ErgoSV: An Environment to Support Usability Evaluation Using Face and Speech Recognition

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

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

Abstract. Usability test is a group of activities that should be performed by all designers in order to identify interaction problems. Filming and Verbalization are two techniques widely used due to the reason that they provide real information about the software interaction capacity. Filming is performed using one or several cameras and the verbalization is done encouraging the participant to verbalize what he/she is thinking about the software. Both techniques register the data in video and audio files to be analyzed forward. Although these techniques has been widely used, the analysis process is considered slow, difficult and expensive because the evaluator may need to review all the data registered from the first second until the end of the test to identify possible usability problems and this task could take from 2x to 10x the test time. This paper presents the ErgoSV Software, a tool to support usability evaluation test using speech processing that recognize specific keywords pronounced by the participants and face images processed during the test. These data are used to provide organized and relevant information to support the data analysis and the identification of interfaces with possible usability problems. Experiments performed in three different softwares presented that this tool reduced the time of analysis to 1,5 times the test time considering the keywords as the main data.

This research is supported by FAPESP.

Keywords: Usability Evaluation, Usability Test, Face Recognition, Speech Processing, Automatic analysis information.

1 Introduction

Usability is the main feature of interactive systems and, according to ISO 9241 should allow the users to perform their tasks with effectiveness, efficiency and satisfaction. Evaluating the software usability can guarantee that the users performs all their tasks in the system and do not reject the systems [4, 12,13].

Usability tests can be performed by designers in order to analyze whether the interaction has problems and so decrease the interface quality. Two techniques are widely used to test the usability: (1) filming: in this technique, the evaluator places one or several cameras to register images by the user, computer, environment and more information that they consider relevant; (2) think-aloud: the evaluator encourage the participant (final user) to verbalize what he/she is thinking about the system and

M. Kurosu (Ed.): Human-Computer Interaction, Part I, HCII 2014, LNCS 8510, pp. 554–564, 2014.

register the data in paper or audio files. The verbalization can be done simultaneously or consecutive with the test. In simultaneously approach, the participants perform their task and express in the same moment their opinion. In consecutive, the participant verbalizes after finishing the test and due to this reason, the consecutive approach is considered slower [4].

These techniques of test are considered too effectiveness due to the reason that provide real information about the software interaction capacity and so, allows the evaluator do input improvements in the interface besides to submit the software to real situations that could not be predicted by designers. However, the filming and the verbalization analysis data are slow and expensive and according to [12] can take long two to ten times the evaluation time [4, 15].

This paper presents the ErgoSV Software, an application developed by researchers of the University of Sao Paulo (Brazil) to support the usability test using filming and verbalization techniques. This tool was developed to register two events used as data: user face images collected by a image processing framework; keywords pronounced by participants that were registered by a speech recognition software.

The ErgoSV was developed and tested by real participants that performed real activities in three different systems and provided events that allow the evaluator to analysis the software usability.

The next section presents the bibliographic review used in this research.

2 Bibliographic Review

This section presents a bibliographic review performed in order to identify researches related with verbalization/think aloud method concepts and applications.

2.1 Speech Processing

People have 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 [16]. Speech Recognition (SR) is the voice interpretation process performed by a computer. It receives an external signal and through computational algorithms performs the transformation of the input data to obtain an output that can be analyzed as a text [11,17].

There are several methods and techniques to perform the SR. The main difference among the techniques is the number of processes performed to transform 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; and (3) display the final result [11, 17, 19].

The use of speech processing in different areas such as software development and biometrics raised the needs of tools to easily support the recognition activities in such way that developers do not need to know specific models. Aiming to solve this gap, the Laboratório de Processamento de Sinais (LAPS) in Federal University of Para – Brazil had developed the Coruja Application [17]. This application allows the use of complex speech processing functions in development environments such as Visual Studio coding few and little instructions, since the Coruja has all the complex algorithms implemented in low level.

2.2 Verbalization/Think Aloud

The Think Aloud Method (Verbalization) is a widely used technique that supports usability evaluation. However the initial studies using it were performed in the psychology area. Ericsson and Simon encouraged this technique and began using it similar way of the usability evaluation technique [1,3].

In an evaluation supported by verbalization, the evaluator encourages the participants (traditional users) to verbalize (speak) what they are thinking about the system allowing the evaluator collect real data about the user satisfaction with the system [4,2,13].

The verbalization can be performed according to two strategies [4]: (1) Simultaneously: the participants verbalize what they are thinking about the software in the same time that execute the task. This approach is considered effective because the participants are using the system and all their ideas can be clear in their minds. However this technique requires mental workload due to the reason that the users need to share attention with the verbalization and the use of the system, converting what they are seeing in an word and pronounce it; (2) Consecutive: the participants verbalize what they are thinking about the software after finishing all the tasks. This approach is considered less intrusive because users only perform their tasks using the system and, after finishing, they verbalize what they were thinking about the system, but it is considered slow due to the reason that the participants need to verbalize after the test and so, retarding the evaluation process. Although the participants do not share attention in using the system and in the verbalization, they need to remember what and why they did the activities, requiring a high mental workload.

Using this strategy, the evaluator should work as a manager in order to guarantee that the participants always verbalize some words. Whether the participant keeps more than sixty seconds without pronouncing a word, the evaluator should notify them with several terms such as ‘Keep talking’, ‘Is there any Doubt ?’, or ‘Do not stop talking’ [1,4].

The participants can pronounce any word or phrase according to their opinion such as ‘it is good’, ‘ I did not understand this screen’, ‘the colors are not good’ and any other that they think appropriate [1,4,12,13].

The data collection can be done in papers which the evaluator writes what the participants pronounce. The use of microphones, computers and voice recorders are also considered in order to facilitate de collecting and the processing [4]. The use of simultaneously or consecutive verbalization approach is a choice of the evaluator and the results of each test can vary according to user, software and test environment [1].

The verbalization is considered one of the most effective usability evaluation techniques and is used and encouraged by many researchers and specialists due to the reason that it provides real data about user satisfaction. The results of evaluations performed using this technique can vary according to specific contexts that are defined by the evaluators, tasks, participants and software contexts. The use of simultaneous or consecutive approach is an evaluator decision and must be done according to the evaluations needs as well as the data interpretation can be influenced by this choice [1,2,4].

A research performed by [8] presents that the use of the Think Aloud technique to support usability evaluation is considered as suitable by a great number of HCI designers and evaluators. In this study, ninety percent of the researchers and students used the verbalization in order to perform usability evaluation, as well as, seventy percent of the developers.

The next section presents the ErgoSV environment.

3 ErgoSV Software

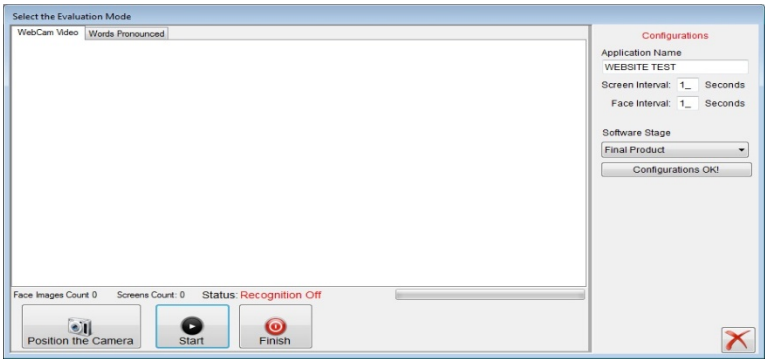
The ErgoSV Software was developed in order to support usability test using face and speech recognition as data to providing inputs that should allow the evaluator to identify interface with possible usability problems easily and safe.

Aiming to perform the data collect the system was developed using the Microsoft Visual C# Express Edition and contained two resources supported by frameworks: (1) Coruja [11,17]: this framework was used in ErgoSV to perform the speech recognition and write in the software a text with the word pronounced by user; (2) OpenCV [7]: used to perform the face recognition and the image processing activities. These frameworks were chosen because it can be used into the Visual C# Express easily and provide all the resources to access the image and speech recognition functions using few procedures and functions.

ErgoSV was divided in two modules in order to improve the evaluation/monitoring process. One module is used to performing the data collect and process initial data and; the other provides the information processed and organized with the data registered in the first module besides available screen images and details about the user´s events.

3.1 ErgoSV – Collect and Initial Processing

The data collecting stage is performed by the in order to support speech and face images recognition. The first step of the test is to fill a form with user self data. These data are used in the analysis stage to create cross reference information and identify who performed the tests. The next step is the configuration of the ErgoSV. This activity is necessary to guide the monitoring system in the test, and can be done in the ErgoSV main interface, presented in the Figure 1.



Application Name: WEBSITE TEST
Screen Interval: 1_ Seconds
Face Interval: 1_ Seconds
Software Stage: Final Product
Face Images Count 0
Screens Count: 0
Status: Recognition Off

Fig. 1. ErgoSV Initial Screen

The main interface is composed by three sections: tabs; form panel and button panel. The first section has two tabs: the firs tab (WebCam Video) presents the face images during the test in order to position the camera in front of the participant face and verify whether the application is recognizing the face; the other tab (Words Pronounced) presents a list of words recognized by ErgoSV using Coruja application. The tab also shows the register time and the confidence rate.

The form panel has fields to be filled by test self data such as the software tested name, the Screen Interval and the Face Interval. These data should receive inputs to guide the ErgoSV in the images registration activities. The intervals fields receive the time that the software will collect images of the participant face and software snapshots. There are not specific values, however, the ErgoSV suggests the time of three seconds due to reason that according to [12,13] is the middle time of a emotion expression, but the time should by a chosen of the evaluator.

To start the test, the participant should click in the button “Position the Camera” in order to position the webcam in front of their face and after this stage they can select the option “Start”. The Start Function starts the ErgoSV monitoring activities and minimizes the application aiming a less interference in the user activities.

The ErgoSV recognizes five keywords pronounced by participants: “Excellent”, “Good”, “Reasonable”, “Bad” and “Terrible” by default but can be replaced by any other group of words as wishes the evaluator.

3.2 ErgoSV Analyzer – Data Analysis and Information Generation

The data analysis is performed using the Analyzer module. The main objective of ErgoSV is the decreasing of the analysis time allowing the evaluator to identify easy and safe interfaces and resource with possible usability problems.

Initially, the software was tested using three different analysis approaches: only words data; only face images data and both words data and face images simultaneously. However, the use of the face image was not considered safe due to the reason that did not allow the easy and safe usability problems identification.

The use of words pronounced was considered appropriate to identify usability problems and provide safe information allowing the identification of possible interface to be reviewed. The analysis data should be performed according to the following steps: (1) select relevant words; (2) Insert interval value; (3) View interfaces or face images; (4) View face images easily; (5) Visualize and analyze interface or face images.

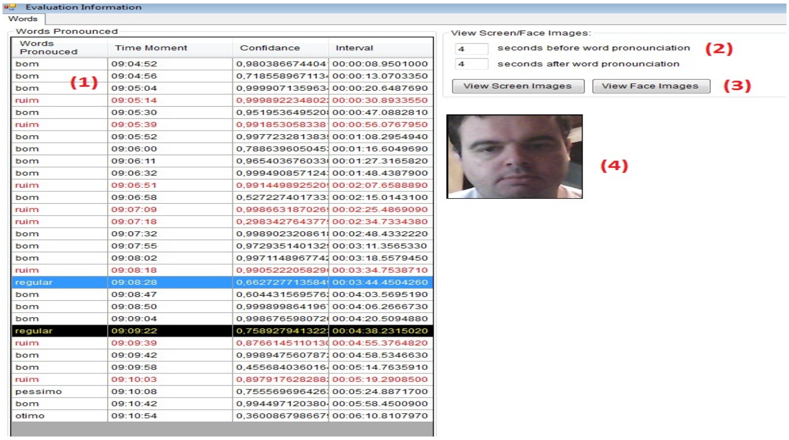
(1) Select relevant words
The evaluator should select a word pronounced by the participant that he/she considered relevant to analyzing. The ErgoSV highlights words considered as bad opinions such as “Regular”, “Bad” or “Terrible” due to the reason that these words can present interfaces that must be reviewed by designers.

(2) Insert Interval Value
The interval is the value of time that the ErgoSV should consider to select interfaces and/or face images from the word pronounced time. For example, a word “Bad” was pronounced in the time 10m20s after start the test. If the evaluator input the interval time as 4 seconds, the ErgoSV must select all interfaces image from 10m16s until 10m24s. The same search is performed using face images data.

(3) Select interfaces or face images
After the interval time had been defined, the evaluator can visualize the interfaces used by the participants or their face images from a word pronounced or from a specific interface.

(4) View face images easily;
The evaluator can access the participant face images in the pronounced moment using the image present in the words list right side.

Figure 2 presents the ErgoSV Analyzer interface with highlights to the four resources previously explained named as (1)..(4) in red colour.

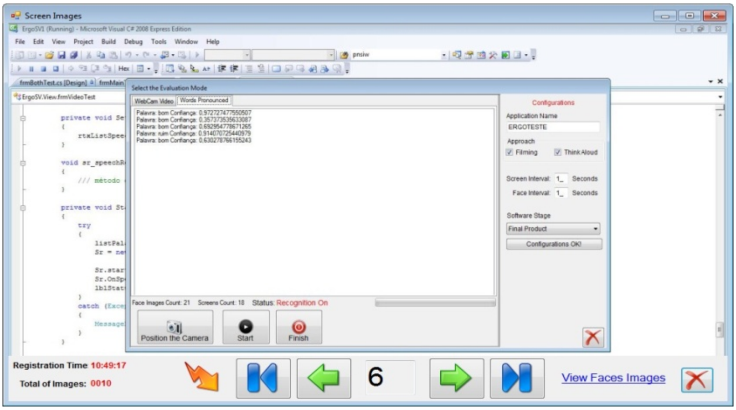


Words Pronouced	Time Moment	Confidance	Interval
bom	09:04:52	0,980386674404	00:00:08.9501000
bom	09:04:56	0,718558967113	00:00:13.0703350
bom	09:05:04	0,999907135963	00:00:20.6487690
ruim	09:05:14	0,999892234802	00:00:30.8933550
bom	09:05:30	0,951953649520	00:00:47.0882810
ruim	09:05:39	0,991853058338	00:00:56.0767950
bom	09:05:52	0,997723281383	00:01:08.2954940
bom	09:06:00	0,788639605045	00:01:16.6049690
bom	09:06:11	0,965403676033	00:01:27.3165820
bom	09:06:32	0,999490857124	00:01:48.4387900
ruim	09:06:51	0,991449892520	00:02:07.6588890
bom	09:06:58	0,527227401733	00:02:15.0143100
ruim	09:07:09	0,998663187026	00:02:25.4869090
ruim	09:07:18	0,298342764377	00:02:34.7334380
bom	09:07:32	0,998902320861	00:02:48.4332220
bom	09:07:55	0,972935140132	00:03:11.3565330
bom	09:08:02	0,997114896774	00:03:18.5579450
ruim	09:08:18	0,990522205829	00:03:34.7538710
regular	09:08:28	0,662727713584	00:03:44.4504260
bom	09:08:47	0,604431569576	00:04:03.5695190
bom	09:08:50	0,999899864196	00:04:06.2666730
bom	09:09:04	0,998676598072	00:04:20.5094880
regular	09:09:22	0,758927941322	00:04:38.2315020
ruim	09:09:39	0,876614511013	00:04:55.3764820
bom	09:09:42	0,998947560787	00:04:58.5346630
bom	09:09:58	0,455684036016	00:05:14.7635910
ruim	09:10:03	0,897917628288	00:05:19.2908500
pessimo	09:10:08	0,755569696426	00:05:24.8871700
bom	09:10:42	0,994497120380	00:05:58.4500900
otimo	09:10:54	0,360086798667	00:06:10.8107970

4 seconds before word pronounciation
4 seconds after word pronounciation

Fig. 2. ErgoSV Analyzer Interface

(5) Visualize and analyze interface or face images
This resource presents the interfaces or participant face images in moments near the moment of the word pronunciation. Figure 3 shows the resource to visualize the interfaces images.


Screen Images
ErgoSV1 (Running) - Microsoft Visual C# 2008 Express Edition
File Edit View Project Build Debug Tools Window Help
Select the Evaluation Mode
WebCam Video
Words Pronounced
Configurations
Position the Camera
Start
Finish
View Faces Images

```
Palavra: bom Confiança: 0,972727477550507
Palavra: bom Confiança: 0,357373535633087
Palavra: bom Confiança: 0,692954778671265
Palavra: ruim Confiança: 0,914070725440979
Palavra: bom Confiança: 0,630278766155243
```

Face Images Count: 21 Screens Count: 18 Status: Recognition On

Application Name: ERGOTESTE

Approach: ☑ Filming ☑ Think Aloud

Screen Interval: 1 Seconds

Face Interval: 1 Seconds

Software Stage: Final Product

Configurations OK!

Registration Time 10:49:17

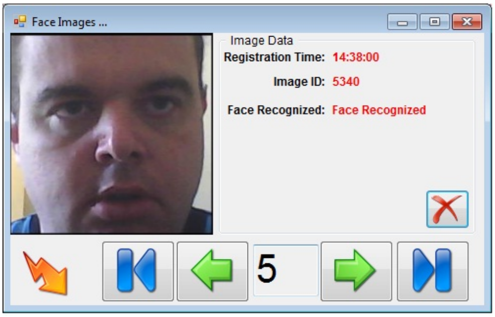
Total of Images: 0010

6

Fig. 3. Interface Visualization Resource

The interface visualization presents the snapshots registered in the moment of an event and near to it (considering the interval value). This resource has a main panel to present the image and a panel in the bottom of the interface with some information such as Registration Time, Total of Images loaded, navigation bar, and a image of a ray that highlight the interface used in the exactly moment of the event of pronounced the word.

A hyperlink name "View Face Images" provides a second interface that allows the visualization of an array of face images based on the time of interface registration and the interval input in the resource. Figure 4 presents the interface to visualize and navigate through participant face images.



Registration Time: 14:38:00

Image ID: 5340

Face Recognized: Face Recognized

5

Fig. 4. Face Images Visualization Interface

The use of the keywords as parameters to identify participant's opinions and the use of interfaces and faces images to support the analysis stage allowed the creation of the approached named as "Environment Tree" presented in the Figure 5.

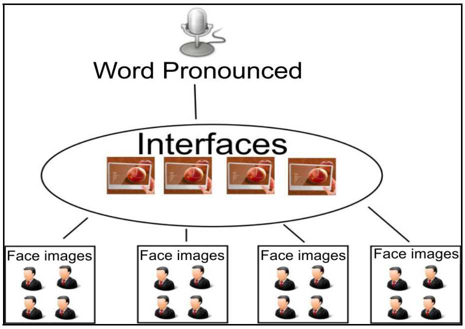


Fig. 5. Environment Tree by ErgoSV

4 Validation and Results

The validation of the ErgoSV tool was performed using three different software: an e-commerce website, a school website and a photo editor desktop software. All of participants performed several activities according to the application type such as searching for a product, buying a product, visualizing professor profile, modified a picture color or create a new pictures based on other images. Two hundred and one words were registered in the tests and distributed according to Table 1.

Table 1. Words Pronounced and Recognized by ErgoSV

	Excellent	**Good**	**Reasonable**	**Bad**	**Terrible**
E-Commerce	1	59	14	1	2
School	1	16	6	1	0
PhotoEditor	3	27	10	1	0
Total	5	102	30	3	2

Besides the words pronounced, the ErgoSV registered the participants' face images. Table 2 presents the numbers of images registered.

Table 2. Face images registered

Software	**Images**
E-Commerce	364
School	618
Photo Editor	800
Total	**1778**

Table 3 presents the total time of each application and the time limit of analysis presented in minutes.

Table 3. Time of test and analysis

Software	**Test Time**	**Analyze limit**
E-Commerce	71 min.	107 min.
School	25 min.	38 min.
Photo Editor	57 min.	86 min.

The analysis time limit was determinated as 1,5x the time test. There is no scientific parameter to this value, it was choosed due to the reason that it is less than the time presented by [12,13] as minimum time to analyze data and generate relevante information.

The analysis sequence was definied considering the keywords that could be collected in the experiments as tha main parameter to identify possible usability problems interface. Thus, the analysis was done studing:

- Keywords that meant opinions such as "Reasonable", " Bad" or "Terrible";
- Keyword that meant the great user opinion: Excellent;
- Special cases with the keyword "Good" .

The data analysis was performed using two different approaches: only words data; and both words data and faces images. Both approaches used the snapshots registered during the test.

The "Only Words" analysis approach was considered satisfactory, easy to use and fast. So, it allowed the evaluator to identify which interfaces were not good according to participants opinions. The time to identify the interfaces was low and the keywords selected to the experiments supported this activities appropriatedly because all the words had clear means, i.e., it was easy to identify whether the participant liked or disliked the interface.

However, this approach presented a problem: the evaluator did not identify which was the user focus in the moment of an event. For example, it was possible to find a bad interface, but this interface had several resources and the evaluators did not identify what resource were used by the participants. This problems was solved using the face imagens since the images provide the eyes position and so, it was possible to identify what was the focus of the participant in the moment of an event reducing the area to be analysed by the evaluator and providing a safe information about the resource classified by participant.

Figure 6 presents the time analysis comparing to time test and analysis time limit.

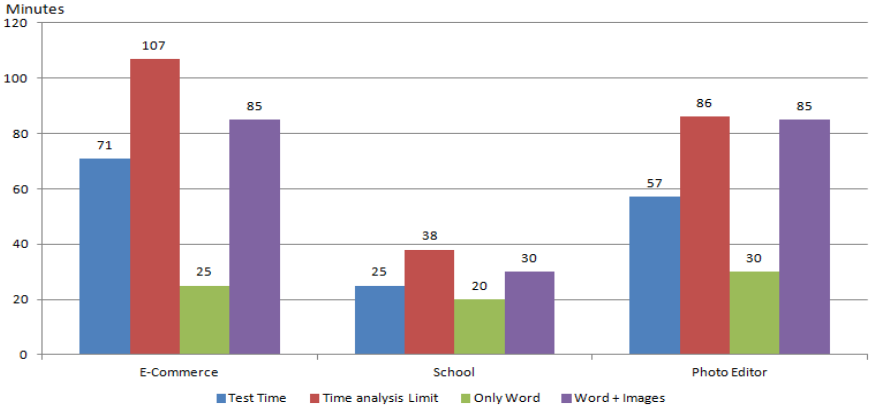


Fig. 6. Times achieved in the tests

5 Conclusions

Usability evaluation is a group of activities that must be performed in order to verify whether the interface has usability problems. The usability test is a technique of usability evaluation that must be realized to test the software interaction capacity, i.e., how the interface/interaction interfere in the participant activities.

Filming and verbalization are two widely used techniques, however are considered slowed due to the reason that the evaluator needs to review a vary amount of data manually and sequentially.

This paper presented the ErgoSV software, a tool that uses speech and face images recognition to support the collection and data analysis in usability tests. The focus of this research was the decrease of the time to identify possible usability problems in the interfaces. The use of keywords with significant means supported the identification of users opinions reducing the time to identify possible problems. The ErgoSV provided a highlight to keywords that could be relevant for analysis and so, the evaluator could easily and safely identify the problems. The interface and face images visualization resource allows the evaluator to accomplish what happen in the moment of an event and few seconds before and after this moment.

Finally, the experiments presented that this tool reduced the time to identify the interfaces with possible usability problems from 2 to 10 times the test time to 1,5 times. The use of face images allowed the identification of the user focus supporting the analysis of the interface and the classification of which resources were used by the user.

Acknowledgments. The authors are grateful to Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

References

1. Boren, M.T., Ramey, J.: Thinking aloud: Reconciling theory and practice. IEEE Transactions on Professional Communication, 261–278 (2000)
2. 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)
3. Cooke, L.: Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. IEE Trans. Prof. Commun., 202–215 (2010)
4. Cybis, W.A., Betiol, A.H., Faust, R.: Ergonomia e Usabilidade: conhecimentos, métodos e aplicações, 2nd edn., Novatec, São Paulo (2010)
5. Gonzalez, R.C., Woods, R.E.: Digital image processing. Addison-Wesley, Reading (1992)
6. Hertzum, M., Hansen, K.D., Andersen, H.H.: Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? Behavior & Information Technology, 165–181 (2009)
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. Mcdonald, S., Edwards, H.M., Zhao, T.: Exploring think-alouds in usability testing: An international survey. IEEE Transactions on Professional Communication (2011)
9. Morandini, M.: Ergo-Monitor: Monitoramento da Usabilidade em Ambiente Web por Meio de Análise de Arquivos de Log. Tese (Doutorado) - Universidade Federal de Santa Catarina (2003)
10. Morandini, M., de Moraes Rodrigues, R.L., Cerrato, M.V., Chaim, M.L.: Project and Development of ErgoCoIn Version 2.0. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part I, HCII 2011. LNCS, vol. 6761, pp. 471–479. Springer, Heidelberg (2011)
11. Neto, N., Patrick, C., Klautau, A., Trancoso, I.: Free tools and resources for Brazilian Portuguese speech recognition. J. Braz. Computing Society, 53–68 (2011), doi:10.1007/s13173-010-0023-1
12. Nielsen, J.: Usability Engineering. Morgan Kaufmann, Moutain View (1993)
13. Nielsen, J.: Designing Web Sites - Designing Web Usability, Campus (2000)
14. Nunes, F.L.S.: Introdução ao processamento de imagens médicas para auxílio a diagnóstico – uma visão prática. In: Livro das Jornadas de Atualizações em Informática, pp. 73–126 (2006)
15. Preece, J., Rogers, Y., Sharp, H.: Design de Interação: Além da interação homem-computador, Bookman, Porto Alegre, Rio Grande do Sul – Brasil (2005)
16. Shariah, M.A., et al.: Human computer interaction using isolated-words speech recognition technology. In: International Conference on Intelligent and Advanced Systems (2007)
17. Silva, P., et al.: An open-source speech recognizer for Brazilian Portuguese with a windows programming interface. In: The International Conference on Computational Processing of Portuguese (PROPOR) (2010)
18. `http://www.laps.ufpa.br/falabrasil/` (accessed in December 2011)
19. Agus, T., et al.: Characteristics of human voice processing. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 509–512. [S.l.: s.n.] (2010)
20. ISO9241. Ergonomic requirements for office work with visual display terminals