and information about the subject; this meta-data is referred as ground truth. In this article and in the machine learning literature, corpora and datasets refer to the same concept and can be freely interchanged. Developers either create their custom datasets or reuse datasets created by other research groups when possible. The advantages of using an existing dataset are not negligible: beyond the time and cost spared by not acquiring the dataset, it also provides means to quantitatively compare the results on common material (benchmark platform).

A few surveys have been published for the field of human action and activity recognition and some of them specifically on the available datasets; however the sub-domain of gesture recognition has been explicitly omitted from those surveys of datasets for comprehensibility [7]. Recently, a startup called *ARB Labs* has tried to initiate a commercial transition in the field by selling its corpus of gestures to companies and researchers [21]. Note that the KUG database already attempted a first step in this direction in 2006 [22]. A recent and comprehensive survey reviews the video datasets available for human action and activity recognition [7] and another article listed the few datasets available for multimodal activity recognition [23]; however a comprehensive survey specifically dedicated to corpora for the field of gesture recognition was still missing. A few recent research papers have been focusing on theoretical and practical considerations required in order to produce quality datasets for research. In [24], they identified three requirements: a natural gesture set, a gesture set containing enough instances of each class and an analysis of the recording conditions and their potential effects. In [25], they studied the impact of using particular semiotic modalities such as text, image or video during subjects instructions on the quality of the acquired gestures as training data for machine learning algorithms. The study highlighted the need to have a trade-off between coverage and correctness of the motion of the acquired gestures for optimal learning performances of the algorithms.

A review of the most recent and/or popular datasets for gesture recognition is shown in Table 1. It resumes their main characteristics: sensors, recording conditions, involved body-parts, ground truthing methods and material, applicability and availability. The table illustrates the increase in the number of datasets being released these last years, mainly due to the novel sensors but also due to the growing number of potential fields of application. By analyzing the corpora listed in Table 1, several considerations can be postulated. A rapid change of the main vision sensors used in the field is clearly visible: from color cameras and webcams to depth cameras. Three main types of datasets can be distinguished according to the "body-parts" visible from the sensors: hand(s) only for hand gesture recognition with only the hand(s) and sometimes part of the arm(s) being visible from the sensor [26–30], upper-body where the subjects use their arm(s) and hand(s) to interact [31–36] and full-body where the whole body is tracked and potentially used for interaction [25,37,38]. Note that upper-body and sometimes full-body may also involve tracking the head and the face of the subject. The intended domain of applicability of each dataset also varies depending on the type of ground truth data available, recording conditions and the chosen set of gestures. The datasets are mainly produced for the training and the evaluation of recognition algorithms and therefore include at least the name of each gesture as ground truth. Some datasets also include temporal segmentation and can be used for gesture spotting [31,33–35,37,38], meaning that it allows researchers to train and evaluate algorithms to automatically detect the start and end of gestures. They can also contain 2D or 3D spatial position of body-parts in the ground truth and therefore may be used to train

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

Table 1
Dataset survey: the table lists the most recent and popular corpora in the field of gesture recognition. The main characteristics of the datasets are listed. The abbreviation for the characteristics can be read in the headers of the columns (table extended from Ruffieux et al. [39]).

Corpus name	Year	Body-parts (Full-Body, Upper-Body, Arms, Hands)	Sensor view (Front-View, Top-View, Lateral-View, Moving-View)	User position (Standing, Sitting, User Moving)	#Classes	#Instances	Sensors (Color, Depth, Skeleton, Sound)	Ground truthing (Automatic, Semi-Automatic, Manual, User, Not Available)	Ground truth (Gesture Labels, Temporal Segm., Spatial Segm.)	Applicability	Availability (Public, Public on Request, Not Yet)
ASL Dataset [31] "NoName"	2013	UB	FV	St.	1300+	1300+	$Kinect^{CD}$ 640 × 480@25 Hz	M	GL TS SS	Hand detection and tracking; gesture recognition	NY
CGD2013 [40] ChaLearn Gesture Dataset	2013	FB	FV	St.	20	13000+	$Kinect^{CDSSo}$ 640 × 480@20 Hz	M	GL TS	Challenge "Multiple instances, user independent learning"	P
ChAirGest [32]	2013	UB	FV	Si.	10	1200	4 IMU (50 Hz) $Kinect^{CDS}$ 640 × 480@30 Hz	A	GL TS	Multimodality and fusion evaluation; temporal segmentation	PR
SKIG [26]	2013	H and A	TV	Si.	10 (30)	1080	$Kinect^{CD}$ 320 × 240@10 Hz	N/A	GL	Improve algorithms taking advantage of RGB-D to recognize gestures	P
3DIG [36]	2012	UB	FV	St.	20	1739	$Kinect^{CDS}$ 640 × 480@30 Hz	SA	GL	Iconic gestures (primitive and complex objects)	P
6DMG [27]	2012	H	–	–	20	5600	Wii Optical Tracker 60 Hz	U	GL TS SS	Implicit vs. explicit information	P
CGPD12 [25] (MSRC-12)	2012	FB	FV	St.	12	6244	$Kinect^{S}$ 30 Hz	N/A	GL $TS^{(start)}$	Recognition evaluation, variations across users	P
G3D [37]	2012	FB	FV	St.	20	600	$Kinect^{CDS}$ 640 × 480@30 Hz	M	GL TS	Improve recognition of gestures for games	PR
MSR datasets [28] (MSRGesture3D)	2012	H	FV	St.	12	336	$Kinect^{D}$ ~130 × 130@20 Hz	N/A	GL	Explore depth data for ASLR focusing on hand only	P
CGD2011 [35] ChaLearn Gesture Dataset	2011	UB	FV	St.	30	50000	$Kinect^{CD}$ 320 × 240@10 Hz	M	GL TS $SS^{partial}$	One-shot learning for gesture recognition systems	P
NATOPS Aircraft Handling Signals Database [33]	2011	UB	FV	St.	24	9600	Stereo Camera 320 × 240@20 fps Vicon System (1 subject)	M	GL TS SS (1 subject only)	Gesture recognition and segmentation Skeleton tracking evaluation Simultaneous body and hand (pose) gesture	P PR
NTU Dataset [29]	2011	H	FV	Si.	10	1000	$Kinect^{CD}$ 640 × 480	N/A	GL	Cluttered background, accuracy and efficiency	P
Keck Gesture Dataset [38]	2009	FB	FV MV	St. UM	14	126 (static) 168 (moving)	Color camera 640 × 480@15 Hz	N/A	GL TS	Compare algorithms (notably resistance to perturbations)	P
ASLLVD [34] The American Sign Language Lexicon Video Dataset	2008	UB	1FV 1LV 1HV	St.	3300+	9800	3 Color cameras 640 × 480@60 Hz	M	GLTS	ASL reference database	PR
CHGD [30] Cambridge Hand Gesture Dataset	2007	H	TV	Si.	9	900	Color camera 320 × 240	M	GL	Hand segmentation; gesture recognition	P

and to evaluate tracking algorithms [27,31,33,35]; the hands and head are the body-parts most commonly tracked. A few datasets also contain varying or cluttered background, varying light conditions or additional subjects wandering in the background in order to test reliability of algorithms to perturbations and real conditions [29,38]. Some datasets also focus on specific types of gestures such as sign language [28,31,34], gaming gestures [37], iconic/deictic gestures [36] or hand pose [29].

It is also interesting to observe the ratio "number of instances per class" which largely varies between datasets; generally, the more classes collected the less number of instances per class. For training and evaluations of algorithms, a high ratio is generally required to obtain valid results or to generalize them [24]. Another interesting piece of information indicates how ground truth data has been acquired, a characteristic which is generally not clearly explained; because most corpora are annotated manually by experts after the recordings. The different ground truthing methods are detailed in Section 2.3. In the above list of datasets, seven have been annotated manually, one automatically, one semi-automatically, one has been annotated by users during data acquisition and for five datasets the method has not been reported although they probably have been manually annotated. Note that the only semi-automatically annotated dataset provides only the name of the gesture as ground truth, which corresponds to the minimal ground truth information. In this article, we present a novel automatic scripted ground truthing approach which has been used to annotate the name and the temporal segmentation of the gestures occurrences contained in the ChAirGest dataset [32]. This approach is described in details in Section 4.

2.3. Ground truthing: tools and approaches

Various software and approaches have been developed to facilitate the ground truthing task. Most of them have been initially developed in the context of voice and video processing and then reused or extended for human action, activity and gesture recognition. This section reviews some of the software, user interfaces and tools that have been applied for ground truthing and then lists and reviews the existing approaches developed to facilitate ground truthing and illustrate each of them with examples of application from research fields close to gesture recognition.

Several software have been developed in order to facilitate the manual ground truthing of large corpora containing multimodal and/or multi-sensors data.

ANVIL is a platform-independent free research tool initially developed to annotate audiovisual data containing multimodal dialogue [41]. Since then it has been widely used and extended amongst the research community for various research fields. ANVIL defines a specific annotation format which enables to hierarchize the information in multiple tracks and to create several depths of annotation such as words spoken, head movement, hand movement, and gesture name. Such a coding system also provides the possibility to add new levels of annotations on pre-existing files. Several studies have been carried to obtain an intuitive interface in order to simplify the labeling process by providing multimodal functionalities. ANVIL has recently been extended to visualize motion capture data using a 3D skeleton representation [42].

ELAN is a platform-independent professional linguistic annotation tool [43]. It has been designed to create and manage text annotations for audio and video files with a strong focus on annotation of language. Similarly to ANVIL, annotations are grouped in layers that can have several relationships between each other: independent, aligned or embedded. The development of the software focused on providing a precise time accuracy of the labeling through specific tools and interfaces.

ViPER is an open-source framework that enables ground truth annotation of video data through a visual interface and also provides systems to evaluate algorithms performances [44]. The toolkit mainly offers two distinct programs: ViPER-GT consists in a graphical user interface (GUI) developed in Java for video annotation focusing on spatial and temporal labeling and ViPER-PE is a command-line performances evaluation tool offering various performance metrics to compare algorithms. SSI is an open-source C++ framework for multimodal signal processing in real-time [20]. It has been developed for human behavior recognition in multimodal contexts and handles most of the steps required for the production of a corpus: processing, synchronization and recording of sensor signals using specific blocks, modules and pipelines which can be easily defined in XML files. A specific user interface has also been developed to visualize and annotate the recorded multimodal data offline.

Recently, VATIC has been developed as an open platform for monetized crowdsourcing of video ground truthing using Mechanical Turk systems [45]. It provides tools to facilitate the development of web-based labeling interfaces in order to take advantage of crowdsourcing possibilities. Finally our framework is being developed to handle all the processes involved when developing multimodal applications [46], including specific tools for facilitating the acquisition of multimodal corpora, as further described in Sections 3 and 4.

In literature, five main approaches can be distinguished to produce the ground truth annotations for videos and multimodal datasets: semi-automatic, crowdsourced, user-annotated and automatic. The following paragraphs describe each approach and illustrate them with practical examples.

The first approach is manual ground truthing; it is the most used method in research, particularly for small project-specific datasets but also for larger datasets. In most cases, manual ground truthing is performed offline by one or more experts spending hours annotating the videos, generally frame by frame, using custom tools or existing framework [47]. Manual ground truthing may also be partially done online during the data acquisition process by observers annotating the events in real-time [48]. Researchers often release partially labeled datasets due to the time required for the manual ground truthing of all the data [49]. Researchers have also worked at making this process faster using better interfaces or lighter representations of the data such as 3D skeleton representation instead of video [42]. In gesture and activity recognition, the manual segmentation often implies splitting videos and/or data in small files where each file contains a single occurrence of an event to recognize; this file segmentation simplifies the ground truthing and the data management processes [35]. The disadvantage of such method is that it removes the possibility to train or evaluate spotting algorithms due to the absence of temporal segmentation. Many datasets have been manually ground-truthed although the information was not reported in related publications.

The second approach is the semi-automatic ground truthing which consists in partially labeling the data using specific algorithms and then requesting an expert to correct, validate or extend the results of the algorithms. This method has been used for face land marking and sign language recognition where annotators were only requested to validate the resulting output of the algorithms [50,51]. In the context of object tracking and recognition in videos, one semi-automatic method consists in providing facilitating tools to the annotators for automatically extracting the contours of objects in each frame [52]. A different approach in the context of multiple subjects interacting for human–robot interaction consists in several steps of annotations [53]. The first step consists in using scripted scenarios to partially label the data during the recordings; the second step uses algorithms to automatically

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

augment the labeling with spatial positions of actors and with a textual translation of their speech; finally, in the last step, human annotators validate the automatic labeling of algorithms and also manually augment the labeling with the actions performed by the subjects.

The third approach consists in crowdsourcing the manual annotation of the datasets. This approach has been popularized these last years with the apparition of crowdsourcing marketplaces on internet such as Amazon Mechanical Turk [54]. On these marketplaces, human workers can be hired to complete a specific task online. This approach is still young and the advantages and disadvantages of using non-expert and untrained annotators have only been partially studied [55]. A recent study showed that using this approach was applicable and efficient for segmenting and labeling activities in videos sequences using specific filtering methods to identify and remove non-serious workers [56]. The interface and requirements to produce good quality annotations for complex videos with dense and closely cropped labels have been studied in Vondrick et al. [45], where they demonstrated that crowdsourcing only should not be used for annotation of videos and that computers should assist humans. Furthermore they have shown the importance of the design of a clear annotation protocol to obtain high quality labeling in the context of crowdsourcing.

The fourth approach consists in user annotations and relies on the subjects to annotate autonomously their data at the time of acquisition. This is only possible in particular contexts and highly relies on the good will of the users. This approach has been applied for activity recognition in real-life using smartphones as input sensors: in Kawaguchi et al. [57], they developed an application where the users or researchers can autonomously annotate accelerometer data in real-time on their smartphone while the activity is being performed. The produced data can then be submitted online. Another project developed the same approach for static object/scene recognition from 3D depth data using a Kinect™ sensor: users can download a specific software enabling the acquisition and labeling of the data acquired by a 3D depth sensor and then submit their scans using a dedicated website [58]. Another classical user annotation approach consists in the user pressing a button in real-time to indicate the start and end of the gestures that he performs. However, from our experience, user annotation tends to produce many errors: subjects often label the class incorrectly or simply forget to press the button to indicate start and stop of gestures. Correcting the errors manually offline may also be a time-consuming task.

The fifth and more challenging approach consists in methods enabling the fully automatic ground truthing of the data. Several methods have been developed to automatically annotate the data: creation of known data through programs or scripts, using devices with high precision to produce ground truth data concurrently with less reliable sensors or using high confidence algorithms to produce reliable ground truth data. In [59], they use computer animation to produce scripted 3D realistic scenes in order to obtain pre-labeled video surveillance datasets. In [60], they use a specific program to produce realistic hand images from multiple viewpoints to obtain a dataset for hand pose recognition. Using scripted-scenarios is also an interesting method; scripts can be produced by the researchers [61] or researchers can reuse existing video sequences where scripts are already available such as for movies. In [62], they use sequences from Hollywood movies and parse the scripts of the movies to produce the ground truth. Another approach consists in using precise motion capture systems to produce reliable tracking data as in Sigal et al. [63]; this method has often been used for action and activity recognition but requires expansive and constraining setups. The last approach consists in relying on algorithms to automatically label or to extend the labeling of data; it has notably been applied for enhancing sign language annotations based on state-of-the-art automatic sign language recognition algorithms [64] or for video annotations based on speech and object recognition [65]. Note that most of these automatic ground truthing algorithms rely on pre-existing labeled datasets for their training and high confidence algorithms.

3. FEOGARM framework

The FEOGARM framework has been built in order to provide tools and methods allowing developers to handle the complete chain of operations when developing a gesture-based multimodal application: from designing and building corpora to the evaluation of the performances of the developed algorithms while maintaining functionalities for rapid prototyping of applications. The aim is to provide a single modular and reusable framework to facilitate the development, training, testing and deployment of application or prototypes for gesture recognition. The added value of the framework is that it has also been built to support the creation and management of corpora and to facilitate the quantitative evaluation of gesture recognition algorithms. The general goal is to simplify the acquisition and processing of data and the comparison of algorithms in the context of multimodal and/or multi-sensors gesture recognition.

Fig. 1 briefly resumes the generic chain of operations required when building a new gesture-based prototype application. The first line refers to the conceptualization of an application: several sensors are considered, a specific vocabulary of gestures is chosen, several algorithms are envisioned, evaluation measures are defined and potentially the final intended application is specified. The second line represents the practical operations that have to be developed to build a prototype application usually based on live trials and reduced dataset. These operations are supported by most frameworks reviewed in this article, including FEOGARM. Finally the third line represents the extended practical operations supported by FEOGARM where more emphasis has been put on the corpus thus facilitating the comparison and optimization of algorithms in the next phases.

The FEOGARM framework, similarly to other rapid prototyping frameworks, facilitates the work of the developers by providing tools and modules to rapidly produce a working prototype application. The added value of our framework is that it also supports specific tools and modules to handle the recording, management and ground truthing of a dataset, data visualization and analysis modules and methods to compare different algorithms or sets of sensors. One of the main difference is that most frameworks focus on the rapid development of prototype applications in order to demonstrate the feasibility of a concept while our approach aims at producing applications where multiple algorithms or combination of sensors can be tested and quantitatively compared on one or more datasets in order to produce an optimal experience in the final application and we therefore claim that our framework can handle the whole chain of operations.

3.1. Description

The framework has been developed in C++/C# and exhibits the standard pipelines, modules and distributed architecture shared amongst most frameworks for multimodal applications [17]. FEOGARM has been built upon ARAMIS, a framework developed in our research group and intended for contextual hybrid gesture multimodal interaction [66]. The FEOGARM framework has been developed for experienced developers; therefore it does not provide a simplified user interface to prototype applications but enables full customization and modification of every modules and drivers through well-documented code and examples.

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

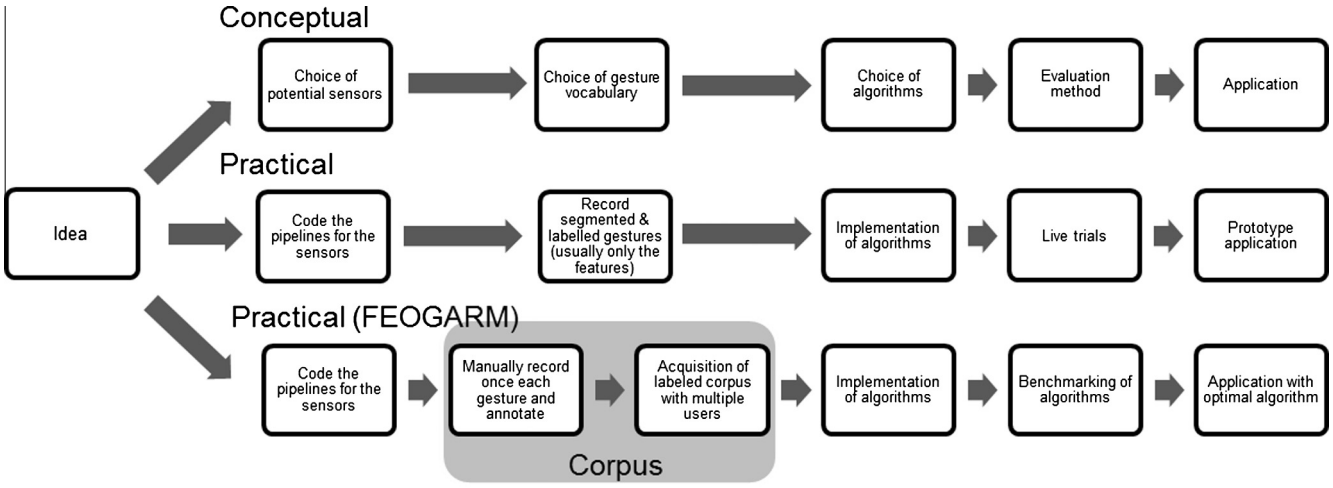

Fig. 1. Chain of operations: conceptual and practical chain of operations when building an application based on gesture recognition. The gray background on the third line (practical evaluation) corresponds to the corpus creation and management.

For each application, a new project must be created: a specific class defines the global structure by specifying the modules and their interconnections as well as potential options. As illustrated in Fig. 2, the framework supports several drivers, processing, recording and evaluation modules and simple graphical visualization interfaces for sensor data or algorithm performances. It also provides specific graphical user interfaces for data *producers* and for data *consumers*: the data producer graphical user interface facilitates the process of acquiring ground-truthed corpora through a presentation style interface which is further described in Section 4. The consumer graphical user interface provides tools to manage and visualize the data of the corpora; the visualizer notably provides an interface to visualize the data streams and the ground truth synchronously and specific tools to analyze the data contained within the corpus. To handle recording and replaying of multimodal corpora containing multiple sensors, some internal processing is performed during acquisition to temporally tag the data streams recorded from each distinct sensor and then an automatic resynchronization of the data is performed when reloading the data. These mechanisms ensure temporal synchronization even if the sensors have different frame rates. Furthermore the process is transparent for the consumers. A current limitation of this solution is that the corpus must have been recorded by the framework in order to be correctly replayed with our system.

The framework requires the implementation of specific drivers for each new sensor. All drivers are based on a common structure and only a few methods that are required to obtain the data from a specific sensor must be modified. The processing modules allow developers to record, to replay or to visualize the data streams from the sensors. They also enable to automatically process the data in real-time: features extraction, filters and segmentation modules. It is worth noting that additional modules can easily be coded and then reused. Classifiers modules correspond to machine learning algorithms which can be trained, used in real-time and evaluated within the framework. Currently several types of Hidden Markov Models (HMM), Support Vector Machine (SVM) and Neural Networks (NN) have been implemented. They are based on the open-source Accord.NET framework [67]. The FEOGARM framework is currently only available internally and should be released as an open-source project in 2014.

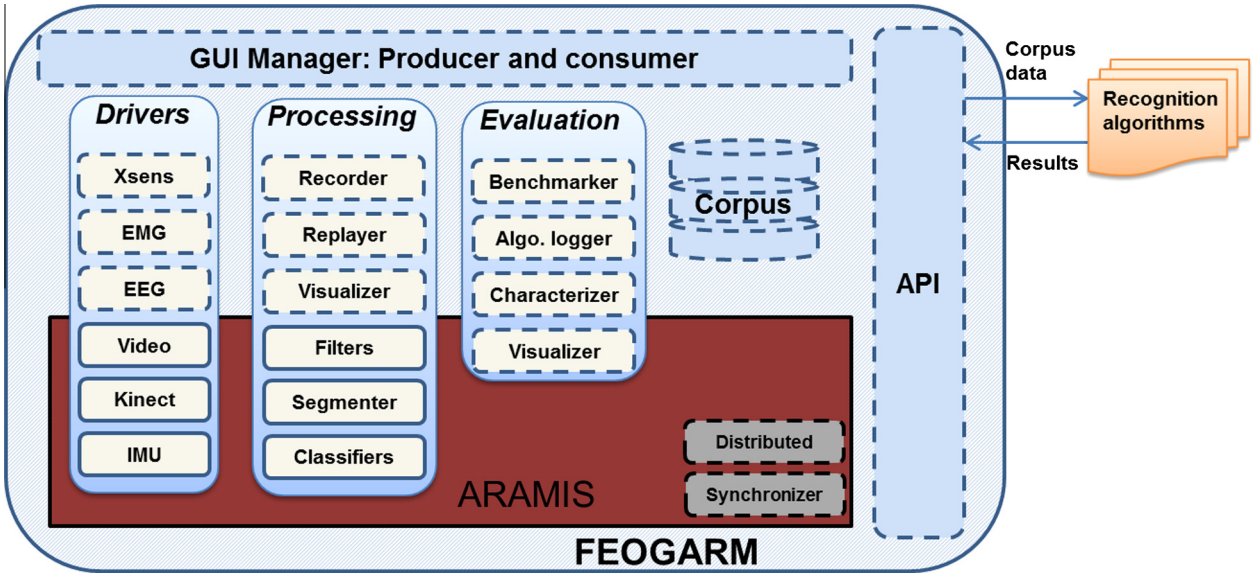

Fig. 2. Framework overview: this schema provides an overview of the FEOGARM framework and its modules.

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

3.2. Application examples

3.2.1. Rapid prototyping: wheelchair natural pointing system

A Natural Pointing System has been developed based on previous work and implemented using the FEOGARM framework [3]. The developed system facilitates the interaction between people with mobility impairments and their surrounding environments; two different paradigms have been evaluated; interaction through a touch interface using a smartphone and through natural pointing gestures. The implementation of the natural pointing detection system [68] in the application has been facilitated by the rapid prototyping features of our framework. It enabled to easily interconnect the different devices required for the experiment and to rapidly perform the evaluation procedure in a laboratory setting.

3.2.2. ChAirGest corpus

The ChAirGest corpus and the related open challenge have been respectively acquired and managed using the FEOGARM framework [32]. The steps in order to produce the corpus have been largely facilitated through the use of the framework. The first step was to design the corpus by choosing the desired physical setup, the set of sensors and the set of gestures for the corpus. Based on these choices, a specific application has been implemented interconnecting the drivers, recording modules and the visual *producer* interface. Once all the pieces connected, preliminary tests have been performed to assess the usability of the application and to improve it. As further described in Section 4, a first recording of each unique gesture and their manual labeling allowed us to subsequently record and label automatically all the gestures performed by the subjects during acquisition. Note that the handling of the subject information, folder grouping of the acquired files and all the metadata were also managed by the framework. Finally the corpus visualization and management tools allowed us to check and clean the raw data rapidly and then to convert it into several lighter formats for release and distribution purposes. The data has been released in the context of an open challenge.[1] Several tools and an API have been released along with the data in order to allow the participants of the challenge to visualize, to access and to process the raw data of the corpus. Finally the evaluation modules have been used to quantify the results of the contestant of the open challenge and to provide quantitative feedbacks.

3.3. Discussion

FEOGARM follows the main standards and requirements for multimodal frameworks: modular, reusable and distributed. It is very similar to other existing frameworks but focus on developers and researchers rather than standard users. Most of the multimodal frameworks reviewed in this article focus on rapid application prototyping and deployment rather than corpus acquisition [4,6], some of them also provide specific interfaces to simplify the manual ground truthing task but none provide semi-automatic approaches [6]. Finally, most reviewed frameworks provide simplified visual interface to manage modules and their connections but do not always provide the possibility for experienced programmers to extend or create new modules easily [5]. The proposed framework has been developed for programmers and researchers: it contains specific modules to simplify corpora acquisition, semi-automatic ground truthing, algorithm development, evaluation of performances and instead of providing a simplified UI limiting the functionalities, specific templates of modules, application and code demonstrate the potential usage of the framework and allow developers to fully customize and extend the framework according to their specific needs.

The main strengths of the framework:

- Fully modifiable and customizable.
- Specific modules for corpus acquisition and management.
- Allow acquisition of data at full rate with Microsoft Kinect™ sensor.
- Modular and reusable programming paradigms.
- Distributed sensors synchronization mechanisms.
- Tools and libraries for algorithm development.
- tools and metrics for performance evaluation.

The main limitations of the framework:

- Works only on Microsoft Windows™ systems.
- Currently not available as an open-source distribution.
- No compatibility with other frameworks.
- Requires programming skills.

4. Scripted ground truthing

In the context of the acquisition of the ChAirGest corpus, the creation of the ground truth was a problem; indeed the ground truthing of the 1200 instances contained in the corpus would have required about 60 h of manual work to an expert according to our estimations and more according to previous research [48]. As we intended to record more than a corpus, we believed that a better solution than manual or crowdsourced ground truthing should be developed for the field of gesture recognition. We developed the directed corpus acquisition approach to reduce the time required to ground truth corpora in the context where users intentionally interact with a specific device or their environment.

4.1. Approach

The directed corpus acquisition approach described in this work can probably be categorized in the automatic approach even if it requires some manual labeling during the initial step. The proposed method simplifies the acquisition of large corpora for gesture recognition. It consists in a randomized scripted scenario which enables the automatic ground truthing during the data acquisition process of the subjects. Such scripted method may already have been applied for gesture labeling but has not been applied for temporal segmentation to our knowledge.

The notable advantages of the proposed method compared to other semi-automatic approaches are the possibility to label the sub-phases of the gestures [10] and to add distracter gestures in the dataset by providing the possibility for subjects to perform movements freely between acting the requested gestures. The temporal accuracy, precision and robustness of current spotting algorithms would not be sufficient to obtain a valid temporal segmentation using traditional semi-automatic or automatic approaches. The approach that we propose consists in performing once a controlled initial acquisition of every unique gesture present in the dataset and then to manually annotate each of them in order to be able to automatically label all subsequent recordings. The approach depicted in this paper has been applied to acquire a corpus dedicated to research on spotting and recognition techniques of mid-air gestures in the context of close human–computer interaction. The corpus contains 10 different gestures performed by 10 subjects. Each gesture is initiated from 3 resting postures: hands on table, hands on belly button and hands under chin. The whole corpus contains 1200 gesture occurrences although only 30 of them required manual ground truthing.

[1] https://project.eia-fr.ch/chairgest/.

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

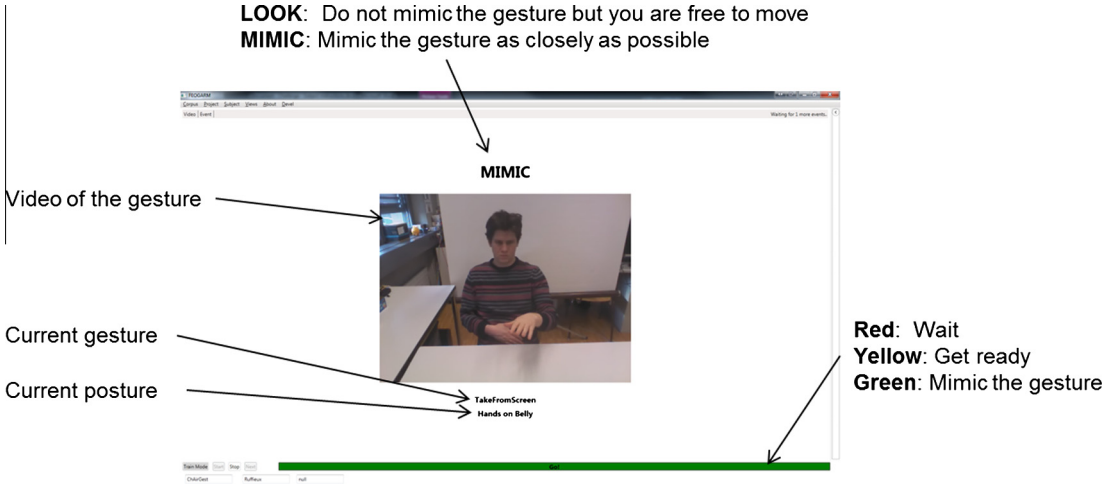

Fig. 3. Acquisition graphical interface: the graphical user interface as viewed by subjects during acquisition.

A custom module has been developed to handle the acquisition and the management of corpora. The application consists in a simple graphical user interface providing views to manage projects, subjects and recordings in the context of multimodal corpus acquisition. The user can create new projects and manage existing ones: each project defines the sensors used and the stimuli shown to the subjects using specific XML files. Through the interface, the user can also add subjects and their information to a specific project; all the information is stored in a structured XML file and can be browsed using the interface. Finally users can handle the recording process by selecting a project and a subject and then directly start the acquisition process through this same interface, shown in Fig. 3.

The exact acquisition and ground truthing approach can be resumed by the following three steps: baseline acquisition, baseline processing and iterative recording of the subjects. The baseline acquisition consists in recording each unique gesture contained in the dataset once by an expert "actor" who knows precisely the gestures to perform. In the context of our corpus, this consisted in recording the 10 gestures for each of the three resting posture; for a total of 30 unique gesture being recorded. The baseline processing consists in one expert splitting the 30 occurrences in short video sequences. Each video has been split to start and finish on an image representing a resting posture. Then the video sequences have been manually annotated with timestamps and labels in order to be replayed to the subjects during the acquisition.

The temporal labeling of a gesture is decomposed in three phases as described in previous research on hand gestures: pre-stroke, nucleus and post-stroke [10]. The segmentation of these three phases is illustrated on the left of Fig. 4: the event "1" corresponds to the start of the replayed video sequence, when the actor is in the resting posture; the event "2" corresponds to the start of the nucleus; the event "3" corresponds to the end of the nucleus and the event "4" corresponds to the end of the replayed video sequence, when the actor is back in the resting posture. Then during the acquisition of the gestures performed by the subjects, our approach takes advantage of the baseline labeled sequences previously created. These baseline videos are replayed in a randomized order to the subject during the recording sessions and the subject has to mimic the gestures shown on the screen. For each recorded gesture of the subject, the baseline video of the gesture is replayed twice consecutively: the first time, the subject is requested to watch the video without moving and the second time the subject is requested to mimic the movements in the video as accurately as possible. This two steps process is illustrated on the right of Fig. 4.

Note that the baseline videos are shown mirrored so the subject really has to follow the on-screen motion of the actor. As the baseline videos were previously annotated, the program takes advantage of these annotations to automatically label the recording of the subject as shown on the right of Fig. 4, where T'_0 (nucleus start event) and T'_1 (nucleus stop event) are labeled on the subject video by using the information from the replayed baseline video. When the baseline video starts, a signal containing the gesture information is sent to specific modules which are responsible to save the annotations along with the current "Frame ID" of the streams of the sensors. Assuming the subject imitates accurately the on-screen motion of the "actor", the gesture of the subject is automatically labeled. A small issue has been identified with this method: it produces a short time delay between the video being displayed and the user initiating the corresponding gesture; the time for the brain to interpret images and send commands to perform the motion. The validation will quantify this issue; however some preliminary methods have been applied to minimize this delay. The first method consists, as previously said, in displaying twice the gesture to perform. This method ensures that the subject understands and plans his motion before actually starting to perform it. The second method consists in providing visual and sound signals to let user anticipate the start of a mimicking gesture. The signals used in our method are a "traffic light" like colored bar (red, yellow and green) as shown in Fig. 3 and periodic sound signals similar to the one used in sports to indicate the start of a race.

4.2. Evaluation

In order to evaluate our automatic ground truthing method and its impact on recognition, we used two approaches. In the first approach, an expert manually annotated a subset of the dataset in order to provide a comparison in terms of temporal segmentation accuracy between a manual approach and the proposed approach. Then we investigated the impact of the temporal errors on the performance of a standard machine learning algorithm based on the data from the inertial motion units of the subset. In the second approach, we artificially shifted the ChAirGest dataset with controlled shift values and evaluated the impact on the recognition rate of a standard machine learning algorithm in order to determine the acceptable range of frame shifts.

4.2.1. Temporal inaccuracies analysis

A subset of the dataset has been manually ground-truthed by one expert; this subset is composed of all the occurrences for three gestures from three subjects. Each gesture occurs 12 times per

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

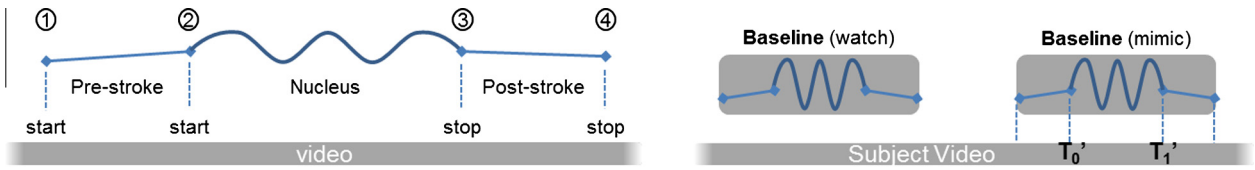

Fig. 4. Temporal segmentation. Left: temporal segmentation labeling format of a gesture. Right: temporal segmentation of the subject video using the baseline video sequences. The events "nucleus start" (T_0') and "nucleus stop" (T_1') are reported from the baseline sequence to the video recorded for the current subject.

subject; four times per resting posture, for a total of 108 occurrences. The three gestures that have been chosen for the subset are "WaveHello", "SwipeLeft" and "CirclePalmDown". They represent one complex gesture ("WaveHello") and two simpler gestures; the complex gesture is harder to imitate as its motion is faster and less constant and the number of forth and back movements may differ between baseline videos. The two other gestures are simpler to imitate when mimicking the video as they have slow and constant patterns of motion. The nucleus duration of the "CirclePalmDown" gesture is longer than the two others: about 3 s while the two other gestures last approximately 2 s.

The manual ground truthing of the subset was performed by an annotator using a software developed internally which provides an interface to visualize the RGB and depth video streams synchronized with the stream from the accelerometer located on the hand of the subject as shown in Fig. 5. We chose to use a custom program due to our custom raw data format; however the functionalities of our software are similar to standard manual annotation systems such as ANVIL or ELAN. The software has the capabilities to play and stop the data streams, move in the stream at the desired position using a slider and move frame by frame. The current Frame ID is displayed textually on the interface. A PhD student from our laboratory was chosen as an expert to annotate the subset. He received precise instructions on how to segment each gesture based on pictures. He had the task to label the name of the gesture, the pre-stroke start event when the hand was leaving its resting posture, the gesture start event when the hand was initiating the movement of the gesture, the gesture stop event when the gesture was finished before going back to the resting posture and finally the post-stroke stop event when the hand was reaching the resting posture. The expert took slightly more than 5 h to annotate these 108 occurrences which represent 9% of the complete dataset. Extrapolating this measure, the annotation of the complete dataset would have taken about 55 h. The expert annotator noticed that one subject (Subject 13) had the tendency to have larger delays than the others when performing the gestures. The two others had average to good performances.

We compared the results obtained from the automatic and manual segmentation methods using frame error and absolute frame error metrics for the start (T_0) and stop (T_1) events of the nucleus for each gesture of the subset as shown in Fig. 6. Note that, in this article, frame refers to a Kinect frame which corresponds to 33 ms and 1 Kinect frame is equivalent to 1.68 Xsens frame due to the differences of frame rate between sensors. We did not analyze pre-stroke start and post-stroke stop events in this study and concentrated on nucleus temporal segmentation labeling. The summary of the statistics obtained is shown in Table 2. By computing the average error over all three gestures, we can observe that, when using the automatic method, the nucleus of a gesture is labeled as starting 5.5 ± 5.5 frames earlier and stopping 0.2 ± 5.6 frames later than manual labeling which is considered as the truth. These results are illustrated for each gesture on the left of Fig. 7. This seems to indicate that users tend to partially catch up with the time delay occurring at the beginning of the gesture.

An in-depth analysis of all occurrences of the subset indicated that the average duration remains roughly constant amongst a same occurrence of gesture, indicating that subjects are generally consistent: if they start the gesture later, they also tend to finish the gesture later. By analyzing the average of absolute errors we can observe a difference of 6.5 ± 4.2 frames for the start event and a difference of 4.8 ± 4.2 frames for the stop event as shown on the right of Fig. 7. The average duration of the nucleus in the subset is 69 frames using the automatic method and 8% shorter using the manual annotation system; this indicates that the automatic system tends to consider additional data compared to the real gesture. Our analysis showed that most of the error is occurring at the nucleus start event.

4.2.2. Validation of the approach

To determine a potential impact of the temporal inaccuracies of the proposed method, we compared the performance of a classification algorithm using the manually annotated subset and the corresponding subset annotated by our approach. Both subsets contain the same 108 gesture data sequences with labels and temporal segmentation. The recognition rate for the 3 gestures has been evaluated by testing the same algorithm based on Hidden Markov Models (HMM) on each subset. The features sent to the HMM were the three-axis raw acceleration and raw angular velocity from the inertial motion units located on the wrist and on the elbow of the subjects. These 12 features were then processed by the HMM and tested using a repeated 10-fold cross-validation to obtain the final classification accuracy. Repeated K-fold cross-validation method is used to measure test error without sacrificing data; it consists in randomly partitioning the original data in K subsets of equal size and then use one subset as test set and the $K - 1$ others as training sets. The final accuracy corresponds to the average of the repeated evaluations. The algorithm was based on an ergodic HMM with four hidden states and implementing the Baum–Welch estimation algorithm for continuous sequences with a normal multivariate distribution [69]. For the training phase, the convergence threshold has been set to 0.0001.

The results of the evaluations showed a final recognition accuracy of 100% on both subsets, both confusion matrices exhibited a perfect diagonal, showing no error at all during recognition. This demonstrates that on the small subset containing three gestures, the temporal inaccuracies of our ground truthing approach are not problematic for recognition algorithms.

4.2.3. Impact of temporal segmentation inaccuracies on recognition

In order to see the potential impact of temporal segmentation errors on the whole dataset, we generated temporally pre-segmented datasets from the original ChAirGest. These temporally-segmented datasets contain only the data corresponding to the portion labeled as the nucleus part of a gesture. We introduced Gaussian frame shifts to the nucleus start and end events of the automatically ground-truthed dataset. We used a range of Gaussian means from −50 to 50 with a constant standard deviation of 5 for the Gaussian parameters to introduce random frame shifts. We used the Box–Muller method to generate the values [70]. The impact of the shifting is shown in Fig. 8 on a representative "WaveHello" gesture for the acceleration data of the inertial motion unit located on the wrist of a subject. The graphics clearly illustrate the

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

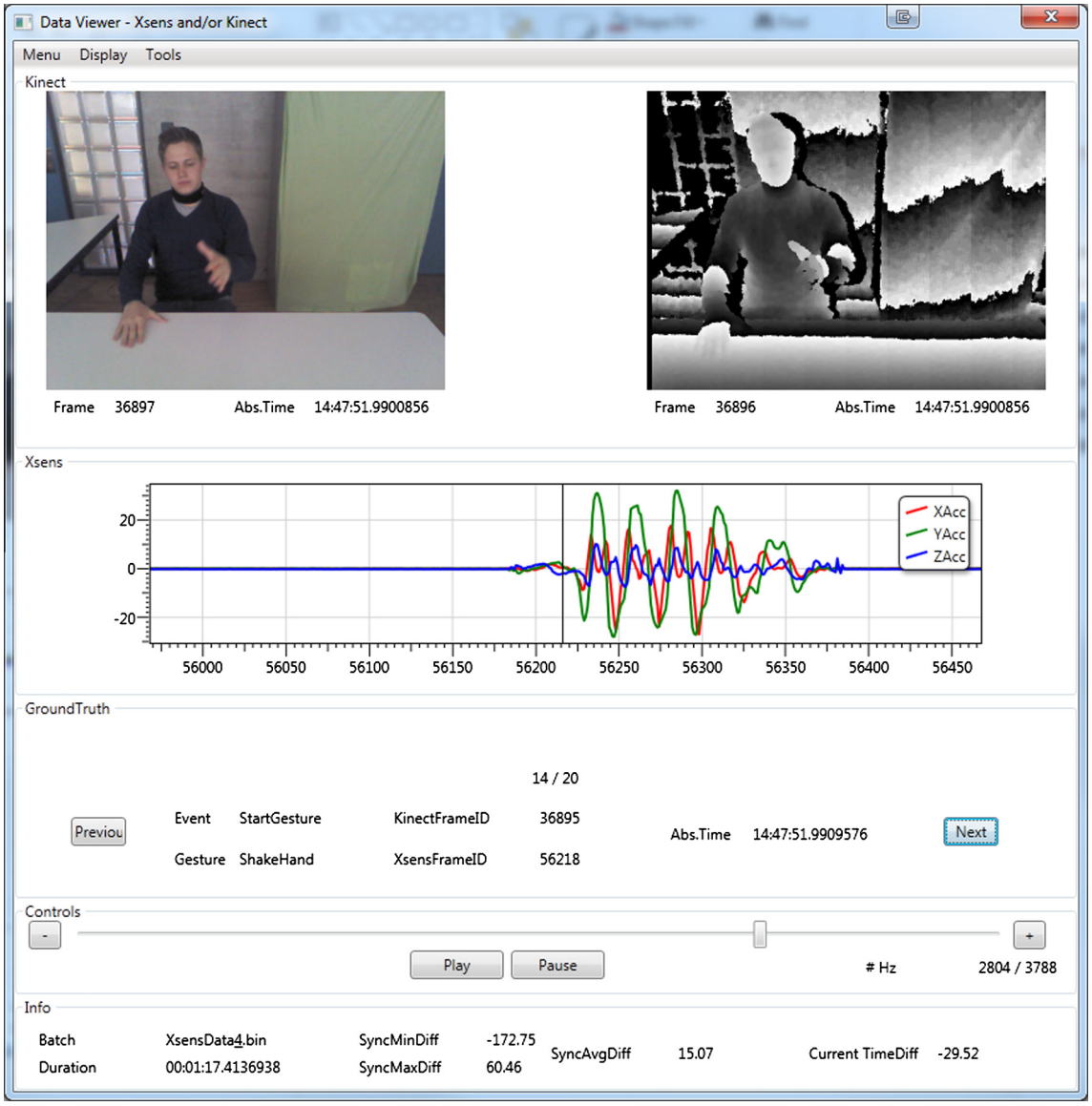

Fig. 5. Visualization interface: the custom software used by the expert for the manual ground truthing of the subset. The image on the interface illustrates the "Hands on table" resting posture of an upcoming "WaveHello" gesture. On the upper left and right section of the window, respectively the color and depth streams are shown. On the second section, the acceleration of the hand is shown with the current value indicated with a vertical black bar. Then ground truth data, controls and statistics are shown on the subsequent sections. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

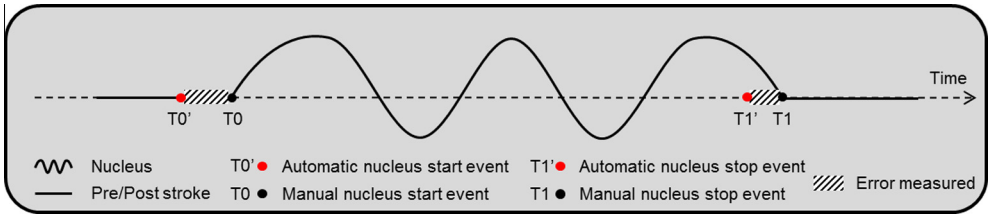

Fig. 6. Error measurement: this schema illustrates the error measured between automatic and manual ground truthing. A positive error indicates that the automatic event happened to early, while a negative error indicates that the automatic event occurred too late. Both error measured in this example are positive.

shift of the data and the corresponding partial loss of information from the region of interest (nucleus). We can also observe that the −50 mean frame shift contains less nucleus data than the +50 mean frame shift; probably due to the 5 frame shift error present in the dataset which is visible in Fig. 7.

Using the same HMM implementation as in the previous study, we analyzed the impact of the different frame shifts on the Macro F1-Score metric [71]. Five temporally pre-segmented datasets have been generated from original ChAirGest dataset for each shift value. The averaged results of the 10-Fold cross-validation for each shift value are shown in Fig. 9 with the averaged values (left) and their plot (right). The graphic shows that the best recognition score has been obtained for a mean frame shift of +5. The plot on the right of Fig. 9 shows that shifting the frames from −10 to +20 does

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

Table 2
Comparison of methods: the summary of the statistics for the three gestures considered in the subset. T_0, T_1, and T'_0, T'_1 correspond respectively to the manual and automatic ground truthing methods. T_0 corresponds to the beginning of the nucleus and T_1 corresponds to the end of the nucleus. The values in the tables represent the error in terms of number of Kinect frames (1 frame corresponds to ~33 ms).

	CirclePalmDown				SwipeLeft				WaveHello			
	$T_0-T'_0$	$\|\|T_0-T'_0\|\|$	$T_1-T'_1$	$\|\|T_1-T'_1\|\|$	$T_0-T'_0$	$\|\|T_0-T'_0\|\|$	$T_1-T'_1$	$\|\|T_1-T'_1\|\|$	$T_0-T'_0$	$\|\|T_0-T'_0\|\|$	$T_1-T'_1$	$\|\|T_1-T'_1\|\|$
Avg	5.1	6.1	−3.3	4.8	6.1	7.2	1.5	3.6	5.2	6.3	1.5	6.1
Max	19.0	19.0	9.0	16.0	19.0	19.0	15.0	15.0	16.0	16.0	12.0	14.0
Min	−6.0	1.0	−16.0	0.0	−7.0	1.0	−5.0	0.0	−9.0	0.0	−14.0	1.0
Range	25.0	18.0	25.0	16.0	26.0	18.0	20.0	15.0	25.0	16.0	26.0	13.0
St Dev	5.4	4.2	5.2	3.8	5.9	4.5	4.5	3.0	5.8	4.5	7.0	3.7

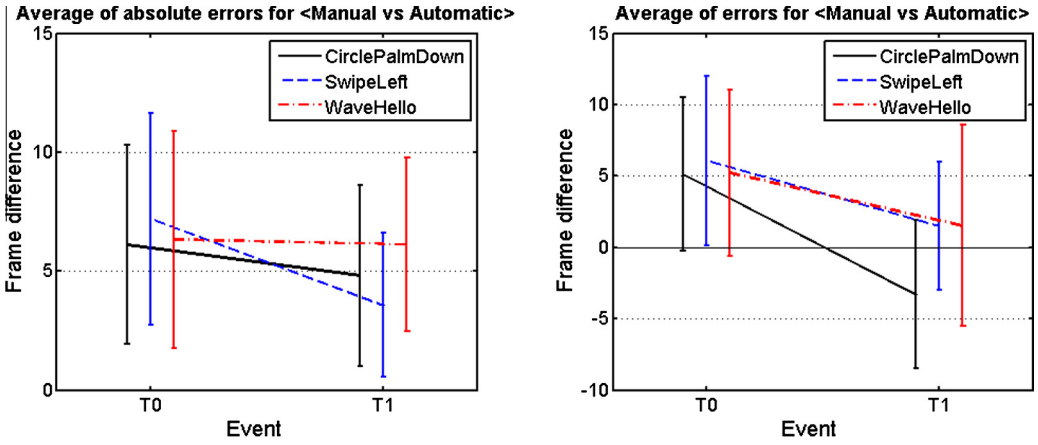

Fig. 7. Plots of errors: the two graphics represent the error in terms of frames between the automatic and manual segmentation methods for the start event (T_0) and end the event (T_1) of the nucleus parts of the gestures. For the three gestures, the average errors ($T_n - T_n'$) is shown on the left graphic and the average of absolute errors ($||T_n - T_n'||$) is shown on the right graphic.

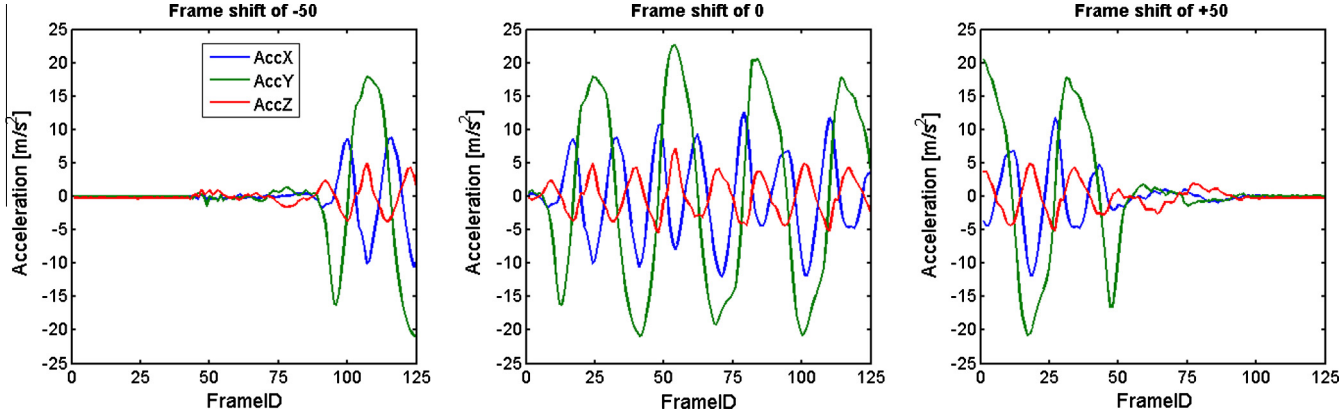

Fig. 8. Shift effect on data: these three plots illustrates the effect of the Gaussian shift on the data of a "WaveHello" gesture using a frame shift of −50 (left), no shift (middle) and a frame shift of +50 (right). The plots represent the acceleration data for the inertial motion unit located on the wrist of one of the subjects.

not produce a significant change of the recognition rate for our specific dataset. Two reasons may explain the absence of symmetry for the F1-Score between positive and negative shifts: the original dataset is already shifted due to the automatic ground truthing method as illustrated in left of Fig. 7 or "better" information is contained at the end of the nucleus.

Note that our analysis is valid for a dataset containing gestures which do not solely differ at the very beginning or end of the nucleus; applying an equivalent frame shift to such dataset would impair the recognition rate for much lower values of frame shift. Hypothesizing that our dataset is indeed shifted by 5 frames compared to reality; this would indicate that shifts due the automatic ground truthing method comprised between −15 and 15 frames are acceptable and do not significantly impair the recognition rate on our dataset.

4.3. Discussion

The novel ground truthing approach described in this article can be considered as an intermediary solution between semi-automatic and automatic approaches. The following paragraphs discuss the scripted and controlled acquisition scenario compared to an approach where the recordings take place during a natural interaction scenario and the specificities of our ground truthing method compared to other traditional methods. Finally, the last paragraph

Frame shift	Macro F1-Score
-50	0.548
-40	0.755
-30	0.776
-20	0.839
-15	0.889
-10	0.924
-5	0.926
0	0.927
5	0.933
10	0.932
15	0.929
20	0.922
30	0.891
40	0.827
50	0.700

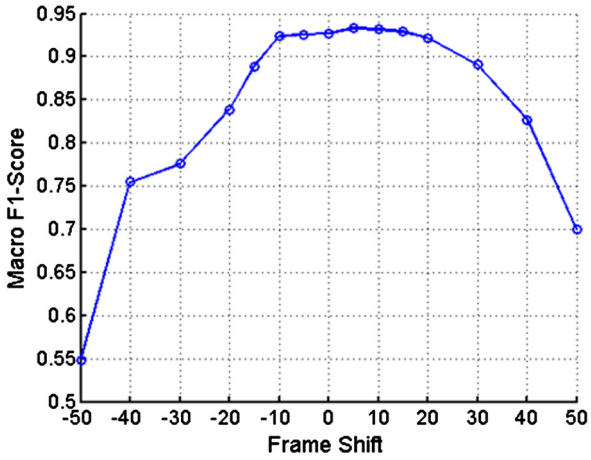

Fig. 9. Impact of frame shift on the recognition accuracy. Left: Macro F1-Score according to the mean frame shift of the Gaussian. Right: F1-Score plot.

of the section highlights the positive and negative facets of the proposed approach.

The proposed approach requires a controlled acquisition scenario; such controlled setup implicates several constraints compared to scenarios were the subject would interact naturally. Our main context of application is however limited to close human–computer interaction in which the user always intends his gestures and expects them to perform specific commands; the definition of the application context nuances the limitations compared to an application intended for broader contexts. In this context, the main advantage of the controlled approach is the possibility to automatically label the gestures and their sub-phases with a good accuracy and with only few labeling errors compared to a natural acquisition scenario where all gestures would need to be labeled manually a posteriori by experts and where an experimenter should potentially be present during every acquisition to be the "Wizard" of the graphical user interface [72]; assuming a Wizard of Oz experiment to yield a natural interaction scenario. Furthermore, the proposed approach does not disable the possibility to include distractors in the background such as moving objects or persons; it only has to be well planned beforehand such as for the changing light condition in ChAirGest [32]. The proposed approach also has several limitations and weaknesses. The constrained environment required during acquisition due to the subject facing a display and having to act exactly as planned largely reduces the possible scenarios of applications. Our approach requires a strong focus of the subject during the acquisition phases in order to perform the gestures timely with the videos. The main weakness of our approach is the reduced naturality of the gestures due to the subject being asked to accurately mimic on-screen gestures. As shown in the evaluation of the approach, the accuracy of the annotations is lower than a manual ground truthing approach.

As previously mentioned, the proposed ground truthing method can be considered as an intermediary solution between semi-automatic and automatic methods and is much more cost and time efficient than manual or crowdsourced methods. Compared to traditional semi-automatic methods, the proposed method limits the need of experts annotating missing labels or correcting the errors performed by an algorithm. An automatic approach based on algorithms would be complex due to current limitations of the existing spotting algorithms; correcting the results of such algorithms a posteriori would require extensive manual post-processing. An Automatic approach based on sensor data could have been considered; however it would have removed the possibility to label sub-phases of gestures and would have created problems with distractor gestures. The existing studies on crowdsourced approaches demonstrated several issues and they showed that crowdsourced approaches still need improvements and specific interfaces to obtain valid results. User-annotation is also an interesting approach although many practical studies showed its limitations in terms of number of errors and accuracy due to the increased cognitive task for the subject; additionally the gestures also tend to be less natural due to the subjects planning their gestures. To summarize, the proposed approach proposes to trade some naturality in the motion of the gestures, some potential contexts of application and some accuracy for a facilitated approach of corpora acquisition and annotation.

The main strengths of the approach:

- Time and cost efficient (efficiency increases with size of dataset).
- Temporal sub-phases labeling.
- Easily reproducible setup.
- Potential inclusion of distractors in the background.
- Distractor gestures allowed (like head scratching).

The main limitations of the approach:

- Limited to acquisition of command gestures.
- Subject must face a display.
- Subject must be focused during the whole acquisition.
- Controlled acquisition in laboratory settings.
- Limited temporal accuracy of annotation.

5. Conclusion

This article focused on three main area of the gesture recognition field: the available corpora, the frameworks and the ground truthing of datasets. A general conclusion of the article is presented and then the three topics are addressed in their respective order. Finally the last subsection points toward potential improvements and future works.

This article presented a framework supporting the rapid prototyping of multimodal applications, the creation and management of corpora and the development and quantitative evaluation of classification algorithms. The state-of-the-art and the analysis of available corpora highlighted the importance of a framework dedicated specifically to gesture recognition and supporting the creation and management of corpora through standards. The article also described a novel method that has been developed to facilitate

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

the cumbersome process of creating a corpus. The proposed approach enables the automatic temporal segmentation and labeling of gestures through the use of scripted scenarios to instruct subjects. The analysis of the temporal segmentation errors, produced by our method and compared to manual annotation, has demonstrated that it did not impact significantly the recognition rate of a machine learning algorithms based on a standard HMM. The proposed solution offers an efficient approach to reduce the time required to manually ground truth corpora of natural gestures in the context of human–computer interaction.

5.1. Gesture recognition, frameworks and corpora

The second section of the article reviewed the current state of the art in gesture recognition and highlighted the main requirements, notably in terms of corpora.

The differentiation of the gesture recognition field from activity and recognition highlighted the need for specific frameworks, toolboxes and corpora focusing solely on the field of gesture recognition. Specifically, high quality multi-purposes and multi-sensors corpora should be developed and released in order to provide quantitative benchmarks for the field. The structure of these datasets should be standardized in order to simplify reusability of developed algorithms and thus facilitate the comparisons amongst them. Better description and specifications of the publicly released datasets should also be encouraged in order to facilitate the choice of a dataset. We provided the reader a list of the most recent and popular datasets specifically designed for gesture recognition, highlighting their main characteristics and features.

5.2. FEOGARM framework

The third section described FEOGARM, a multimodal framework which has been developed to enclose all necessary operations when working on gesture recognition applications: creation and management of corpora, facilitating modules for development of algorithms, tools and novel metrics for performances evaluation and finally support for rapid prototyping of applications.

The proposed framework has been built using the standards for multimodal frameworks: distributed, modular, reusable and synchronization mechanisms between sensors. It has notably been developed with special attention to a currently popular depth-camera in research: the Kinect™ sensor. The framework supports the acquisition of the color, depth and skeleton streams at full rate and quality and the possibility to reuse the raw recorded data through standard methods from the official Microsoft Kinect™ SDK. Such framework should help developers to create standard datasets and share them to the community. Two projects that have already been completed using the functionalities of the framework have been presented; a prototype of application to study how to improve the life of people with disabilities through mid-air gestures and a dataset for gesture recognition. The advantages and limitations of the proposed framework have been discussed compared state-of-the-art framework. FEOGARM is still continuously being improved and augmented with new modules and tools; however it requires further commissioning in order to be publicly released as an open-source framework.

5.3. Scripted ground-truthing

The fourth section and central part of the article presented a time and cost-efficient method based on scripted scenarios to support the automatic generation of ground truth during the acquisition of datasets in the context of gesture recognition for human–computer interaction.

The method proved to be efficient in labeling and segmenting the gestures of the ChAirGest dataset. Using the data collected in our first evaluation and extrapolating them, we evaluated that an expert using a traditional approach such as manual labeling method to annotate the whole dataset would have spent about 55 h to complete the task. This roughly corresponds to a factor 30 compared to the proposed method when applied to our small dataset.

The analysis of the results demonstrated that the recognition tasks were not significantly impacted by the range of temporal segmentation errors produced by the proposed ground truthing method. The analysis of a subset of the ChAirGest dataset demonstrated that the proposed method produces less precise annotations than manual labeling with an average error of about 8% on the gesture duration. The analysis also showed a variable time-shift of the gesture of 180 ± 180 ms for the nucleus start event and of -4 ± 180 ms for the nucleus end event. Our evaluations demonstrated that the temporal segmentation errors occurring with the automatic method are not lowering the recognition rate of algorithms based on Hidden Markov Models on the specific subset used in this study. Frame shifts tests on the complete dataset demonstrated that shifts comprised between -15 and 15 frames do not significantly impact the recognition performances on our dataset.

Further tests should be performed with several other datasets, algorithms and type of gesture in order to generalize the results. The main advantages of the method presented in this paper are the huge gain in time and/or cost for the ground truthing (labeling and temporal segmentation) of large datasets, the low rate of errors of recorded subjects due to the simple mimicking task and the possibility to easily replicate the setup. The main disadvantages and limitations of the method are the constrained application to scenarized acquisitions where subjects have to mimic videos on a screen and thus potentially perform less natural gestures. Such system also tends to produce more homogeneity between occurrences of a same gesture because the temporal and the spatial motion of the subject are partially imposed by the videos.

5.4. Future work

This work provided a first solution toward the development of a simple automatic ground truthing approach for large corpora of gestures. Additional studies should be performed to fully validate and enhance the proposed method. A study should investigate the impact of mimicked gestures compared to intentional gestures when generalizing to real-life situations. A study should also evaluate the impact of the temporal errors of the proposed approach on the performance of spotting algorithms: how spotting and spotting and recognition algorithms are impacted by the temporal inaccuracy of the proposed method. Future studies should also further explore the possibilities to improve the proposed ground truthing method. A solution to automatically improve the accuracy of the segmentation could be to use an optimization algorithm based on a class separation metric as in Kirkham et al. [73]: by trying to slightly shift the temporal segmentation events it can be possible to find the optimal class separation. Similarly, the incorrect gestures (acquisition errors) could be automatically removed by using a gesture recognition algorithm and rejecting the ambiguous gestures.

Acknowledgments

This research has been supported by the Hasler Foundation within the project entitled "Living in Smart environments: Natural and Economic gesture-based HCI".

References

[1] T.A. Nguyen, M. Aiello, Energy intelligent buildings based on user activity: a survey, Energy Build. 56 (2013) 244–257.
[2] A.A. Chaaraoui, P. Climent-Pérez, F. Flórez-Revuelta, A review on vision techniques applied to human behaviour analysis for ambient-assisted living, Expert Syst. Appl. 39 (2012) 10873–10888.
[3] M. Caon, S. Carrino, S. Ruffieux, O. Khaled, E. Mugellini, Augmenting interaction possibilities between people with mobility impairments and their surrounding environment, in: Adv. Mach. Learn. Technol. Appl. SE – 18, Springer, Berlin Heidelberg, 2012, pp. 172–181.
[4] D. Bannach, O. Amft, P. Lukowicz, Rapid prototyping of activity recognition applications, IEEE Pervasive Comput. 7 (2008) 22–31.
[5] A. Camurri, P. Coletta, G. Varni, S. Ghisio, Developing multimodal interactive systems with EyesWeb XMI, in: Proc. 7th Int. Conf. New Interfaces Music. Expr. – NIME '07, 2007, p. 305.
[6] J. Wagner, F. Lingenfelser, E. André, The Social Signal Interpretation framework (SSI) for real time signal processing and recognition, INTERSPEECH (2011).
[7] J.M. Chaquet, E.J. Carmona, A. Fernández-Caballero, A survey of video datasets for human action and activity recognition, Comput. Vis. Image Underst. 117 (2013) 633–659.
[8] S. Mitra, T. Acharya, Gesture recognition: a survey, IEEE Trans. Syst. Man Cybern. C: Appl. Rev. 37 (2007) 311–324.
[9] T. Schlömer, B. Poppinga, N. Henze, S. Boll, Gesture recognition with a Wii controller, in: Proc. 2nd Int. Conf. Tangible Embed. Interact. TEI 08, 2008, p. 11.
[10] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human–computer interaction: a review, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 677–695.
[11] H. Hasan, S. Abdul-Kareem, Human–computer interaction using vision-based hand gesture recognition systems: a survey, Neural Comput. Appl. (2013).
[12] J.M. Palacios, C. Sagüés, E. Montijano, S. Llorente, Human-computer interaction based on hand gestures using RGB-D sensors, Sensors (Basel) 13 (2013) 11842–11860.
[13] J. Suarez, R.R. Murphy, Hand gesture recognition with depth images: a review, in: 2012 IEEE RO-MAN 21st IEEE Int. Symp. Robot Hum. Interact. Commun., 2012, pp. 411–417.
[14] R.A. Bolt, "Put-that-there", in: Proc. 7th Annu. Conf. Comput. Graph. Interact. Tech. – SIGGRAPH '80, ACM Press, New York, New York, USA, 1980, pp. 262–270.
[15] A. Jaimes, N. Sebe, Multimodal human–computer interaction: a survey, Comput. Vis. Image Underst. 108 (2007) 116–134.
[16] D. Roggen, F. Kilian, A. Calatroni, T. Holleczek, Y. Fang, G. Tr, et al., OPPORTUNITY: towards opportunistic activity and context recognition systems, Networks (2011).
[17] L. Tang, Z. Yu, X. Zhou, H. Wang, C. Becker, Supporting rapid design and evaluation of pervasive applications: challenges and solutions, Pers. Ubiquitous Comput. 15 (2010) 253–269.
[18] T. Weis, M. Knoll, A. Ulbrich, G. Muhl, A. Brandle, Rapid prototyping for pervasive applications, IEEE Pervasive Comput. 6 (2007) 76–84.
[19] A. Camurri, B. Mazzarino, G. Volpe, Analysis of expressive gesture: the eyesweb expressive gesture processing library, Lect. Notes Artif. Int. (2004) 460–467.
[20] J. Wagner, E. Andr, M. Kugler, D. Leberle, SSI/ModelUI – a tool for the acquisition and annotation of human generated signals smart sensor integration ModelUI, in: DAGA 2010, TU Berlin, Berlin, Germany, 2010, pp. 2–3.
[21] ARB Labs. <http://www.arblabs.com/> (accessed 10.23.13).
[22] B. Hwang, S. Kim, S. Lee, A full-body gesture database for automatic gesture recognition, 7th Int. Conf. Autom. Face Gesture Recognit., 2006, pp. 243–248.
[23] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Berkeley MHAD: a comprehensive multimodal human action database, IEEE Work Appl. Comput. Vis. (2013) 53–60.
[24] P. Glomb, M. Romaszewski, S. Opozda, A. Sochan, Choosing and modeling hand gesture database for natural user interface, in: Proc. 9th Int. Gesture Work., Athens, Greece, 2011, pp. 72–75.
[25] S. Fothergill, H. Mentis, P. Kohli, S. Nowozin, Instructing people for training gestural interactive systems, in: Proc. 2012 ACM Annu. Conf. Hum. Factors Comput. Syst. – CHI '12, 2012, p. 1737.
[26] L. Liu, L. Shao, Learning discriminative representations from RGB-D video data, Proc. Int. Jt. Conf. Artif. Intell. (2013).
[27] M. Chen, G. AlRegib, A new 6D motion gesture database and the benchmark results of feature-based statistical recognition, Emerg. Signal Process. (2012) 131–134.
[28] A. Kurakin, Z. Zhang, Z. Liu, A real time system for dynamic hand gesture recognition with a depth sensor, Signal Process. Conf. (2012) 1975–1979.
[29] Z. Ren, J. Meng, J. Yuan, Z. Zhang, Robust hand gesture recognition with kinect sensor, in: Proc. 19th ACM Int. Conf. Multimed. – MM '11, ACM Press, New York, New York, USA, 2011, p. 759.
[30] T.-K. Kim, S.-F. Wong, R. Cipolla, Tensor canonical correlation analysis for action classification, in: 2007 IEEE Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.
[31] C. Conly, P. Doliotis, P. Jangyodsuk, R. Alonzo, V. Athitsos, Toward a 3D body part detection video dataset and hand tracking benchmark categories and subject descriptors, in: Pervasive Technol. Relat. to Assist. Environ., 2013.
[32] S. Ruffieux, D. Lalanne, E. Mugellini, ChAirGest: a challenge for multimodal mid-air gesture recognition for close HCI, in: Proc. 15th ACM Int. Conf. Multimodal Interact. – ICMI '13, ACM Press, Sydney, Australia, 2013, pp. 483–488.
[33] Y. Song, D. Demirdjian, R. Davis, Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database, in: Proc. IEEE Int. Conf. Autom. Face Gesture Recognit., IEEE, Santa Barbara, 2011, pp. 500–506.
[34] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, A. Thangali, The American sign language lexicon video dataset, in: 2008 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., IEEE, 2008, pp. 1–8.
[35] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, H.J. Escalante, ChaLearn gesture challenge: design and first results, in: IEEE Conf. Comput. Vis. Pattern Recognit. Work., IEEE, 2012, pp. 1–6.
[36] A. Sadeghipour, L. Morency, S. Kopp, Gesture-based object recognition using histograms of guiding strokes, in: Proc. Br. Mach. Vis. Conf. 2012, British Machine Vision Association, 2012, pp. 44.1–44.11.
[37] V. Bloom, D. Makris, V. Argyriou, G3D: A gaming action dataset and real time action recognition evaluation framework, in: 2012 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., 2012, pp. 7–12.
[38] L.S. Davis, Recognizing actions by shape-motion prototype trees, in: 2009 IEEE 12th Int. Conf. Comput. Vis., IEEE, 2009, pp. 444–451.
[39] S. Ruffieux, D. Lalanne, E. Mugellini, O. Abou Khaled, A survey of datasets for human gesture recognition, in: M. Kurosu (Ed.), Human-Computer Interact. Adv. Interact. Modalities Tech. SE - 33, Springer International Publishing, 2014, pp. 337–348.
[40] S. Escalera, C. Sminchisescu, R. Bowden, S. Sclaroff, J. Gonzàlez, X. Baró, et al., ChaLearn multi-modal gesture recognition 2013, in: Proc. 15th ACM Int. Conf. Multimodal Interact. – ICMI '13, ACM Press, New York, New York, USA, 2013, pp. 365–368.
[41] M. Kipp, ANVIL – a generic annotation tool for multimodal dialogue, in: INTERSPEECH, ISCA, 2001, pp. 1367–1370.
[42] Q. Nguyen, M. Kipp, Annotation of human gesture using 3D skeleton controls, in: Proc. Int. Conf. Lang. Resour. Eval. Lr. 2010, European Language Resources Association, Valleta, 2010.
[43] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, H. Sloetjes, Elan: a professional framework for multimodality research, in: Proc. Lang. Resour. Eval. Conf., 2006.
[44] D. Mihalcik, D. Doermann, The Design and Implementation of ViPER, 2003.
[45] C. Vondrick, D. Patterson, D. Ramanan, Efficiently scaling up crowdsourced video annotation, Int. J. Comput. Vis. 101 (2012) 184–204.
[46] S. Ruffieux, E. Mugellini, D. Lalanne, O.A. Khaled, FEOGARM: a framework to evaluate and optimize gesture acquisition and recognition methods, in: Work. Robust Mach. Learn. Tech. Hum. Act. Recognition; Syst. Man Cybern., Anchorage, 2011.
[47] Z. Liu, Real world activity summary for senior home monitoring, in: 2011 IEEE Int. Conf. Multimed. Expo, IEEE, 2011, pp. 1–4.
[48] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, et al., Collecting complex activity datasets in highly rich networked sensor environments, in: 2010 Seventh Int. Conf. Networked Sens. Syst., IEEE, 2010, pp. 233–240.
[49] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.L. Reyes-Ortiz, Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine, Proc. 4th Int. Conf. Ambient Assist. Living Home Care, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 216–223.
[50] B. Barbour, K. Ricanek, An interactive tool for extremely dense landmarking of faces, in: Proc. 1st Int. Work. Vis. Interfaces Gr. Truth Collect. Comput. Vis. Appl. – VIGTA '12, 2012, pp. 1–5.
[51] R. Yang, S. Sarkar, B. Loeding, A. Karshmer, Efficient generation of large amounts of training data for sign language recognition: a semi-automatic tool, Comput. Help. People with Spec. Needs. 4061 (2006) 635–642.
[52] I. Kavasidis, S. Palazzo, R. Di Salvo, D. Giordano, C. Spampinato, A semi-automatic tool for detection and tracking ground truth generation in videos, in: Proc. 1st Int. Work. Vis. Interfaces Gr. Truth Collect. Comput. Vis. Appl. – VIGTA '12, 2012, pp. 1–5.
[53] X. Alameda-Pineda, J. Sanchez-Riera, J. Wienke, V. Franc, J. Čech, K. Kulkarni, et al., RAVEL: an annotated corpus for training robots with audiovisual abilities, J. Multimodal User Interfaces 7 (2012) 79–91.
[54] Amazon Mechanical Turk. <https://www.mturk.com/mturk/welcome> (accessed 11.20.13).
[55] P.-Y.P. Hsueh, P. Melville, V. Sindhwani, Data quality from crowdsourcing: a study of annotation selection criteria, in: Proc. NAACL HLT 2009 Work. Act. Learn. Nat. Lang. Process., Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 27–35.
[56] L.-V. Nguyen-Dinh, C. Waldburger, D. Roggen, G. Tröster, Tagging human activities in video by crowdsourcing, in: Proc. 3rd ACM Conf. Int. Conf. Multimed. Retr. – ICMR '13, ACM Press, New York, New York, USA, 2013, p. 263.
[57] N. Kawaguchi, N. Ogawa, Y. Iwasaki, K. Kaji, T. Terada, K. Murao, et al., HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings, in: Proc. 2nd Augment. Hum. Int. Conf. AH 2011, Tokyo, Japan, March 13, 2011, ACM, 2011, p. 27.
[58] A. Aydemir, D. Henell, P. Jensfelt, R. Shilkrot, Kinect@ home: crowdsourcing a large 3D dataset of real environments, in: 2012 AAAI Spring Symp. Ser., 2012, pp. 8–9.
[59] S. Brutzer, B. Hoferlin, G. Heidemann, Evaluation of background subtraction techniques for video surveillance, CVPR 2011 (2011) 1937–1944.

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004

[60] P. Natarajan, R. Nevatia, View and scale invariant action recognition using multiview shape-flow models, in: 2008 IEEE Conf. Comput. Vis. Pattern Recognit., IEEE, 2008, pp. 1–8.
[61] C. Jaynes, A. Kale, N. Sanders, E. Grossmann, The terrascope dataset: scripted multi-camera indoor video surveillance with ground-truth, in: 2005 IEEE Int. Work. Vis. Surveill. Perform. Eval. Track. Surveill., 2005, pp. 309–316.
[62] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, M. Marsza\lek, Learning realistic human actions from movies, in: Conf. Comput. Vis. Pattern Recognit., IEEE, 2008, pp. 1–8.
[63] L. Sigal, A.O. Balan, M.J. Black, HumanEva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, Int. J. Comput. Vis. 87 (2009) 4–27.
[64] P. Dreuw, H. Ney, Towards automatic sign language annotation for the ELAN tool, in: Lr. Work. Represent. Process. Sign Lang. Constr. Exploit. Sign Lang. Corpora, Marrakech, Morocco, 2008.
[65] V. Lavrenko, S.L. Feng, R. Manmatha, Statistical models for automatic video annotation and retrieval, in: 2004 IEEE Int. Conf. Acoust. Speech, Signal Process., vol. 3, IEEE, 2004, pp. 1044–1047.
[66] S. Carrino, E. Mugellini, O.A. Khaled, R. Ingold, ARAMIS: toward a hybrid approach for human–environment interaction, Proc. 14th Int. Conf. Human-Computer Interact. Towar. Mob. Intell. Interact. Environ., vol. Part III, Springer-Verlag, 2011, pp. 165–174.
[67] Cesar Souza, Accord.NET Framework, 2013.
[68] S. Carrino, A. Péclat, E. Mugellini, O. Abou Khaled, R. Ingold, Humans and smart environments, in: Proc. 13th Int. Conf. Multimodal Interfaces – ICMI '11, ACM Press, New York, New York, USA, 2011, p. 105.
[69] L.E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Stat. 37 (1966) 1554–1563.
[70] E.R. Golder, J.G. Settle, The Box–Muller method for generating pseudo-random normal deviates, Appl. Stat. (1976) 12–20.
[71] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manage. 45 (2009) 427–437.
[72] J.F. Kelley, An iterative design methodology for user-friendly natural language office information applications, ACM Trans. Inf. Syst. 2 (1984) 26–41.
[73] K. Kirkham, A. Khan, S. Bhattacharya, N. Hammerla, S. Mellor, D. Roggen, T. Ploetz, Automatic correction of annotation boundaries in activity datasets by class separation maximization, in: ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. HASCA Work., 2013.

Simon Ruffieux is a PhD student in cooperation between the University of Applied Sciences and Arts of Western Switzerland and the University of Fribourg. His PhD thesis focuses on developing methods for facilitating the research in gesture recognition. He completed his diploma in 2008 at the Swiss Federal Institute of Technology in Computer Sciences (BSc/MSc) with a specialization in Bio-Computing. He also works as a part-time scientific collaborator in the ICT Department of the University of Applied Sciences and Arts of Western Switzerland. His research domains include Human–Computer Interaction, Multimodal Interaction, Robotics and Systems Automation.

Denis Lalanne is a senior Researcher/Lecturer at the University of Fribourg, Switzerland. He is member of the DIVA group at the University of Fribourg. He completed his PhD in the Swiss Federal Institute of Technology in 1999, then worked as a postDoc in the USER group in IBM Almaden, as teacher and researcher in the University of Avignon and as a usability engineer in a Swiss start-up. His research domains include Human–Computer Interaction, Multimodal Interaction Engineering and Information Visualization.

Elena Mugellini is Professor at the ICT Department of the University of Applied Sciences and Arts of Western Switzerland, Fribourg. She is the leader of HumanTech research group (formerly MISG). She holds a PhD in Telematics and Information Society received from the University of Florence in 2006, and a Master in Telecommunication Engineering from the same university received in 2002. Her current research interests are on the areas of Ambient Intelligence, Multimodal Interaction, Tangible User Interface, Personal Information Management and Document Engineering.

Omar Abou Khaled is Professor in ICT Department of the University of Applied Sciences and Arts of Western Switzerland, Fribourg. He holds a PhD in Computer Science received from the Perception and Automatic Control Group of HEUDIASYC Laboratory of "Université de Technologie de Compiègne". He is Director of International Relations Office, Head of MRU "Advanced IT Systems Architects" at EIA-FR. Until 2007, he was leader of MISG (Multimedia Information System Group). He is responsible of several projects in the field of Document Engineering, Multimodal Interfaces, Context Awareness, Ambient Intelligence, Blended Learning, and Content-Based Multimedia Retrieval.

Please cite this article in press as: S. Ruffieux et al., Gesture recognition corpora and tools: A scripted ground truthing method, Comput. Vis. Image Understand. (2014), http://dx.doi.org/10.1016/j.cviu.2014.07.004