Analyzing Face and Speech Recognition to Create Automatic Information for Usability Evaluation

Thiago Adriano Coleti, Marcelo Morandini, and Fátima de Lourdes dos Santos Nunes

University of Sao Paulo, Brazil
{thiagocoleti,m.morandini,fatima.nunes}@usp.br

Abstract. Observe users perform their tasks in a software is an important way to performing usability evaluation due to the reason that provides real data about the interaction between user and system. Filming and verbalization are very used techniques and they must be a concern for all designers. However, the needs of reviewing all registered data manually became these techniques slow and difficult. This paper presents an approach that uses face recognition and speech processing to generate relevant information about a system under test such as what moments the user had specific reaction and which ones. The ErgoSV software supported the experiments that were performed using an e-commerce website. The results showed that this approach allows the evaluator identify interfaces with usability problems easily and quickly as well as present information using percentages that supported the evaluator making decision.

Keywords: Usability Evaluation, Human-Computer Interaction, Speech Recognition, Face Recognition.

1 Introduction

Evaluating software usability is one of the most important activities of the design development process and performing it with real users should be a concern to all developers. In some sense, this task should be irreplaceable since it provides real information about the interaction between user and software and how one interferes in other to the evaluator [2].

This strategy to evaluate Human-Computer Interaction (HCI) usability is also known as usability test, and usually is performed by observing the user performing their tasks in a prototype or in a full software release. Two techniques are widely used to supporting the test: filming and verbalization. Filming consists in the positioning of one or several cameras near the user in order to collect images of face, keyboard, mouse, environment and other locations that can be considered important by the evaluator. In verbalization tests, the participant is encouraged to verbalize (pronounce) what he is thinking about the system and the evaluator can collect this data writing or registering them in audio files. The participant can verbalize during the evaluation (simultaneously verbalization) or verbalize after the test (consecutive verbalization). These techniques are widely used by researchers and developers and both of them

M. Kurosu (Ed.): Human-Computer Interaction, Part I, HCII 2013, LNCS 8004, pp. 184–192, 2013.

may present either qualitative or quantitative results for analyzing the recorded interaction. However, they are considered slow and expensive techniques due to the reason that evaluators and designers should review all the images and voice data as a film or a music to identifying whether happens some usability problem. According to Nielsen [2,3] this task can take two to three times the time of evaluation.

This paper presents the development, implementation of a usability evaluation approach based on observation method, filming and verbalization techniques and supported by face recognition and speech recognition. In this approach software collects face images and words pronounced by participants. Then, it processes these data and indicates what time specific user's face reaction occurs or when they pronounced a word. Thus, these data can be used to produce other relevant information about the interaction such as level of confidence and satisfaction on the results presented and efficiency/efficacy of the interactions performed.

2 Bibliographic Review

This section presents the bibliographic review performed in order to collect data about the subjects dealt with this research. Three issues are discussed: Usability evaluation supported by face and speech recognition; Image Processing/Face Recognition; and Speech Processing.

2.1 Usability Evaluation Supported by Face Recognition and Speech Recognition

The usability evaluation is a systematic process aimed to collect data in order to produce qualitative and/or quantitative information about the interface, users and interaction process, allowing the evaluator to provide corrections or establish a interface pattern [2,4]. Two methods are used by the designer to perform usability evaluations: (1) Usability inspection, where an interface is compared with guidelines, such as Ergonomic Criteria [1] or Heuristics [2,3]; (2) Usability Test, where real users are encouraged to use a software prototype or a full release and submit it to real situations in order to analyze whether the interaction between user and software has problems [1]. Filming the interaction process using one or several cameras or request to user for verbalizing what they are thinking about the software are two widely used techniques. The first technique aims to register images about the interaction between user and software. The evaluator places one or several cameras in strategies position in order to collect images from user, software, computer and environment. The images are used as data and analyzed by evaluator in order to identify interfaces with usability problems. The analysis is performed manually and consists in watch all the video since the first recorded second until finishing the test. The second technique is known as verbalization and consist in encourage the user (participant) verbalize what they are thinking about the software and consequently, the evaluator registers it in note or audio files [1,2,4]. The analysis of audio files or notes is performed in the same way of

filming technique [1]. Due to this reason these techniques are considered slow and this difficulty leads some evaluators ignore this stage of usability test causing interaction problems.

2.2 Face Recognition/Image Processing

An digital image is the representation of a physic object that can be recorded, processing and interpreted according to user's needs. Image processing is composed by a set of techniques aiming at manipulating images using computational algorithms in order to extract information from them [6].

The image processing is an activity usually used in several areas such as medicine, geography, physical and human-computer interaction. In medicine area the image process has being highlighted in several activities such as X-Ray and ultrasound as a resource to supporting the medical decision-making. Beyond medicine, the image processing is used in other studies and task, such as entertainment, design, security and aviation and involves a broad class of software, hardware and theory [5].

The face detection/face recognition is one of the most important and known image processing activities. This technique use algorithms to identify were a face is located in a image where a human being is represented. [7].

The image processing and more specifically the face recognition could be an important resource in order to support usability evaluation to collect and processing user´s face images during the evaluation and processing it to generating information about test.

2.3 Speech Processing

The human being has several mechanisms to express their emotions and one of the more important ways is the voice. Due to the importance in human life, the voice became an important area of research in computing [12]. Speech processing is the process of voice interpretation by computer, receiving an external signal and through computational algorithms performing the transformation in an output like a text [9,10]. The approach of converting voice signal in a text is also defined by authors as Automatic Speech Recognition (ASR).

There are several methods and techniques to perform the speech recognition such as Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP) and Mel-Frequency Cepstral Coefficient (MFCC) and Hiden Markov Models (HMMs). The main difference among them is the number of processes performed to transforming the voice signal in text, but the basic activities are the same: (1) collect sounds using a resource such as microphone; (2) processing the signal and generating the text; (3) display the final result [9, 10].

The use of speech processing in different areas such as software development and biometrics raised the needs of tools to supporting the recognition activities easily and quickly in such waythat developers do not need to know models. Aiming solving this gap, the Laboratório de Processamento de Sinais (LAPS) in Federal University of Para – Brazil has developed the Coruja Application [11]. This application allows the

use of speech processing functions easily and quickly in development environments such as Visual C# and can recognize both English and Brazilian Portuguese language. Researches using this tool [1,12] concluded that the Coruja Application recognizes between sixty and ninety percent of tested words. Tests were also performed before starting the ErgoSV development that also had the recognition rate greater then seventy percent. Due to this reason the Coruja Application was chosen to supportthe ErgoSV development.

3 Usability Evaluation Supported by Face and Voice Recognition

The usability evaluation framework supported by face and voice recognition was developed in order to support observation method. The main novelty of this approach is the use of face recognition, image processing and speech recognition as a resource to collect and process data, generating relevant information such as confidence and satisfaction on the results presented, efficiency/efficacy of the interactions performed, the interfaces and moments when the user had specific reactions. In this way, evaluator does not need review all data storage in order to obtain these data.

Aiming analyzing the effectiveness of the approach, experiments were performed using two specifics software developed for this research, called ErgoSV Software and ErgoSV Analyzer [8]. These applications aimed to collect data about user such as face image and words pronounced beyond collecting screenshot images, processing the data and generating relevant information. Figure 1 presents the ErgoSV Software Approach.

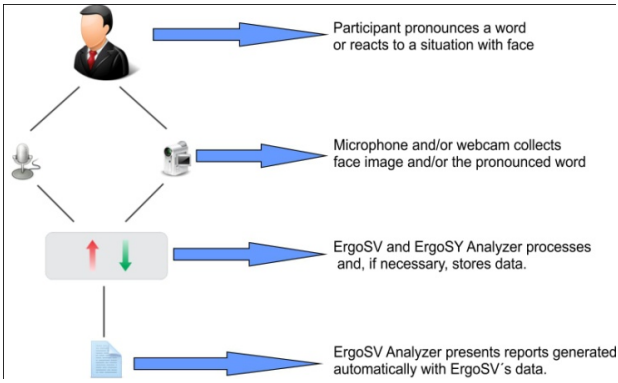

Fig. 1. ErgoSV ErgoSV Approach

3.1 ErgoSV Software

The first application was called ErgoSV and aimed to collect evaluation data. This software was installed in a computer and used to performing observation in a website usability test. In order to develop this software two resources were used to support

face and speech recognition: (1) OpenCV Library: a free computational library that has several image processing functions and is easily integrated with development environment such as Visual C#. This library allows the easy access to face recognition function, avoiding the development of recognition algorithms [7]; (2) Coruja Library:a free library that allows developers to use speech recognition functions easily and also allows the integration with development environments. This tool is able to recognize any pronounced word and/or specific words configured in a word files (specific dictionary) [11].

Therefore, to perform the approach experiments, we choose the user face as a data to be collected and configured the ErgoSV and the OpenCV Library in order to collect only the face image and register it. Regarding to speech data collection, we choose four words that represent quality concepts: Excellent, Good, Reasonable and Bad.

Besides these settings, before start any test, the evaluators configured other parameters such as application name, approach (Only Filming, Verbalization, Both), Images Interval (for screen and face collecting) and Words. A face image was requested for all participants in order to collect a default picture to be compared to others face images collected during the test.

We used an e-commerce website and we established a series of activities to be performed by participants. The activities were related with the buying process such as Searching for a product, Visualizing Products, Buying Process and Informing payment details. Figure 2 presents the ErgoSV interface used to performing the tests and collecting data.

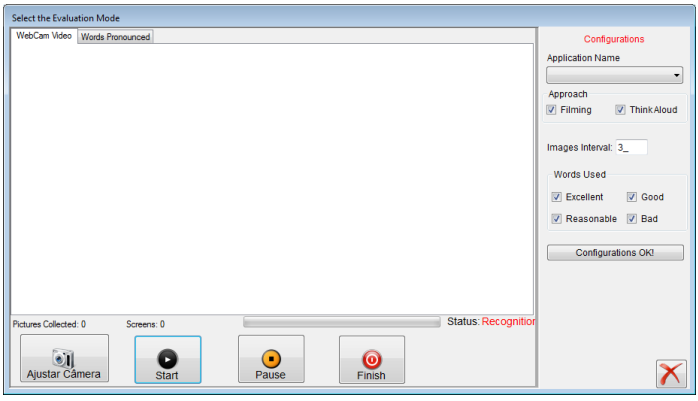

Fig. 2. ErgoSV Interface

A series of experiments was performed by four participants using ErgoSV who executed several tasks provided by evaluators. Each test took about fifteen minutes and in all tests the ErgoSV collected one face image and one screen image per second and all the pronounced word. The test data, processing and results are present in next subsection.

3.2 ErgoSV Analyzer

The ErgoSV Analyzer is a software developed in order to make the processing of collected data and displaying relevant information about software tested such as the words pronounced, the face and screenshot images collected, the words confidence and the software usability rate.

This application is also used to analyze whether the face image, words pronounced and screenshot images are adequate to generating usability information, mainly considering which moments the user had specific reactions.

The information exhibition was divided in three parts: the first interface is related with Words Pronounced and displays which words were pronounced, the confidence rate and the time (minutes after starting test) the specific reaction has happened. Also, this interface presents a chart containing the percentage of each word pronounced during the test. When the evaluator selects a pronounced word, he/she can access some images of screenshot that were used by participant when they had that reaction. The quantity of images displayed is configured according to user needs who must inform how many seconds before and after the word is pronounced he/she hope see the screens. Figure 3 presents the Words Pronounced Interface and Figure 4 presents the Results Interface.

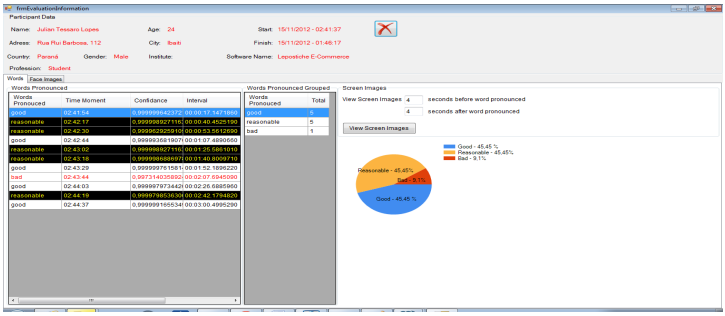

Fig. 3. Words Pronounced Interface

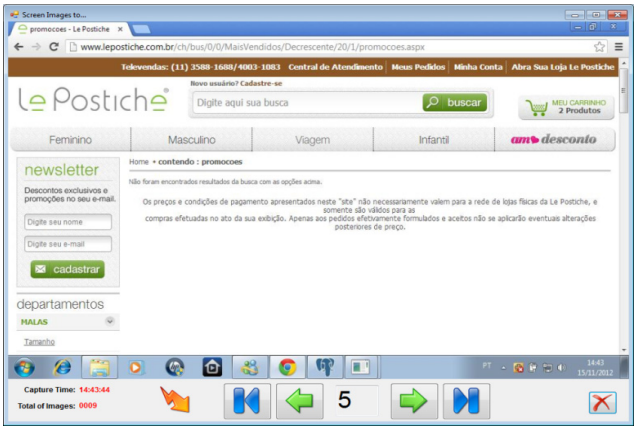

Fig. 4. Results Interface

The interface showed in Figure 4 displays information about the face images collected in the test such as Time Moment, Situation of the image (whether the interface is different from default image), Status (Discarded or Face Recognized). The Status Information refers to the capacity of recognizing or not a face and it was necessary because due to several reasons such as distractions, phone calls and others, the participant can be not looking for the camera and thus, the system is not able to recognize the face. Initially these images are discarded; however it must be important to analyze what moment the participant was not looking to computer. Two charts present the percentage of discarded images and faces that were recognized.

As well as the words pronounced, the face images displayed also allow the evaluator access the screen images that presents what the participant has did when the system collected that image. Figure 5 presents the Face Recognition Screens.

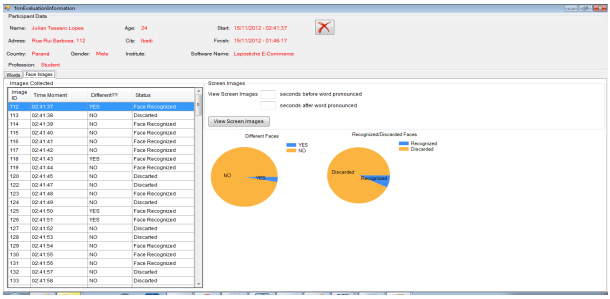

Fig. 5. Face Recognition Screens

The approach supported by ErgoSV Analyzer allows the evaluator (re-searcher) identify interfaces with usability problems in two ways: using words information or using faces information:

(1) Using words pronounced information: words such as Reasonable and Bad are highlighted calling attention of evaluator to possible usability problems and providing a series of interfaces in order to be analyzed. This resource avoids the need of reviewing all registered data to finding problems. For example, whether the participant pronounced the word “Bad” after ten minutes from starting the test, this word will be presented in the ErgoSV information allowing verify some interfaces before and after the pronunciation. In this way it is not necessary to review ten minutes of registered data to finding this information. In this case, in less than one minute the evaluator can know what interface have usability problems. Charts presenting quantitative information using the percentage of each word pronounced can provide real inputs about the general user opinion about the application;

(2) Using Face Recognition Information: after performing the evaluation, an image processing is performed to compare the captured face image to the default image, captured on the beginning of the test. A specific algorithm of image processing is used to performing the images comparison. Thus, two different images can be an important parameter that the participant had some reaction in this moment and so, something happened with this user. A series of screen images and words pronounced

can be accessed from the image register, supporting the evaluator to identifying the problems. Similarly, the information that the user was not looking for the camera highlights that some action had turned the user's attention and this situation may have been caused by the interface. Percentages of images collected, face recognized and images discarded can support evaluator making decisions quantifying the user's behavior through face image. The experiments also presents that face reaction is more involuntary then the pronouncement of a word leading the system processing a large number of different faces.

4 Discussion

The use of Speech and Face Recognition was considered satisfactory due the reason that it facilitates the data collecting processing allowing the participants perform their tasks without have to to note or mark something in a book or other software. The processing of pronounced words generated relevant information about software usability and allows user to identify problems interface in an easy and fast way.

The same results were noted in speech recognition similarly to face recognition. The ability to recognize different faces and whether the user was not looking for camera allowed the evaluator to identify possible problems interfaces beyond identify situation that distracted the participant. However the algorithm used to compare images still needs some calibration because it indicates some similar images as different. An improvement in this resource is being performing in order to provide a better image comparison.

Therefore, the results of proposed approach and the application used to support the experiments were considered satisfactory due to the reason that usability problems were identified based on specific user reactions, providing what moment and/or interface that needs improvement, beyond relevant information generated automatically by software avoiding the full registers review.

5 Conclusion

The observation method is an important and effective way to perform usability evaluation, mainly because it allows that real users test the application submitting it to situations similarly to real environment. Filming and Verbalization are two techniques widely used due to reason that collect the opinion and behavior of participants during the software using.

This paper presented the first results of a research to automate the generating information process using face and speech recognition. The use of these resources allowed identify easily and quickly which interfaces had usability problems due by processing specific user's reactions collected during the test, reducing time and cost in the review process. The data processing also allowed quantifying the usability test through the percentage of words and images, providing to evaluator a general idea about the users' opinion and their reactions. Currently we work in order to improve the cross reference information based on parameters such as age, gender, education

and other that can be created in the future. Also we intend to use the ErgoSV and ErgoSV Analyzer to performing evaluation in other software such as prototypes and Ecological software.

Acknowledgment. Financial Supported by FAPESP.

References

1. Cybis, W.A., Betiol, A.H., Faust, R.: Ergonomia e Usabilidade: conhecimentos, métodos e aplicações, 2nd edn. Novatec, São Paulo (2010)
2. Nielsen, J.: Usability Engineering. Morgan Kaufmann, Moutain View (1993)
3. Nielsen, J.: Designing Web Sites - Designing Web Usability. Campus (2000)
4. Preece, J., Rogers, Y., Sharp, H.: Interaction Design Beyond Human-Computer Interaction (2005); John-Wiley & Sons, Ltd. (2011)
5. Gonzalez, R.C., Woods, R.E.: Digital image processing. Addison-Wesley, Reading (1992)
6. Nunes, F.L.S.: Introdução ao processamento de imagens médicas para auxílio a diagnóstico – uma visão prática. Livro das Jornadas de Atualizações em Informática, 73–126 (2006)
7. Lima, J.P.S.M., et al.: Reconhecimento de padrões em tempo real utilizando a biblioteca OpenCV. Técnicas e Ferramentas de Processamento de Imagens Digitais e Aplicações em Realidade Virtual e Misturada, 47–89 (2008)
8. Coleti, T.A., Morandini, M., Nunes, F.L.S.: The Proposition of ErgoSV: An Environment to Support Usability Evaluation Using Image Processing and Speech Recognition System. In: IADIS Interfaces and Human Computer Interaction 2012 (IHCI 2012) Conference, Lisbon, vol. 1, pp. 1–4 (2012)
9. Neto, N., Patrick, C., Klautau, A., Trancoso, I.: Free tools and resources for Brazilian Portuguese speech recognition. In. J. Braz. Computing Society, 53–68 (2011), doi:10.1007/s13173-010-0023-1
10. Shariah, M.A., et al.: Human computer interaction using isolated-words speech recognition technology. In: International Conference on Intelligent and Advanced Systems 2007 (2007)
11. `http://www.laps.ufpa.br/falabrasil/` (accessed in December 2011)
12. Silva, P., Batista, P., Neto, N., Klautau, A.: An open-source speech recognizer for Brazilian Portuguese with a windows programming interface. In: Pardo, T.A.S., Branco, A., Klautau, A., Vieira, R., de Lima, V.L.S. (eds.) PROPOR 2010. LNCS, vol. 6001, pp. 128–131. Springer, Heidelberg (2010)