International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.5 (2015), pp.123-140
http://dx.doi.org/10.14257/ijmue.2015.10.5.13

A Systematic Review of Data Exchange Formats in Advanced Interaction Environments

Celso A. S. Santos[1], Estêvão B. Saleme[1] and Juliana C. S. de Andrade[1]

[1]*Department of Informatics, Federal University of Espírito Santo, Brazil*
saibel@inf.ufes.br, estevaobissoli@gmail.com, julianacristina.ti@gmail.com

Abstract

The advent of advanced user interface devices has raised the interest of industry and academia in finding new modes of Human-Computer Interaction. Advanced interfaces employ gesture recognition, as well as motion and voice capturing to enable humans to interact naturally with interactive environments without utilizing any of the traditional devices like mice, joysticks or keyboards. Many approaches have been developed using a large variety of sensors to capture human interaction information and then provide further processing and recognition of the acquired information. However, the majority of these approaches usually focus on the actual implementation of the various stages that comprise an advanced interaction environment. Thus, the need for defining common data formats for improving integration and reutilization of these solutions are typically not addressed. On the other hand, this study aims at surveying existing research on integrating devices into interactive environments, at different interoperability levels and in data formats, identifying techniques and patterns of conveying information from the real world to the virtual world, in order to synthesize results, organize applicable documents by similarities and identify future research needs.

Keywords: *Advanced Interaction, Natural User Interface, Data Interoperability, Data Exchange Format, Device Integration*

1. Introduction

Despite the evolution of computer system interfaces from command line based to sophisticated graphic systems, interaction between users and these systems haven't yet become natural. To improve this scenario, advanced interaction interfaces based on recognition of gestures, voice, eye-movement and/or touch have been proposed to allow human-computer interaction to become closer to human communication [1][2][4].

Advanced interaction environments have become a reality nowadays, mainly due to more affordable prices, easiness of use and popularization of gaming devices such as Kinect and Wiimote [5]. Kinect, which offers depth sensors that help tracking real time movements, has been frequently used in human-computer communication research, through gestures, robotics and graphic computing [6][7]. Another reason for its success is the availability of proprietary and open source libraries for application development. The Kinect SDK [8] is a development environment provided by Microsoft that includes drivers and libraries needed for gesture and voice recognition, using the Kinect sensor. The OpenNI framework [10] and NITE middleware [9] are two other known initiatives that allow the creation of sophisticated interfaces with a higher level of abstraction, based on the combination of several sensors provided by Kinect [10].

The development of interactive environments demands great time and effort in interacting with all of the software components and input devices (camera sensors, microphones, RFID sensors, etc.) [11]. Although building advanced interfaces

ISSN: 1975-0080 IJMUE

benefited from the use of sensors from the gaming domain, integrating these sensors with other user and environment capturing devices is still an unresolved issue. Integration of data coming from several devices within the same interactive system or application, in general, works in two directions: (i) improving performance and accuracy when obtaining information from sensor data processing and (ii) allowing for reuse and integration of solutions based on only one type of sensor, in an isolated manner.

An example of the first direction is in Caputo *et al.*[7], where the authors propose the fusion of multiple Kinect depth sensors with HD camera color sensors as a means of improving gesture recognition from greater distances. This approach combines arm and hand movement dynamics recognition, provided by Kinect, with static hand gesture recognition, from color sensors. The integration of sensor captured data (in this case RGB values) is done at application level and no information about format or structure of the data is provided. On the other hand, Carvalho *et al.* [6] seeks to standardize the description and handling of position data obtained from accelerometers, facilitating the integration of these types of sensors to an application. Integration of accelerometer data with other sensors, however, it is not addressed by the authors.

Focusing on reuse and integration, Rehem *et al.* [11] presents a framework named TTAir to facilitate the implementation of gesture based interfaces, starting from the standardization and reutilization of code and tasks commonly done by developers. The framework has a data access layer that provides a standard way for retrieving information from several sensors, creating an information database for future processing and inferences. The integration of new sensors into the environment depends on specific driver implementations, at the TTAir access layer. In a similar manner, the OpenNI provides interfaces to software components (called middlewares) and physical devices through an API. This API enables the registration of multiple components (modules) inside OpenNI which are responsible for producing or processing physical sensor data, in a more flexible manner.

For example, suppose a scenario where it is desired to build an interface that combines gesture recognition from (i) Kinect captured data (using the solution in [11]) and (ii) information from user hand movements, with accelerometer data captured from Wiimote (solution in [6]) in order to control, by gestures, remote robots or unmanned vehicles. In this scenario, creating an integrated solution would be greatly simplified if data associated to different (and complex) sensor data recognition processes was generated considering interoperable solutions.

The following situations illustrate other advanced interaction scenarios where interoperability should be considered:

- Playback of audiovisual content integrated with several sensing effects (wind, heat, vibration, smell, etc.). For universal access, players and actuators, from different manufacturers, should be able to communicate with each other using a common language.
- A board game between a virtual character and a real user, where the user's game pieces are moved through gestures. To increase the number of games and recognition sensors, actions from the real and virtual worlds would need to be synchronized according to a given pattern.
- A gesture based universal TV control. Porting applications to and from different manufacturers without requiring huge customization efforts would require gesture recognizers to reproduce the same information from a set of gestures, to actuators.

The ISO/IEC/IEEE 24765:2010 [12] presents the following three definitions for the term interoperability: 1) the ability of two or more systems or components to exchange information and use exchanged information; 2) the ability of two or more

ORBs (Object Request Broker) to cooperate in delivering requests to their appropriate objects; 3) the ability to communicate, execute programs and transfer data between several functional entities, requiring few or no knowledge about unique characteristics of those entities. ECIS states that interoperability is one of the main pillars of the TIC industry and any device that cannot interoperate with other dissimilar products has no value [13]. In addition, the ISO/IEC 25010:2011 [14] establishes interoperability as a determinant quality of software products, defining it as the capability of a software product to interact with one or more specific systems.

Hanseth *et al.* [15] believes that, in order to communicate, partners should use common patterns such as language or protocols, which is absolutely necessary for an information infrastructure to exist. Generally, patterns are described through standards which define common elements, such as user interfaces, system interfaces, data representation, data exchange protocols and data access interfaces or system functions [16]. Moving towards interoperability, standards are of great importance because they can be adopted by several manufacturers, thus increasing the probability of a set of systems from different sources to be able of interoperating with each other. MPEG-V [18] and MPEG-U [19] standards, for example, focus in interoperability between information that is shared, with aggregated semantics or not, between real and virtual worlds in advanced interaction environments.

In fact, data interoperability between systems can be implemented through [16]:

- Unifying the definition of all system data for all involved parts. This approach is problematic when applied to complex and large scale systems, as it would require great standardization efforts;
- Use of object orientation to specify data definitions, encapsulating internal details, such as in JSON (JavaScript Object Notation);
- Use of extensible data models and standard interfaces. The SNMP (Simple Network Management Protocol) is an example;
- Use of markup languages to describe data at several semantic levels.

Usually, data needs to be captured, transferred and processed by different components belonging to a distributed system. The use of common patterns and protocols to exchange information between all the involved parts (sensors, actuators and data processors in interactive environments) is vital for the correct operation of the system as a whole [17]. Hence, mechanisms should be proposed to transform data into structured meaningful information, allowing it to be processed by applications that form an interactive environment [2][4].

This paper investigates how research that involves implementation of advanced interaction environments deals with information interoperability issues. Thus, a systematic review was taken to identify frequent techniques and patterns used in: (i) transmitting information generated during the recognition process of an interactive environment state from sensor captured data and (ii) how information is processed by devices and applications belonging to the same environment. As the main contribution, this paper brings a detailed and critic analysis of previews work that involves data interoperability issues inside advanced interaction environments. For a second contribution, we propose the categorization of found work based on the adaption of the Lawson model [38], aimed at the control-command domain inside advanced interaction environments.

The remainder of this paper is organized as follows: Section 2 presents the methodology employed to perform a systematic review; Section 3 shows results from the aforementioned review; Section 4 shows results from the study, followed by Section 5, which presents our conclusions.

2. Applied Research Method

The applied research method is based on a broad and deep search for scientific papers related to advanced interaction environments and sensor data integration. The systematic review process presented here is based on references [20] and [21]. The process was divided in four stages: research protocol (planning), research execution, data extraction and analysis of the results.

Starting from the analysis, characterization, evaluation, comparison and synthesis of researched papers, this review offer, as its main result, a representative set of solutions approaching data and information integration within advanced interaction environments. In the next subsection the adopted method will be detailed.

2.1. Research Protocol

The planning phase produces, as its main result, a research protocol which defines criteria for selecting sources (digital libraries), query strings identification (with Boolean operators related to the investigated subject), as well as criteria for including and excluding papers under classification.

2.2. Search Selection Criteria

Indexed digital databases were used for researching, for being more specific, reliable and for presenting less redundant results than generic search engines (*e.g.* Google Scholar). Furthermore, databases would need to be related to the research area, allowing unrestricted access and supporting the use of query strings that were the result of a logical combination of metadata words including title, abstract and other keywords. In addition to the systematic review, a manual research was planned for the execution phase in order to minimize the possibility of missing relevant work.

Based on these criteria, the following digital libraries were selected: IEEEXplore, ACM Digital Library, Scopus, Engineering Village and Science Direct. Despite meeting applicable requirements, they all possessed different features for executing advanced queries. Thus, it became necessary to adapt our query strings to match the specific languages presented by each environment. For the manual research, Google Scholar was used.

2.3. Identifying Papers

The paper identification process is an important phase for collecting evidence about the research subject. Starting from the main objective, more specific questions were raised to determine which string combinations should be submitted for querying the digital databases. These questions were:

- What is the current state of the art regarding techniques for device integration in advanced interaction environments?
- Which multimodal interfaces present some sort of interoperability techniques?
- What techniques and/or patterns are used for describing sensor data that will be processed by other devices or systems?
- What is the ratio between recognized information and captured physical data of identified techniques and/or patterns?

Starting from these questions, several string combinations were evaluated until a logical expression, suitable to the purpose of the research, was found, in terms of scope and precision. Basically, a suitable query string should result in a reasonable number of papers and further fetching, within the top ten positions, should contain relevant contents to our review (previously known by the authors), specially, references to [2] and [17]. Keywords were related to natural, advanced or multimodal interaction interfaces which are recognized by sensors, systems or devices. All selected papers should contain at least title, abstract and keywords in English. Thus, the adopted expression was:

```
(natural or advanced or multimodal)
and
(interaction or "user interaction")
and
(interface or "user interface")
and
(sensor or sensed or device or system)
and
(recognize or recognition)
```

For the manual research, the same query string was used, occasionally suppressing some words, in an attempt to retrieve different papers from those found on indexed databases.

2.4. Inclusion and Exclusion Criteria

Criteria were defined according to the questions presented in section 2.3 and should be applied independently at the research execution stage (Section 3).

Inclusion criteria were:

- Uses natural, advanced or multimodal interaction interface;
- Describes sensors, devices or systems inside advanced interaction environments;
- Uses patterns to represent recognized physical information inside advanced interaction environments;
- Presents formatting of recognized data and/or information inside advance interaction environments.

Papers with the following characteristics were excluded:

- Paper contents not entirely available;
- Paper language other than English, Portuguese or Spanish;
- Limited to sensor capturing;
- Limited to processing of sensor signals;
- Limited to signal recognition inside physical environments and without human interface;
- Requiring specialized hardware (joystick, glove, helmet, etc.) to interact.

Absence of a layout for the presented data and information recognition inside advanced interaction environments.

3. Protocol Execution

Once the research protocol was defined, the collection of papers and primary data started. The execution phase was divided into three stages, each refining results from previous stages until a group of relevant papers was retrieved.

At stage 1, the query string presented in section 2.3 was submitted to each of the search engines from the selected sources separately, generating an initial result set containing 1813 papers. Bibliographic entries found while executing the query string on selected digital libraries are available at http://goo.gl/E3z3sq.

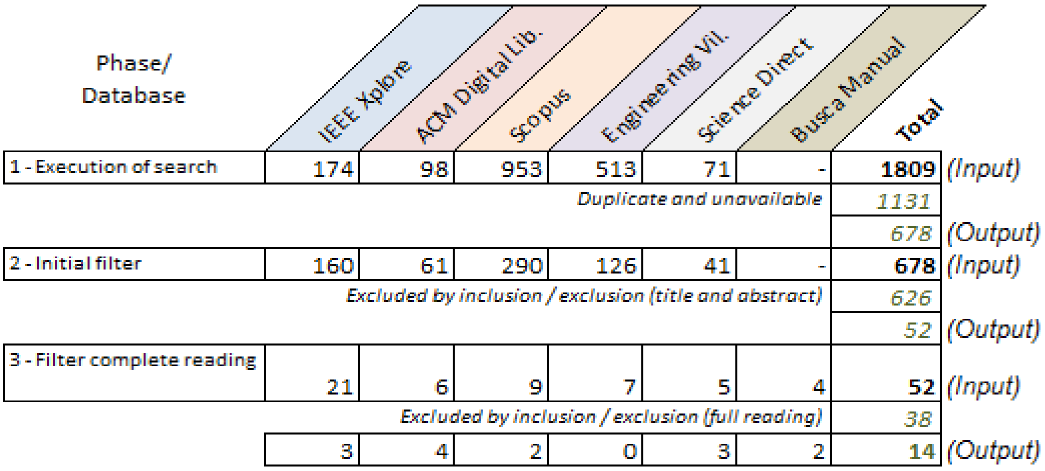
Stage 2 started by applying a filter that eliminated duplicated and unavailable papers, resulting in a new group comprised of 682 papers.

After stage 2, inclusion and exclusion criteria were applied to papers belonging to the 682 element group. Based on information from title, abstract and keywords, 52 papers remained after this filtering, where most of the exclusions happened because of the lack of information about formatting of captured data, therefore making it unfeasible to produce interoperable sensor data.

When it was impossible to define the inclusion or exclusion of a paper based on its title, abstract and keywords, at the second stage, the complete reading of the paper content was needed, in order to determine if it should be included in the final group of selected papers. At the end of stage 3, a manual search was conducted, using Google Scholar, for generating additional results. Only 12 papers from the starting group, plus 2 from the manual search, were considered relevant to the survey, as they expressed some concern regarding data interoperability inside advanced interaction environments. They are: [2], [5], [6], [17][22][23], [31].

It is important to notice that the manual search was done after the execution of all stages, to collect only papers that were relevant to the subject, and had not been retrieved by querying the indexed databases.

Figure 1 presents the number of selected papers after each of the filtering stages.



Phase/ Database	IEEE Xplore	ACM Digital Lib.	Scopus	Engineering Vil.	Science Direct	Busca Manual	Total	
1 - Execution of search	174	98	953	513	71	-	1809	(Input)
	Duplicate and unavailable						1131	
							678	(Output)
2 - Initial filter	160	61	290	126	41	-	678	(Input)
	Excluded by inclusion / exclusion (title and abstract)						626	
							52	(Output)
3 - Filter complete reading	21	6	9	7	5	4	52	(Input)
	Excluded by inclusion / exclusion (full reading)						38	
	3	4	2	0	3	2	14	(Output)

Figure 1. Number of Selected Papers at every Stage after Filtering

Figure 2 shows the distribution, by year, of the group of 682 papers, demonstrating a growing trend of research related to user advanced interaction since 2004.

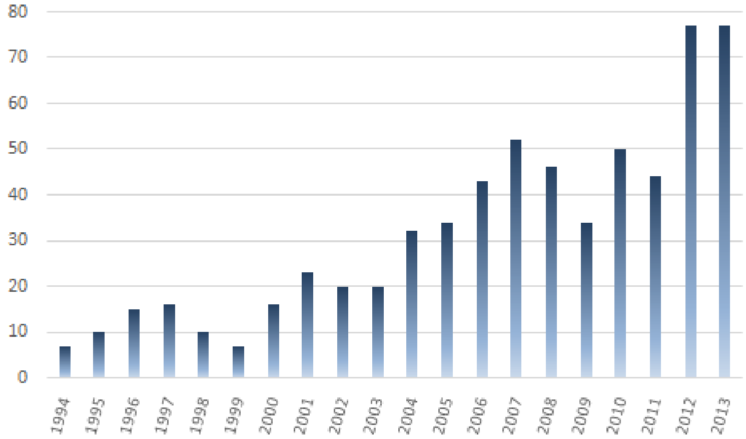


Figure 2. Evolution of Number of Publications Related to the Subject between 1994 and 2013

3.1. Data Extraction

Data extraction aims to categorize groups of papers obtained while subsequent filters were applied during the execution of the systematic review protocol. Groups were categorized according to the following aspects:

1. Main contribution (Modelling, sensorial capturing, processing, recognition, performance evaluation and others);
2. Interaction method used (Touch, hand gestures, body gestures, voice, emotions or thoughts, face, using of objects, traditional, look and others);
3. Types of sensors, devices and/or systems supported by each approach (RGB sensor, RGBD sensor, microphones, touch devices, accelerometer, keyboard or mouse, pen, psychological sensor, flexible sensors and others);
4. Proposed evaluation metric (Controlled experiment, formal proof, prototype, simulation, real environment, none and others);
5. Focus on interoperability or not; if so, at which architectural level?

Results from the evaluation of papers obtained in the search and execution stages are presented in the following section.

4. Results

One of the first evidences obtained from this systematic review was the inexistence of a generic architecture for advanced interaction environments. Such architecture should make explicit, for example, the components and interconnections utilized, as well as concerns related to user interaction, utilized sensors, middlewares, frameworks, APIs and actuators, providing feedback to user interaction.

Despite some architectural abstractions for , the majority of advanced interaction applications that have been proposed for gesture-based systems ([2],[32] e [33]), voice ([23], [34], [36] and [37]]), body movements ([5] and [34]) and touch ([37]), only emphasize specific modalities of interaction, showing little concern with the interoperability between the different layers and also do not addressing reuse. However, despite the specificities, the cited works possess similarities in their construction and operation, possessing capture phases, processing, recognition and action/performance. Based on this assertion, this article proposes an adaptation of the Lawson’s model, for generically describing interactive environments, [38] originated in the control area. The adapted model in the Figure 3 covers all the components of an advanced interaction system, comprising the stages of: (i) capture (sensing) and (ii) processing of the signal in the environment; comparison of the processed information with a foreseen and desired state (recognizing) and (iv) activation of devices due to recognition.

Two opportunities for interoperability between stages were identified from the adapted Lawson’s model in Figure 3. The first opportunity lies between the capture and processing of the data, and the second during the processing/recognition of a desired state, which then sends events to the actuation level. The interoperability, at the first level, would allow, for example, exchanging depth sensors from one manufacturer to another, with little impact in the application structure. At the second level, it would be possible to exchange implementations of voice, touch and gesture recognition, seamlessly as long as standardized representations and structures of recognizable actions (events) exist. These events would generate, for example, similar actions at the output interfaces of real or virtual environments. The idea of an event bus proposed by Rehem et al. [11] goes in this direction, despite not having standardization of event information, sent by recognizers, to the actuators in the real or virtual world.

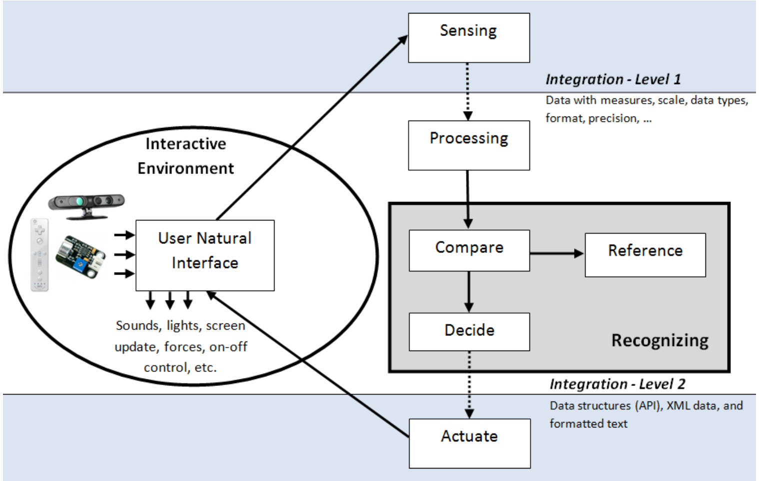

Sensing
Integration - Level 1
Data with measures, scale, data types, format, precision, ...
Processing
Interactive Environment
User Natural Interface
Sounds, lights, screen update, forces, on-off control, etc.
Compare
Reference
Decide
Recognizing
Integration - Level 2
Data structures (API), XML data, and formatted text
Actuate

Figure 3. An Adapted Lawson's Model for Interactive Environments

Following, a statistical analysis from this review will be presented. It will also be shown the characterization and critical analysis under the optic of the analogous model presented in this section.

4.1. Statistical Analysis

The five categories described in the section 3.1 were utilized to group papers retrieved during the search stage. The original 682 papers obtained in the initial phase of this stage after the filtering (removal of duplicates and unavailable), in addition to the 14 final papers obtained after the data extraction phase, were categorized according to the interaction method, in order to provide a general view of how these aspects are related to advanced interaction environments.

The papers were analyzed, looking for the types of interaction used, and then categorized in: touch, hand gesture, body gesture, voice, emotions or thoughts, face, use of objects (like pens), traditional (keyboard and mouse), look and others. Figure 4 (a) and (b) presents the categorization of interaction types found in the 682 and 14 papers, respectively.

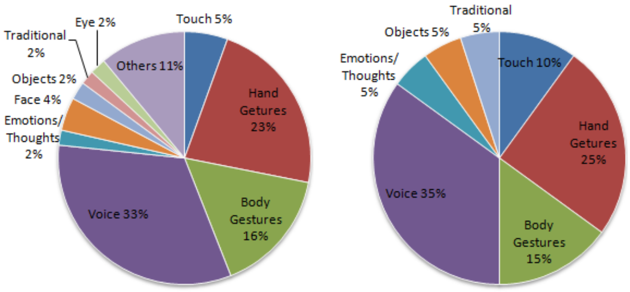


Figure 4. Interaction Methods in the (a) 682 Papers Obtained in the First Stage and (b) 14 Papers Obtained in the Third Stage

According to Figure 4 (a), the most used interaction methods are voice, hand gesture and body gesture, which, together correspond to more than 70% of papers that were analyzed. Similarly, these forms of interaction are also the most used in the 14 papers, selected after the third stage, as shown in Figure 4 (b). Notice that, despite the big difference in size of analyzed groups, the distribution of the applied method is very similar and there is a tendency to use voice and gesture as preferred interaction methods within advanced interaction environments. Another point to be highlighted is that, in some papers, the interaction methods are used as a group, forming a multimodal interface. The percentage of papers that involved multimodal interfaces was of 21% for those that were selected after the third stage and of 29% for those after the first filtering.

The modalities (types of interaction) were grouped, from 1994 until the year of 2013, in four periods of time, as shown in figure 5. It can be seen that since 2004 there has been a growth in the use of modalities based on (hand) gesture due to the popularity of sensors such as Kinect and Wii Remote.

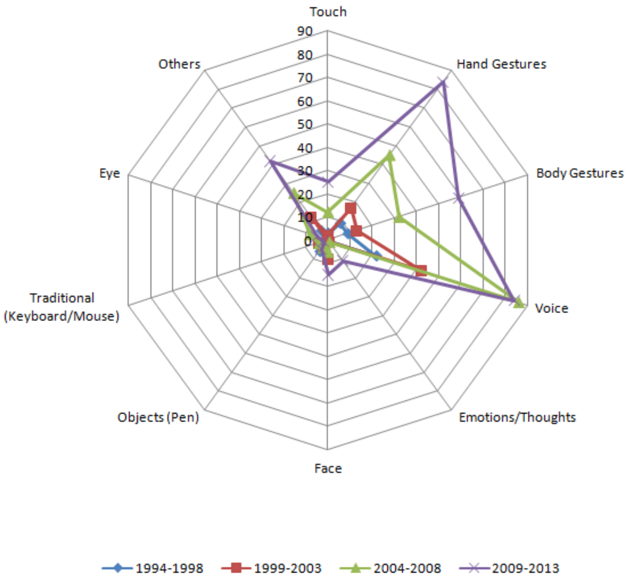


Figure 5. Classification of Interaction Methods, by Year

Directly related with the type of interaction, the information about types of sensors and devices utilized was also analyzed. They were categorized in: RGB sensors, RGBD sensors (Kinect, PrimeSense), microphones, touch (tablet, smartphone, and graphic tablet), accelerometer, traditional devices (keyboard and mouse), computational pen, psychological sensor, flexible sensors (bracelets, gloves) and others. Figure 6 (a) and (b) show the percentage for each type of sensor and device for the groups of 682 and 14 papers.

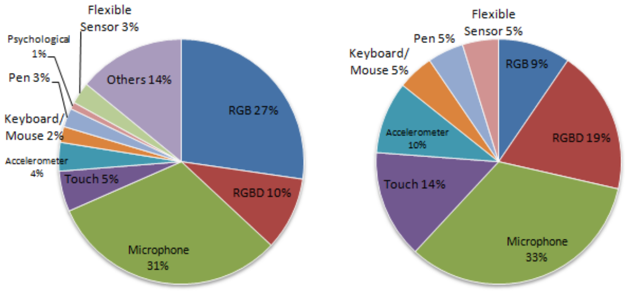


Figure 6. Types of Sensors and Devices in (a) 682 Papers Obtained in the First Stage and (b) 14 Papers Obtained in the Third Stage.

Figure 6 (a) shows that the most used types of sensors and devices in the group of 682 papers were microphones, used by voice based interfaces. Next, the RGB and RGBD sensors were the most used, mainly for gesture interaction. The RGBD sensors provide, additionally to the RGB image of a scene, a Depth map, used by many computational vision techniques. For the 12-paper group, most sensors used are again microphones, RGBD and RGB sensors, according to Figure 6 (b). It is worth mentioning that, as observed in interaction methods, the distribution between categories of sensors is very similar, even with the large difference in the amount of papers per group that were analyzed. Furthermore, it is clearly noticed an inversion on the use of RGB and RGBD sensors from Figure 6 (a) to (b), justified by the easiness of access to this type of sensor and the use of more recent approaches, that is more pronounced in the group of 14 papers, selected after search and filtering.

Regarding to other categorizations from section 3.1 (main contribution, evaluation metric and interoperability level), only the selected 14 papers which somehow approach the issue of data interoperability, will be used in this analysis.

Modeling (45%) was the main form of contribution from the papers that were found. Therefore, a model is presented to explain the characteristics or the operation of a system or application. The other significant forms of contribution were processing (27%) and recognition (18%). These areas are often used as synonyms; however, as shown in the model from Figure 3, the recognition involves comparison with given references and decisions, while processing involves the treatment of data from sensor devices. The remaining 9% of the papers could not be categorized by area of contribution.

Another information extracted from the 14 remaining papers, at the end of the third stage, was related to evaluation metrics, which were categorized as: controlled experiment, formal proof, prototype, simulation, real environment, no evaluation, and other evaluations.

The most used evaluation metric is prototype (55%), a product in testing or planning phase, which is a model that will be commercialized or distributed in the future. Next, controlled experiments (27%) were another evaluation metric, where tests are made in a controlled environment. Finally, the other identified metric is real environments (9%), in which the evaluation of the proposals are made in real environments. Additionally, 9% of research did not show any evaluation metric.

The most important information that was obtained from this review was about interoperability, regardless of the level in which it was done. Out of 682 articles analyzed at the end of the first stage, only 2% (14) showed some type of concern regarding data formatting or the utilization of some pattern for structuring and/or describing of dada, as shown in Figure 7 (a).

The remaining 14 papers were then analyzed specifically for interoperability issues. As shown in Figure 7 (b) most of those (58%) dealt with interoperability at some level, even without using a specific pattern. Lastly, the articles that presented some type of interoperability (58%) were classified accordingly to their level of interoperability, based on the adapted Lawson's model shown in Figure 3: 86% corresponding to interoperability at the second level, and 14% corresponding to interoperability at the first level, as depicted in Figure 7 (c).

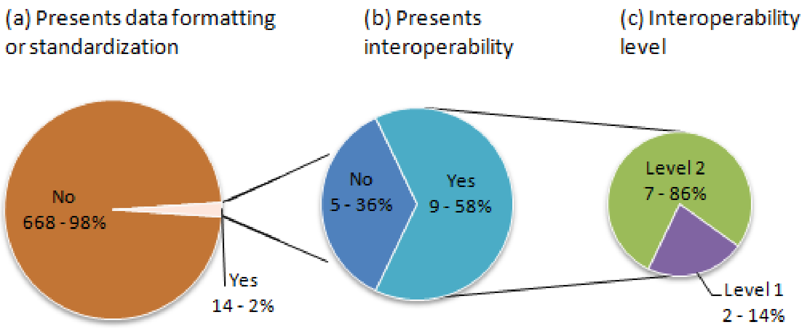


Figure 7. Papers that Present (a) only Formatting of Data; (b) and Treat Interoperability; (c) in which Level

4.2. Characterization and Critical Analysis

In all papers related to voice interaction, it was noticed that there were concerns about interoperability at the second level of integration, done after processing/recognition.

In Conti *et al.* [22] the voice information is received, processed by the SAPI (Speech API) engine and, with the support provided by a semantic repository, it is then normalized and delivered to an application. Its main idea is to recognize text with semantic support and activate functions on the application code. Plain text arrives at the application, which must follow a naming convention for function names that will be activated based on inferences. The SAPI can be considered an integration component in the architecture of advanced interaction applications, but has proprietary technology, restricted to the Microsoft Windows operating system, thus reducing the interoperability of the solution.

The Architecture shown in AT&T [23] proposes a centralized voice processing service (speech to text/text to speech) that overcomes limitations of mobile devices, communicating through the web. The provided services can be exposed under 3 formats – EMMA, JSON and XML – showing a concern with the use of patterns to promote interoperability. EMMA (extensible Multimodal Annotation) is a W3C standard language which provides a rich notation for the interoperability of data formats on multimodal systems. W3C SSML (Speech Synthesis Markup Language) was used specifically by TTS. This way, it promotes interoperability by using data formatting to exchange information between the layers of processing/recognition and action, making the application independent of the upper layer.

Marin et al. [26] also uses voice interaction, proposing an interface and architecture that allow users to command and program two educational robots remotely, through the internet. A microphone captures user-to-robot commands, which are made aware of their surroundings by using several cameras. One of the biggest difficulties of the online robots project was the network latency which ended up causing a great delay between sending commands and the feedback provided to users through cameras. Marin et al. does not show a standard format for voice commands.

Also in the voice interaction area, Sorce et al. [27] shows a multimodal system which provides context aware services. The service provides information to people according to their location in a particular environment, using a PDA equipped with an RFID identifier. Interaction is made through voice, using natural language. The system uses VoiceXML and the formatting of data is done at the second level, after the recognition of voice commands.

More recent researches involving interaction via hand gesture and touch has been showing concerns with the issue of achieving data interoperability at both levels, according to the adapted Lawson's model, that is, between capture and processing/recognition, and also between processing/recognition and the actuation level, as shown in Figure 3.

Han and Choi [2] propose a method for the recognition of hand position using a depth camera and the MPEG-U/Part 2 standard, at the second level of integration. The proposed standard specifies a group of standard data formats for advanced interaction interfaces that does not depend on frameworks or operating systems. The authors developed a proof of concept using the framework Candescent NUI for the recognition part, transforming the acquired information to the MPEG-U format. They concluded that if the development of advanced interaction technologies employed the MPEG-U standard, many applications could benefit from it. In fact, applications at the lowest layer of the architecture (after recognition), would not have to worry with the upper layer, in case it would receive expressions containing interaction semantics (ex. hand movement representing a circle), and it could then focus its efforts exclusively in responding to gestures, presenting a considerable time gain in the project.

In the architecture presented by Carvalho *et al.* [6], the authors propose a model for structuring captured data from accelerometers, making data manipulation via applications easier. The architecture establishes three conceptual layers with the following functions: 1) accelerometer data acquisition; 2) categorization of data acquired by gesture recognition; 3) modeling and adding meaning to the obtained data according to a structured model. The XML format was chosen to provide interoperability of accelerometer data. This is done at the third layer, which is directly connected to the application, after the processing/recognition phase. On the other hand, providing interoperability is only useful to accelerometer based devices, not being a solution that is comprehensive enough to be used in the integration of data from other sensors within interactive environments.

Carrino and Pédat [28] show a multimodal approach for interacting in smart environments. The interaction can be made through modalities like voice and/or gesture. An architecture based on the following layers was proposed by the authors: 1) A sensor layer manages all the technical aspects related to the management of sensor devices; 2) a processing layer, where gesture and voice modalities are managed and integrated, and 3) a feedback layer provides output data to the user. To provide interoperability, the architecture uses a protocol for the exchange of messages based on XML, between the processing layer and the feedback layer that is, at level 2 of the model depicted in Figure 3.

Another study that uses gesture interaction is the Ionescu *et al.* [29]. The authors present a system for operating remote TV control and set-top boxes through hand gesture, captured from a 3D camera. An XML file is used to describe gestures and allows users to create a gesture language. The paper is not concerned about interoperating with other modal or sensorial interfaces, but presents a data format after recognition, which allows integration with other systems.

The SSI (Social Signal Interpretation) [24] is a framework for real time social signals recognition, which supports a variety of sensing devices. The framework complements existing tools, offering special support to the development of online recognition systems using multiple sensors. Therefore, different strategies for combining information of different data modalities and decision levels are available, as well as the acquisition of a generic graphic interface for acquiring data and formatting models. New components can be added using a C++ API or using an XML interface which allows writing and editing applications using a simple text editor. Interoperability is done at the second level, that is, after the recognition of different modalities. It is worth mentioning, however, that the SSI framework is supported only by the Windows platform.

Schröder *et al.* [25] reviews the standardization process of EmotionML (Emotion Markup Language) presenting syntax elements. It was designed to be used in different technological contexts, intending to reflect concepts of affective sciences. The authors report that to achieve interoperability, a standard format of representation should be used. The standard was normalized by the W3C (World Wide Web Consortium). The syntax is described in XML and linked to a vocabulary of emotions; however, it does not use ontologies in its first version. Another challenge that is mentioned in the conclusion refers to the difficulty of expressing scales in EmotinML, because there is still no consensus in the community about the best practices, so a more detailed definition can be created.

Repo and Riekki [31] propose a middleware for multimodal context aware interfaces. To provide multimodal interfaces, third party user interface components are needed. The design and implementation of user interface (UI) applications can use three different techniques: abstract UI (XML), UI plugin (mobile Java code) and

Web based UI (HTML). The abstract UI does not depend on any programing language, type of device, operating system or toolkit. This is accomplished through separation of UI implementation from the actual application, through a XML based scripting language, which is used to describe abstract elements of the UI and their priorities. The formatting of data on XML is done after the recognition process. Being concerned with interoperability at the first level, Han et al. [17] makes use of the MPEG-V formatting pattern to unify interactions between the real and virtual worlds. After the capture of facial expressions, body gestures and hand gestures, the resulting data is formatted with MPEG-V and used to control characters in the virtual world. Authors report that a standardization of the interface is important and needed so as to provide interoperability among different virtual contexts (games, simulations, DVD and sensors). The representation provides a means to structure data in terms of position, orientation, speed and acceleration, and also allows values unity. There is no concern with the represented semantics in the exchanged data, which will be determined only in the processing/recognition phase.

The software middleware presented in Suma *et al.* [5] targets interactions of the entire body within virtual environments, so as to control third party applications. It provides a gesture creation interface, sensor configuration, recognition actions and other mechanisms. It also defines syntax for representing human gestures using sets of rules that correspond to basic spatial and temporal components for a given action. Primitive actions are represented in plain English text, simple and understandable even to lay users, in order to stimulate development of accessible interfaces with fewer efforts. However, interoperability is restricted to the middleware's operating environment.

Shen *et al.* [30] describes a tablet (Force Tablet) designed for human-computer interaction with a goal of acquiring the kinetic (strength) and kinematics (trajectory) of human writing through a stylus pen. Authors report that patterns have been used for digital ink since the nineties, particularly by using UNIPEN and Jot devices for digital writing. In research, the patterns are not directly used because their main concern is about tracking detailed information about calligraphy. However, in interactions between pen/tablet and computer, InkXML is widely used, which is a data format used to represent digital ink from electric pens or stylus and that can be used in writing recognition, signature verification and gesture interpretation. Finally, it is presented a simple application based on a pen interface in order to evaluate handwriting ability. Table 1 presents a general view of the 14 critically analyzed articles that contained some sort of data formatting. For each article is presented the (i) year of publication, (ii) the user interaction modality, (iii) the data interoperability approach, accordingly with session 1, and (iv) the interoperability level according to the Lawson's model adaptation described in Section 4.

Table 1. General View of the Final 14 Selected Papers

Paper	Year	Modality	Data Format	Lawson's Level
[2]	2014	Hand Gesture	XML	Level 2
[5]	2013	Voice	Single data definition	Level 2
[6]	2012	Hand Gesture	XML	-
[17]	2013	Body gesture	XML	Level 1 e 2
[22]	2006	Voice	Single data definition | Object Orientation	-
[23]	2009	Voice	XML | Object Orientation	Level 2
[24]	2013	Hand Gesture | Body gesture | Touch | Voice	XML	Level 2
[25]	2011	Mood / emotions	XML	Level 2
[26]	2005	Voice	Single data definition |	-

			Object Orientation	
[27]	2007	Voice	XML	-
[28]	2011	Hand Gesture | Body gesture	XML	Level 2
[29]	2011	Hand Gesture	XML	-
[30]	2005	Objects (pen)	XML	Level 1
[31]	2004	Touch | Keyboard/Mouse | Voice	XML	Level 2

5. Conclusion

If, during the nineties, human-computer interaction was considered to be at its childhood stage [32], it has evolved substantially in the last years. Restrictions such as recognizing more than one user interacting with their hands in a scene, while processing in real time, were solved in an effective manner, and the use of natural interfaces has been increasingly present among us through the popularization of gesture, movement, touch and voice sensing techniques.

A big part of research in the advanced interactions field has been dedicated to the process of capturing and recognizing signals and data generated by several types of sensors. On the other hand, the integration of most solutions, or part of them, becomes impractical if exchanged information and data is not structured and described in a standardized and open manner. Being concerned about the interoperability of data exchanged between devices and systems, this paper systematically searched and examined a series of publications related to advanced interactions, looking for techniques and patterns used to promote the communication between different devices and applications of such nature. Despite uptrend in researching user advanced interactions, the number of approaches tackling interoperability between devices and systems using data formats inside advanced interaction environments is still small. Of the 682 originally selected papers, using the proposed query strings described in section 2.3, only 14 of them presented some sort of concern regarding data formalization and description, representing only 2% of the collected papers.

It was perceived the utilization of several types of formats, such as XML, JSON and plain text for data. In the field of voice interaction it was noted the utilization of widespread standards from W3C, such as VoiceXML and SSML. Under visual (gestures and poses) and tactile (touch) modalities, it was highlighted the recent emergence of MPEG-V and MPEG-U standards, focusing on the transmission of real world information to the virtual world, with the possibility of associating semantics to the underlying information, all in an interoperable manner.

Papers found during the research were analyzed and synthesized based on the proposed adaptation of the Lawson model. According to the new model, the integration through data formatting could be done at two different levels: level one, which is applied between the capture and recognition of data, and level two, which is applied between the recognition and the event/action. Interoperability at the first level allows a wide range of different sensors (or similar sensors from different manufacturers) to be used for the same purpose, also allowing sensors interchangeability to happen with the less possible impact. Within the analyzed papers, the solution that provides interoperability at this level is the MPEG-V standard, connecting the virtual world (*e.g.* videogames, simulators and applications) to the real world (*e.g.* sensors and actuators). If one sensor is replaced, at the sensing layer of the model presented in figure 3, by another one that implements the same pattern, then lower layers will not suffer any interference. It was possible to perceive that the majority of papers that presented interoperability were more concerned about the second level, where usually actuators are located, after processing/recognition. At this level, actuators can work independently from sensors and recognizers, if interoperability is provided. This considerably reduces project time by reusing upper layers, as developers would need to focus on actuation only. Within the solutions found, the part 2 of the MPEG-U standard was predominant, which specifies a set of data formats with geometrical, symbolical, gesture and hand position patterns, touch patterns and composition patterns, that is independent of frameworks or operating systems, creating a semantic sense after the recognition process. The MPEG-V standard also allows the interconnection of actuators to exchange information about device capabilities, adaptation of devices according to user preferences and device control (e.g. adding a light or controlling the temperature of an air conditioner). While defining the standardization of advanced interfaces for interaction between systems and devices, MPEG evidences its concern with language universality between different data sources and consumers. Because of their involvement in different levels, if necessary, these patterns can be combined into the architecture of the same solution.

Interoperability through patterns allows for the interconnection of systems and devices developed by different manufacturers, further allowing them to freely choose each component within the preferences of each user. Despite the concern with data formats for interoperability, XML or JSON by themselves, for example, are not panaceas for all data integration issues, inside advanced interaction environments. In parallel to data format, other factors should be considered, such as solution complexity, response time and implementation cost. An architectural perspective could allow better organizing the complexity of integrating systems and devices, leading to more cohesive solutions.

It should be noted that this research focused in the issues of achieving interoperability through data formatting, not ruling out the possibility of using middlewares and frameworks to provide interoperability through data messages exchanging, in advanced interaction environments.

Middlewares can also be seen as an interoperability alternative when it is desired to minimize the complexity and heterogeneity of several systems. On the other hand, different requirements and architectures from middleware proposals, even when in the same domain, cannot solve the interoperability problem definitely. For example, it was required the creation of an integration layer, denominated GEM, to allow interoperability between applications produced according to middlewares such as MHP, ATSC, GINGA, etc. inside the Digital TV domain. Several middleware solutions were also proposed for gesture recognition (Candance, TTAir, etc.). The same happens for API frameworks.

For future work, experiments about multimedia integration through patterns found in this survey are suggested. A gateway could link data from different sources and forward it to other devices, allowing easier adaptations in diversified advanced interaction environments.

Finally, it should be highlighted that the limitations presented in this paper are similar to those present in other systematic reviews. Even though an additional manual research was carried out, it is possible that important material, such as dissertations, related books, white papers and other relevant documents, had not be found inside digital databases while using the designed search and selection protocol.

References

[1] G. R. S. Murthy and R. S. Jadon, “A Review of Vision Based Hand Gestures Recognition”, International Journal of Information Technology and Knowledge Management, vol. 2, no. 2**, (2009)** July-December.

[2] G. Han and H. Choi, “MPEG-U based advanced user interaction interface system using hand posture recognition”, Proceedings of 16th International Conference on Advanced Communication Technology, **(2014)** October 16-19, Pyeongchang, Korea.

[3] D. Avola, A. Bueno, G. Gianforme, S. Paolozzi and R. Wang, “SketchML a Representation Language for Novel Sketch Recognition Approch”, Proceedings of 2nd International Conference on Pervasive Technologies, **(2009)** June 9-13, Corfu, Greece.

[4] N. Sebe, “Multimodal Interfaces: Challenges and Perspectives”, Ambient Intelligence and Smart Environments, vol. 1, no. 1, **(2009)**.

[5] E. A. Suma, D. M. Krum, B. Lange, S. Koenig, A. Rizzo and M. Bolas, “Adapting user interfaces for gestural interaction with the flexible action and articulated skeleton toolkit”, Computers and Graphics, vol. 37, no. 3**, (2013)** May.

[6] J. Carvalho, M. Manzato and R. Goularte, “Accelerometers data interoperability: easing interactive applications development”, Proceedings of 18th Brazilian Symposium on Multimedia and the Web, WebMedia, **(2012)** October 15-18, São Paulo, Brazil.

[7] M. Caputo, K. Denker, B. Dums and G. Umlauf, “3D hand gesture recognition based on sensor fusion of commodity hardware”, Mensch and Computer, **(2012)**.

[8] Kinect for Windows, Available in http://www.microsoft.com/en-us/kinectforwindows/. Accessed 13 Oct 2014

[9] PrimeSense NITE, Prime Sensor NITE 1.3 Framework Programmer’s Guide, **(2010)**

[10] A. N. Rehem, C. A. S. Santos and M. V. R. Andrade, “Interfaces para Aplicações de Interação Natural Baseadas na API OpenNI e na Plataforma Kinect”, Proceedings of the XVII Webmedia and XXVI SBBD. **(2011),** October 3-6, Santa Catarina, Brazil.

[11] A. N. Rehem, C. A. S. Santos and L. A. Carvalho, “Touch the air: An event-driven framework for interactive environments”, Proceedings of 19th Brazilian Symposium on Multimedia and the Web, **(2013)** November 5-8, Salvador, Brazil.

[12] ISO/IEC/IEEE 24765:2010(E). Systems and software engineering - Vocabulary.

[13] ECIS - European Committee for Interoperable Systems. Interoperability. Available in http://www.ecis.eu/ecis-interoperability Accessed 07 Sep 2014

[14] ISO/IEC 25010:2011. Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models

[15] O. Hanseth, E. Monteiro and M. Hatling, “Developing Information Infrastructure: The Tension between Standardisation and Flexibility”, Science, Technology and Human Values, vol. 21, no. 4**, (1996)** October.

[16] M. Kasunic and W. Anderson, “Measuring Systems Interoperability: Challenges and Opportunities”, Technical Note CMU/SEI-2004-TN-003, Software Engineering Institute, **(2004)** April, Carnegie Mellon University, Pennsylvania, USA.

[17] S. Han, J. Han, J. D. K. Kim and K. Changyeong, “Connecting users to virtual worlds within MPEG-V standardization”, Signal Processing: Image Communication, vol. 28, no. 2, **(2013)**.

[18] ISO/IEC 23005-5:2013. Information technology - Media context and control - Part 5: Data formats for interaction devices.

[19] ISO/IEC 23007-2:2012. Information technology - Rich media user interfaces - Part 2: Advanced user interaction (AUI) interfaces.

[20] R. M. C. Segundo and C. A. S. Santos, “Systematic Review of Multiple Contents Synchronization in interactive Television Scenario”, ISRN Communications and Networking, **(2014).**

[21] King’s College London, Systematic Reviews User Guide, **(2014)** May. Available in http://www.kcl.ac.uk/library/help/documents/Systematic-Review-User-Guide.pdf

[22] G. Conti, G. Ucelli, and R. De Amicis, “Verba Volant Scripta Manent” a false axiom within virtual environments. A semi-automatic tool for retrieval of semantics understanding for speech-enabled VR applications. Computers and Graphics, vol. 30, no. 4, **(2006)**.

[23] G. Di Fabbrizio, T. Okken and J. G. Wilpon, “A speech mashup framework for multimodal mobile services”, International Conference on Multimodal Interfaces, **(2009)**. November 2-4, Massachusetts, USA.

[24] J. Wagner, F. Lingenfelser and E. André, “The social signal interpretation (SSI) framework”, Proceedings of the 21th ACM International Conference on Multimedia, **(2013)** October 21-25, Barcelona, Spain.

[25] M. Schröder, P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter and E. Zovato, “EmotionML– an upcoming standard for representing emotions and related states”, Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction, **(2011)** October 9-12, Tennessee, USA.

[26] R. Marin, P. J. Sanz, P. Nebot and R. Wirz, “A Multimodal Interface to Control a Robot Arm via the Web: A Case Study on Remote Programming”, IEEE Transactions on Industrial Electronics, vol. 52, no. 6**, (2005)** December.

[27] S. Sorce, A. Augello, A. Santagelo, G. Pilato, A. Gentile, A. Genco and S. Gaglio, “A multimodal guide for the augmented campus”, Proceedings of the 35th Annual ACM SIGUCCS Conference on User Services, **(2007)** October 7-10, Florida, USA.

[28] S. Carrino and A. Péclat, “Humans and Smart Environments: a novel multimodal interaction approach”, Proceedings of the 13th International Conference on Multimodal Interfaces, **(2011)** November 14-18, Alicante, Spain.

[29] D. Ionescu, B. Ionescu, C. Gadea and S. Islam, “An Intelligent gesture Interface for Controlling TV Sets and Set-Top Boxes”, Proceedings of the IEEE International Symposium on Applied Computational Intelligence and Informatics, **(2011)** May 19-21, Timisoara, Romania.

[30] F. Shen, L. Kang, B. Fang and Z. Wu, “The Pen-Computer Interface Based on Force Tablet: Active Information Acquisition for Human”, Proceedings of the IEEE International Conference on Information Acquisition, **(2005)**, June 27 - July 3, Hong Kong and Macau, China.

[31] P. Repo and J. Riekki, "Middleware support for implementing context-aware multimodal user interfaces", Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia, **(2004)** October 27-29, Maryland, USA.

[32] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: a review", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7**, (1997)** July.

[33] R. O'Hagan and A. Zelinsky, "Visual gesture interfaces for virtual environments", Proceedings of the 1st Australasian User Interface Conference, AUIC, **(2000)** 31 January - 3 February, Canberra, Australia.

[34] C. Duque, F. De la Rosa, and J. T. Hernandez, "Multimodal interaction architecture applied to navigation in maps", Proceedings of the 8th Computing Colombian Conference, **(2013)** August 21-23, Armenia, Colombia.

[35] S. Sultana, M. A. H. Akhand, P. K. Das and M. M. H. Rahman, "Bangla Speech-to-Text conversion using SAPI", Proceedings of the International Conference on Computer and Communication Engineering, **(2012)** July 3-5, Kuala Lumpur, Malaysia.

[36] M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman and I. Smith, "Unification-based multimodal integration", Proceedings of the 35th Annual Meeting on Association for Computational Linguistics, **(1997)** July 7-12, Madrid, Spain.

[37] M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction", Proceedings of the IEEE, vol. 91, no. 9**, (2003)** September.

[38] H. C. Builder, C.S. Bankes and R. Nordin, "Command Concepts: A Theory Derived from the Practice of Command and Control", Rand, Washington **(1999)**.

Authors

Celso A. S. Santos, he is associated professor at Federal University of Espírito Santo (UFES), currently with the Department of Informatics. He holds a Doctorat in Informatics (1999) from the University Paul Sabaiter, Toulouse France. His PhD Thesis was carried out at the Laboratory of Analysis and Architecture of Systems (LAAS) of National Scientific Research Center (CNRS). He also received his MSc. degree from University of São Paulo (1994) and his Bachelor's Degree in Electrical Engineering, from University Federal of Espírito Santo (1991). His research interests include multimedia application design, multimedia synchronization, digital video, human computer interface evaluation.

Estêvão B. Saleme, he is an Information Technology Analyst at Federal Institute of Espírito Santo. He received his Graduation's Degree in Information Systems at FAESA (2008), and a post-graduate degree in Software Engineering at Federal University of Lavras (2010). Currently pursuing a Masters in Computer Science at Federal University of Espirito Santo. His research interests include multimedia systems and web applications.

Juliana C. S. de Andrade, she is networks technologist at the Federal Institute of Espírito Santo (IFES) and professional education teacher of Information Technology in the National Industrial Apprenticeship Service (SENAI). She received her Graduation's Degree in Technology Computer Networking from Federal Institute of Espírito Santo (2011), and is currently pursuing a Master's Degree in Computer Science at the Federal University of Espírito Santo. Her research interests include ubiquitous computing, focused on data interoperability and integration facilities in natural interaction user interfaces.